TERM ASSOCIATION MODELLING IN INFORMATION RETRIEVAL

JIASHU ZHAO

A DISSERTATION SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN COMPUTER SCIENCE AND ENGINEERING
YORK UNIVERSITY
TORONTO, ONTARIO
MARCH 2015

Abstract

Many traditional Information Retrieval (IR) models assume that query terms are independent of each other. For those models, a document is normally represented as a bag of words/terms and their frequencies. Although traditional retrieval models can achieve reasonably good performance in many applications, the corresponding independence assumption has limitations. There are some recent studies that investigate how to model term associations/dependencies by proximity measures. However, the modeling of term associations theoretically under the probabilistic retrieval framework is still largely unexplored.

In this thesis, I propose a new concept named Cross Term, to model term proximity, with the aim of boosting retrieval performance. With Cross Terms, the association of multiple query terms can be modeled in the same way as a simple unigram term. In particular, an occurrence of a query term is assumed to have an impact on its neighboring text. The degree of the query term impact gradually weakens with increasing distance from the place of occurrence. Shape functions are used to

characterize such impacts. Based on this assumption, I first propose a bigram CRoss TErm Retrieval ($CRTER_2$) model for probabilistic IR and a Language model based model $CRTER_2^{LM}$. Specifically, a bigram Cross Term occurs when the corresponding query terms appear close to each other, and its impact can be modeled by the intersection of the respective shape functions of the query terms. Second, I propose a generalized n-gram CRoss TErm Retrieval ($CRTER_n$) model recursively for n query terms where $n > 2$. For n-gram Cross Term, I develop several distance metrics with different properties and employ them in the proposed models for ranking. Third, an enhanced context-sensitive proximity model is proposed to boost the $CRTER$ models, where the contextual relevance of term proximity is studied. The models are validated on several large standard data sets, and show improved performance over other state-of-art approaches. I also discusse the practical impact of the proposed models. The approaches in this thesis can also provide helpful benefit for term association modeling in other domains.

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisor Prof. Jimmy Xiangji Huang for his guidence, motivation, enthusiasm, and rigorous attitude. During my Ph.D. study with Prof. Jimmy Xiangji Huang, I was introduced to Information Retrieval and become a researcher in this area. With his patience, I was able to change my major smoothly from a Mathematical Master to a Ph.D. in Computer Science. Prof. Jimmy Xiangji Huang also provided me a lot of opportunities to conferences and introduced me to many famous professors in the field. I have never been so close to the cutting edge of Science. I am grateful to have Prof. Jimmy Xiangji Huang as my Ph.D. supervisor.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Nice Cercone and Prof. Aijun An, for their advice, knowledge, and insightful comments.

I also would like to thank my dissertation Examining Committee: Prof. Walter P. Tholen, Prof. Norbert Fuhr, and Prof. Mahmudul Anam. They provided many

insightful and detailed suggestions for improving my thesis.

I offer my regards and blessings to all my lab mates: Mariam Daoud, Ben He, Vivian Hu, Jun Miao, Justin Wu, Zheng Ye, Xiaoshi Yin, Xiaofeng Zhou, and so on. They are my best friends and study partners.

Last but not least, I would like to thank my family. My beloved father Shuyuan Zhao and mother Fuxia Liu always take care of me and support me in my whole life. My husband, Shicheng Wu, always encourages my study and research. Because of them, I can sit down and write my papers and this thesis with a peaceful mind.

Table of Contents

List of Tables

List of Figures

1 Introduction and Motivation

Information Retrieval (IR) aims to find relevant information resources to a query from a collection of information resources. Queries are statements of information needs, and are usually formed as a series of keywords. An automated information retrieval system takes the query as input and outputs a ranked list of documents with different degrees of relevancy. Due to the purpose of effectiveness and efficiency, the documents in the collection are usually preprocessed into their indexed representations, and the queries are prepossessed into the corresponding representations. Figure 1.1 shows a basic IR system, where an IR weighting model matches document representations with a query representation and generates a list of relevant documents. I focus on the IR model part of the retrieval system, and propose new IR models to promote the retrieval performance, which is, providing more relevant documents.

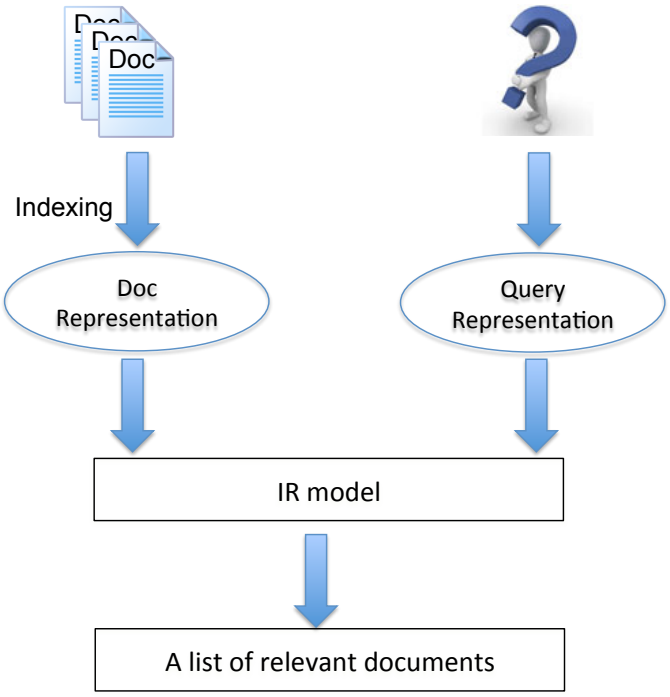


Figure 1.1: A basic IR system

1.1 Motivation

Different IR systems have their own indexing schema. The gap between computer storage and human memory is that the computers can only "read" word by word. When forming the document representation, a standard approach is to extract keywords and to represent a document by the keywords and keyword frequencies. Figure 1.2 shows an example where the documents are processed and stored as terms and corresponding frequencies. In this process, the terms in the documents are assumed to be independent from each other. Meanwhile, the queries are usually processed in

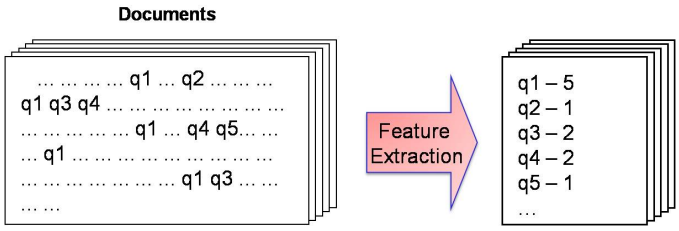


Figure 1.2: An example of document representation

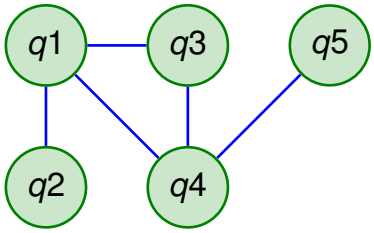


Figure 1.3: An example of associations among terms

a similar way that the query terms are independent.

Many traditional IR models are based on the independence assumption that terms are independent of each other. Although traditional retrieval models can achieve reasonably good performance in many applications, the corresponding independence assumption has limitations. From human being's understanding, the terms may have some dependencies on each other. In other words, there might be some implied associations among the query terms, as shown in Figure 1.3. In order to provide more relevant documents to the users, associations among query terms should be also considered.

Figure 1.4: An Example of Term Association for "Earthquake in Canada"

For a more detailed example, given an input query "Earthquake in Canada" [1], there exists an association between the query terms. Users are looking for neither other events in Canada nor earthquakes in other countries. Figure 1.4 shows two documents with both "earthquake" and "Canada" occurring twice. A traditional IR

[1]The term "in" is usually treated as a stop word and removed during prepossessing. Therefore, "Earthquake in Canada" is treated as two terms "Earthquake" and "Canada".

model will assign these two documents the same weights. However, if we read the documents in detail, we can see that the first document (in Figure 1.4(a)) reports an earthquake in Japan, and the second document (in Figure 1.4(b)) reports an earthquake in Canada. Obviously the second document is more relevant to the user's query than the first document. Therefore, it is necessary to reward the document in which the matched query terms have a stronger association.

Many recent studies in IR are trying to address this problem from various ways. The approaches considering term associations have shown to be effective in many IR applications (Beigbeder and Mercier 2005, Gao et al. 2004, Hawking and Thistlewaite 1995). However, the nature of the associations among query terms still awaits further study. Some proximity approaches only consider adjacency (Song and Croft 1999, Srikanth and Srihari 2002), while non-adjacent terms may also have associations. N-gram models (Ahmed and Nurnberger 2009, Mayfield and McNamee 2003) consider n word sequences, which expand the radius of matching. It is yet hard to determine the optimal n, and the complexity usually grows exponentially with the growth of n. Other proximity based probabilistic weighting models, such as (Broschart and Schenkel 2008, Büttcher et al. 2006), add proximity information into their weighting functions in a heuristic manner. However, their experiments are not conclusive and their retrieval functions are not shown to be effective and robust enough (Tao and

Zhai 2007).

1.2 Main Contributions

In this thesis, a new perspective is provided to address the problem mentioned above. In particular, I focus on proposing new models to integrate the associations among multiple query terms in existing IR models. First, a bigram Cross TErm Retrieval model ($CRTER_2$) is proposed as the basis model for two query terms. The association between two query terms are simulated by a pseudo term, bigram Cross Term. By proposing the concept of Cross Term, term association can be modelled by computing the weight of the corresponding Cross Term in a manner that is theoretically similar to that used for a single query term. Second, the concept of bigram Cross Term is extended to n-gram Cross Term. An n-gram Cross TErm Retrieval model ($CRTER_n$) is proposed for n query terms where $n > 2$. $CRTER_n$ is modelled based on $CRTER_2$,...,$CRTER_{n-1}$ and the n-gram Cross Term. Third, an enhanced context-sensitive proximity is proposed to integrate context information into the proposed $CRTER$ model. Fourth, extensive experiments have been conducted on standard real data sets are conducted to illustrate the effectiveness and the stability of the proposed models. Fifth, I also discuss the practical impact of the proposed models in this thesis. The approaches in this thesis can also provide helpful benefit

for various other domains.

In detail, Cross Term in this thesis is a pseudo term that is generated by two or more query terms occurring close to each other. I assume that an occurrence of a query term has an impact on its neighboring text. This impact attenuates when the position of a neighboring term is farther away. If we try to characterize this impact with a mathematic function, intuitively the function should satisfy the following properties: Non-negative, Continuous, Symmetric, Monotonic, Identity (See details in Chapter 3). I use kernel functions that have been brought to proximity retrieval (de Kretser and Moffat 1999, Lv and Zhai 2009) to estimate the query term occurrences' impact. They are Gaussian Kernel, Triangle Kernel, Circle Kernel and Cosine Kernel. I propose to investigate three more kernel functions that satisfy the above properties: Quartic Kernel, Epanechnikov Kernel, and Triweight Kernel. The kernel functions are normalized by scaling between 0 and 1.

This thesis first illustrates the idea of a Cross Term for two query terms as a special case. We say a bigram Cross Term $q_{i,j}$ occurs, when two query terms, q_i and q_j, occur in close proximity in a document and therefore their impact shape functions have an intersection. q_i and q_j are called the generating terms of $q_{i,j}$. The corresponding impact shape functions' value at this intersection is bigram Cross Term $q_{i,j}$'s occurrence value, which ranges from 0 to 1. According to the impact

shape function's properties, the closer two query terms' occurrences are, the higher the generated bigram Cross Term's occurrence value is. Therefore, the bigram Cross Term's occurrence value indicates to what extent two query terms are correlated.

Many previous term association approaches treat term association as an extra part, for instance, adding an extra score to the weighting function (Tao and Zhai 2007). I evaluate the term association directly by regarding the Cross Term as a special term in the existing IR weighting models. Terms are the most fundamental element in IR. By manipulating Cross Terms, we will have a better understanding of term association and a better approach to controlling its effects.

In the probabilistic weighting model, some variants change from term to term, namely the within-document term frequency (tf), the number of documents containing the term (nd), and the within-query term frequency (qtf), respectively. I define the corresponding variants for Cross Terms: $tf(q_{i,j}, D)$, $nd(q_{i,j})$, and $qtf(q_{i,j})$. For $tf(q_{i,j}, D)$, as a pseudo term, the traditional counting method of the occurrences in a document does not make much sense, especially when we aim to give more weight to occurrences of Cross Terms where the generating query terms are closer. Instead of accumulating the number of occurrences, a Cross Term's value in a document is accumulated as its $tf(q_{i,j}, D)$, which is small when q_i and q_j are far away, and large when q_i and q_j are close to each other. Please note that $tf(q_{i,j}, D)$ is always smaller

than the number of occurrences of a Cross Term in the document. In order to balance other variants with $tf(q_{i,j}, D)$, $nd(q_{i,j})$ and $qtf(q_{i,j})$ are defined correspondingly. $nd(q_{i,j})$ is the accumulated Cross Term's average value on each document over the collection. In a query, since it is normally short and only contains query terms, that terms in a query are assumed to be densely distributed in this thesis. If two terms exist in a query, they are regarded as being adjacent. Then $qtf(q_{i,j})$ is a simplified case of within-document frequency by treating the query as a document.

The Cross Terms' variants are also defined on the basis of another popular IR model, Language Model (LM). The bigram Cross Term in LM shares some variants as in the probabilistic weighting model as discussed above, and the same definitions and estimations are kept for these variants. One extra variant for LM which does not exist in the probabilistic weighting model is $cf(q_{i,j})$, which is $q_{i,j}$'s frequency over the collection. In this thesis, $cf(q_{i,j})$ is estimated by accumulating $tf(q_{i,j})$ over the collection.

The concept of Cross Term is further extended to multiple terms in this thesis, in order to be able to capture the association of more than two query terms. When there are n query terms ($n > 2$), there is no single intersection where all the term impact functions cross each other. Therefore, we have to find an alternative way to define an n-gram Cross Term. A basic component for bigram Cross Term is the distance

between two query terms. In the higher dimensional case, several advanced distance metrics are proposed to characterize the occurrences of multiple query terms, and thereby define n-gram Cross Terms.

With the defined Cross Term variants, Cross Terms are integrated into the traditional IR weighting models by treating them as special terms. BM25 weighting model (Robertson et al. 1996), which usually provides very effective performance, is applied as the basic probabilistic weighting model. For LM, several approaches smooth LM to solve the zero probability problem and to achieve better performance. I focuse on Dirichlet smoothing which is one of the most popular approaches and reported to have the best average precision in most of the cases (Zhai and Lafferty 2001, 2004). In these basic IR models, bigram Cross Terms' weights are computed and linearly combined with query terms' weights as the bigram Cross TErm Retrieval model $CRTER_2$. In order to distinguish the applied basic IR models, $CRTER_2$ denotes the probabilistic BM25 based bigram model, and $CRTER_2^{LM}$ denotes the LM based bigram model. The n-gram Cross TErm Retrieval models $CRTER_n$ and $CRTER_n^{LM}$ are proposed recursively based on the defined n-gram Cross Terms in a recursive way. I also discuss the time complexity of the proposed models and how to implement the model in a more efficient way.

I also propose to enhance association-based probabilistic retrieval models with

more contextual information in this thesis. A term pair with higher contextual relevance of term proximity is assigned a higher weight. Several measures are proposed to estimate the contextual relevance of term proximity I propose a proximity enhancement approach to integrate the contextual relevance of term proximity into the retrieval process. The top ranked documents from a basic weighting model are assumed to be more relevant to the query, and the contextual relevance of term proximity are calculated using the top ranked documents. I propose a context-sensitiveproximity model, and the experimental results on standard TREC data sets show the effectiveness of this enhanced model.

1.3 Outline

This thesis contains seven chapters. The chapters are organized as follows. Chapter 1 presents the research problem and the motivation, and introduces the contributions of this thesis. Chapter 2 discusses prior related work including basic weighting models in IR and existing term association approaches. Chapter 3 introduces the concept of bigram Cross Term, defines its variants, and proposes bigram Cross Term Retrieval models. Chapter 4 extends the concept of Cross Term for modeling multi-term associations, and proposes n-gram Cross Term Retrieval models. Chapter 5 proposes an enhanced term association model by integrating the contextual infor-

mation. Chapter 6 summarizes the work in this thesis and concludes the findings. Chapter 7 concludes this thesis and discusses potential future work.

2 Related Work

The literature review presents other research work related to this thesis. Section 2.1 describes the basic ranking models in IR. Then the existing term association models are introduced in Section 2.2. Section 2.3 summarizes the related work and discusses their connections with this thesis.

2.1 Ranking Models in Information Retrieval

The first applications of Information Retrieval are in mechanized library system (Maron and Kuhns 1960), where the concept of "Relevance" is introduced as a key factor. Instead of string matching, indexing is a faster and more flexible approach to collect, parse and store data. In an index, each document is represented by a set of weighted keywords. The query representations and the document representations are then compared to calculate a score, as shown in Figure 1.3. Researchers have proposed several IR models to calculate the relevance scores between queries and

documents.

Boolean model and Vector Space Model are two representative early weighting models. They are briefly introduced in Section 2.1.1 and 2.1.1. Even though they have disadvantages to some extent, these two models are important to the history of IR and some popular models are form of their extensions. Probabilistic-based model BM25 is presented in Section 2.1.3, which is one of the most effective IR weighting models. Statistic-based Language Model (LM) is shown in Section 2.1.4, which is also one of the most popular IR weighting models.

2.1.1 Boolean Model

The Boolean model in Information Retrieval (BIR) (Goker and Davies 2009, Hsiao and Harary 1970) is the first IR model. The BIR is based on Boolean logic and classical set theory in that both the documents to be searched and the user's query are conceived as sets of terms. Retrieval is based on whether or not the documents contain the query terms. Query terms are combined with three basic operators, the logical product AND, the logical sum OR and the logical difference NOT. The Boolean model gives users a sense of control over the retrieval system. On the other hand, the Boolean model is hard for untrained users to manipulate correctly. Its main disadvantages are: (1) the retrieved documents are either scored 0 or 1, not a

ranked set of documents; (2) all terms are equally weighted; (3) exact matching may retrieve too few or too many documents.

2.1.2 Vector Space Model

In the Vector Space model (Salton et al. 1975), both queries and documents are represented as vectors of terms.

$$Q = (w_{1,Q}, ..., w_{n_v,Q})$$

$$D = (w_{1,D}, ..., w_{n_v,D})$$

where w is the weight for a dimension, and n_v is the dimensionality of the vectors in the Vector Space model. The definition of term depends on the application. Typically terms are single words, keywords, or longer phrases. If the words are chosen to be the terms, the dimensionality of the vector is the number of words in the vocabulary[2]. The Vector Space model assigns non-binary weights to index terms in queries and in documents, and these term weights are used to compute the similarity between a query and documents in the collection (Baeza-Yates et al. 1999). Some similarity functions are: Inner product, Cosine, Dice, and Jaccard (Nassar et al. 2013).

[2]The vocabulary includes distinct terms occurred in the collection

- Inner Product

$$Q \cdot D = \sum_{i=1}^{n_v} w_{i,Q} \times w_{i,D}$$

- Cosine Similarity

$$Cosine(Q, D) = \frac{\sum_{i=1}^{n_v} w_{i,Q} \times w_{i,D}}{\sqrt{\sum_{i=1}^{n_v} w_{i,Q}^2} \times \sqrt{\sum_{i=1}^{n_v} w_{i,D}^2}}$$

- Dice Similarity

$$Dice(Q, D) = \frac{2 \sum_{i=1}^{n_v} w_{i,Q} \times w_{i,D}}{\sum_{i=1}^{n_v} w_{i,Q}^2 + \sum_{i=1}^{n_v} w_{i,D}^2}$$

- Jaccard Similarity

$$Jaccard(Q, D) = \frac{\sum_{i=1}^{n_v} w_{i,Q} \times w_{i,D}}{\sum_{i=1}^{n_v} w_{i,Q}^2 + \sum_{i=1}^{n_v} w_{i,D}^2 - \sum_{i=1}^{n_v} w_{i,Q} \times w_{i,D}}$$

Compared to BIR presented in the early section, the Vector Space model computes a continuous degree of similarity between queries and documents and supports partial matching, where the term weights are not binary. The main disadvantage of the Vector Space model is that it does not define appropriate values to the vector components (Goker and Davies 2009). Some researchers (Salton and Yang 1973) suggested a combination of term frequency tf, and idf, the inverse document frequency,

which is the number of documents that contain the term.

2.1.3 Probabilistic Model

The Probabilistic model (Fuhr 1992) defines term weighting based on probability theory. The probability of a query Q and a document D are relevant is $P(R|Q,D) = \frac{P(D|R,Q)*P(R|Q)}{P(D|Q)}$. BM25 is a probabilistic weighting function employed by the Okapi system (Robertson et al. 1996). As shown by previous TREC experimentation, BM25 usually provides very effective retrieval performance on the TREC collections that are used in (Voorhees and Harman 2005). In BM25, a search term is assigned a weight based on its within-document term frequency and query term frequency (Zhao et al. 2010, 2009). The corresponding weighting function is as follows.

$$\begin{aligned} w = & \frac{(k_1+1)*tf}{K+tf} * \log \frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)} * \frac{(k_3+1)*qtf}{k_3+qtf} \\ \oplus \quad & k_2 * nq * \frac{(avdl-dl)}{(avdl+dl)} \end{aligned} \tag{2.1}$$

where w is the weight of a query term, N is the number of indexed documents in the collection, n is the number of documents containing a specific term, R is the number of documents known to be relevant to a specific topic, r is the number of relevant documents containing the term, tf is within-document term frequency, qtf

is within-query term frequency, dl is the length of the document, $avdl$ is the average document length, nq is the number of query terms, the k_is are tuning constants (which depend on the database and possibly on the nature of the queries and are empirically determined), K equals to $k_1 * ((1-b) + b * dl/avdl)$, and $\oplus$ indicates that its following component is added only once per document, rather than for each term. In our experiments, the values of k_1, k_2, k_3 and b are set to be 1.4, 0, 8 and 0.55 respectively.

2.1.4 Language Model

Language Model (Ponte and Croft 1998) is interested in estimating the probability that document D generates the observed query Q, $p(D|Q) \propto p(Q|D)p(D)$. p(D) is the prior probability that D is relevant to any query, which does not affect document ranking. The weight of a matched term q_i can be identified as

$$w(q_i, D) = \frac{p_s(q_i|D)}{\alpha_d P(q_i|C)} \tag{2.2}$$

where model $p_s(q_i|D)$ used for "seen" words that occur in the document, $P(q_i|C)$ is the collection language model, and α_d is a document-dependent constant. The simplest method to estimate $p(w|D)$ is maximum likelihood $p_{ml}(w|D) = \frac{c(w;D)}{\sum_w cw,D}$. However, the maximum likelihood estimation usually under-estimates the unseen

word in the document. Many smoothing techniques were proposed, and popular smoothing techniques in IR are the Jelinek-Mercer method, Bayesian smoothing using Dirichlet priors, and Absolute discounting (Zhai and Lafferty 2004).

- Jelinek-Mercer

$$P_\lambda(w,d) = (1-\lambda)p_{ml}(w|d) + \lambda p(w|C)$$

- Dirichlet

$$P_\mu(w,d) = \frac{c(w;d) + \mu p(w|C)}{\sum_w c(w;d) + \mu}$$

- Absolute discounting

$$P_\delta(w,d) = \frac{max(c(w;d) - \delta, 0) + \delta|d|_u}{\sum_w c(w;d)}$$

where λ, μ and δ are parameters, and $|d|_u$ is the number of unique terms in d.

2.2 Term Association Models

Since the terms are not independent from each other in the real world, the relevancy between a query and a document can be better characterized by matching based on multiple terms. In this section, I first introduce traditional term association approaches, phrase-based IR and similarity coefficients for IR in Section 2.2.1

and Section 2.2.2. Further, topic models are discussed for term associations in Section 2.2.3. Section 2.2.4 focuses on approaches based on BM25 and LM. In Section 2.2.5, I discuss the term associations in terms of the number of grams: bigram term associations are compared with n-gram associations.

2.2.1 Phrase-based Term Association Approaches

Since the 1990s, some early researchers started to investigate term association approaches in IR, and they found such approaches to be effective. The query term associations have been modeled by different approaches according the distance of the query terms in documents. For example, Allan and Ballesteros (Allan et al. 1995) indexed phrases instead of terms with InQuery (Callan et al. 1992), and obtained improvements in TREC campaigns. The weighting formula of phrases is a $tf \cdot idf$ formula, where the tf is a combination of Okapi and InQuery.

This approach can only handle query terms that are adjacent to each other in the documents. Intuitively, however, the query terms that are not adjacent to each other, but occur closely to each other may also carry some associations.

2.2.2 Similarity Coefficients

This group of approaches builds arrays for terms and utilizes various similarity coefficients to measure term associations. In (Jones and Jackson 1970), the distribution of terms in document descriptions was represented as a co-occurrence array, and a similarity matrix was derived by the weighted and unweighted Tanimoto similarity definition. Qiu and Fei (Qiu and Frei 1993) computed the similarity between any two terms by the vector product, and built elements in the term vector by

$$d_{ik} = \frac{(0.5 + 0.5\frac{ff(d_k,t_i)}{maxff(t_i)}) \cdot iif(d_k)}{\sqrt{\sum_{j=1}^{n}((0.5 + 0.5\frac{ff(d_k,t_i)}{maxff(t_i)}) \cdot iif(d_k))^2}}$$

where $ff(d_k, t_i)$ is the term frequency of term t_i in document d_k, $iif(d_k) = \frac{m}{|d_k|}$ is the inverse item frequency of document d_k, m is the number of documents in the collection and $|d_k|$ is the number of terms appearing in document d_k, $maxff(t_i)$ is the maximum term frequency of t_i in all documents.

Although many techniques in this area have been tested and some interesting results were obtained, most of the techniques have been used to do query expansion, and few studies on document modeling with term similarity coefficients have been conducted (Wei and Croft 2007).

2.2.3 Latent Topic Models for Term Associations

In the latent topic models, the terms are associated with each other with respect to the latent topics. From the collection, these models estimate the probability of a topic generated from a document $P(topic|doc)$ and the probability of a term generated from from a topic $P(term|topic)$. The terms that are associated with a same topic could be recognized as being associated with each other. Standard Latent Semantic Analysis stems from linear algebra and performs a Singular Value Decomposition of co-occurrence tables. Probabilistic Latent Semantic Analysis (PLSA) (Hofmann 1999) is based on a mixture decomposition derived from a latent class model, where its probabilistic variant has a sound statistical foundation and defines a proper generative model of the data. The Latent Dirichlet Allocation (LDA) model (Blei et al. 2003) assumes that the topic distribution has a Dirichlet prior, which results in more reasonable mixtures of topics in a document. However, these latent topic models are very expensive and difficult to apply on large collections even with online learning techniques (Hoffman et al. 2010), and a new training process is required for each new collection (Wei and Croft 2007).

However, latent mixture models are usually very expensive and difficult to apply on large collections. There is often no exact inference techniques for these models and approximation techniques have to be adopted to iteratively approach the solution.

Parameter tuning for these complicated models makes them even more expensive. Furthermore, they require a new training process for each new collection; in contrast, term-term associations can often be used across collections.

2.2.4 BM25-based and LM-based Term Association Approaches

Researchers have developed various term association approaches on different types of IR models. Some studies heuristically integrated word proximity into probabilistic weighting models, such as (Broschart and Schenkel 2008, Büttcher et al. 2006, He et al. 2011, Hu et al. 2012, Rasolofo and Savoy 2003, Tao and Zhai 2007, Wong and Yao 1993, Ye et al. 2012). Some work also found that term proximity was a useful measure for selecting query expansion terms in probabilistic IR (Miao et al. 2012). (Song et al. 2011) used a position dependent term count to represent both the number of occurrences of a term and the term counts propagated from other terms. However, it is still largely unexplored to theoretically model term associations under the probabilistic retrieval framework (Pickens 2000, Robertson and Walker 1994, Wang and Si 2008).

For statistical LM, some recent research work embedded the term association information in LM. For example, (Zhao and Yun 2009) viewed query terms' proximate centrality as Dirichlet hyper-parameters to assign weights to the parameters of the

multinomial document language models. For discriminative models, dependency weights were assigned to term pairs in (Shi and Nie 2010). Our proposed $CRTER$ models can be applied to any traditional IR models, since the term associations are regarded as pseudo terms, and terms are the most fundamental elements in IR.

2.2.5 Bigram and N-gram Term Association Approaches

In general, bigram IR models are popular in modeling terms associations and can make a better balance between effectiveness and complexity. For example, Song and Croft (Song and Croft 1999) proposed a general language model that combined bigram language models with several smoothing techniques including a Good-Turing estimate and corpus-based smoothing of unigram probabilities. The relative contributions of the different models to the query generation probability were determined empirically. Srikanth and Rohini (Srikanth and Srihari 2002) proposed a biterm language model, which approximated the biterm probabilities using the frequency of occurrence of terms. Biterm language models are similar to bigram language models except that the constraint of order in terms is relaxed. In (Alvarez et al. 2004), the authors further proposed a language modeling approach that incorporated word pairs, without a constraint on adjacency or word order, where word pairs were determined by statistical relationships, or lexical affinities, between words. In addition,

Pickens (Pickens 2000) introduced an approach that used non-adjacent biterms, but the particular domain, musical documents, required an emphasis on the order of "words". In particular, our approach differs from the previous studies where we propose the concept of a pseudo term, Cross Term, generated by multiple query terms, to investigate how multiple query terms' occurrence change together. Cross Terms are naturally integrated into basic retrieval models as new terms, and therefore incorporate proximity into the retrieval process.

N-gram term associations have been investigated in IR for years (Ahmed and Nurnberger 2009, Bendersky and Croft 2012, Hou et al. 2013, Mayfield and McNamee 2003, McNamee and Mayfield 2004). Hawking and Thistlewaite (Hawking and Thistlewaite 1995) evaluated "Span" proximity approaches on TREC data sets, which is the text segments containing all query term instances. (Gao et al. 2004) introduced the linkage of a query as a hidden variable, which expressed the term associations within the query as an acyclic, planar, undirected graph. In (Beigbeder and Mercier 2005), the authors proposed a mathematical model based on a fuzzy proximity degree of term occurrences particularly for boolean queries. (Bendersky and Croft 2008) extracted key concepts from long and verbose queries. (Guo et al. 2008) proposed a conditional random field model to predict a sequence of refined query terms for a given sequence of ill-formed query terms. (Bendersky et al. 2009)

showed that query segmentation could reduce query latency without compromising effectiveness. (Park et al. 2011) presented a term dependency model, which was inspired by a quasi-synchronous stochastic process for machine translation. A retrieval model based on Markov random field (Metzler and Croft 2005) was presented for developing a general retrieval framework for modeling bigram and n-gram term associations. In (Bendersky et al. 2010), the Markov random field model was further extended by assigning weights to concepts.

2.3 Summary

The Okapi BM25 and the Language Model are the most investigated models nowadays. New models will be introduced in this thesis by integrating term association into the existing weighting models. In particular, this thesis focuses on boosting the performance of BM25 and Language Model.

3 Bigram Cross Term Retrieval Model

We first introduce a new concept, Cross Term, which is not an actual existing term but a pseudo term. The association among query terms is formalized as a Cross Term. To have a better understanding of Cross Term, we start with a special case when two query terms are considered, namely bigram Cross Term. In particular, we formalize the notion of bigram Cross Term in Section 3.1, and illustrate how to compute the statistics of a bigram Cross Term (e.g term frequency, inverse document frequency, and the number of documents that containing a Cross Term) in Section 3.2. Section 3.3 introduces the kernel functions used to calculate the Cross Term statistics. The BM25-based bigram CRoss Term Retrieval model $CRTER_2$ and LM-based model $CRTER_2^{LM}$ are proposed in Section 3.4. The experimental settings and models for comparison are shown in Section 3.5, and the experimental results are presented and discussed in Section 3.6. Further discusses and analysis are made in Section 3.7.

3.1 A New Pseudo Term: Bigram Cross Term

In this section, we formally define the notion of a bigram Cross Term and the method of computing a bigram Cross Term generated by two query terms. Suppose we have a query, $Q = \{q_1, q_1, ..., q_n\}$ and a document D, where pos is one of the positions of query term q_i in document D. The term q_i will influence the positions between $pos + k$ and $pos - k$. We use an impact function $f_i(pos, k)$ to capture the impact of a matching term q_i at position pos. In order to better describe this impact, $f_i(pos, k)$ should satisfy the following properties.

Property 3.1.1. *Let $f_i(pos, k)$ be the impact function of a query term q_i at position $pos + k$. We assume that $f_i(pos, k)$ follows 5 properties.*

1. *Non-negative: $f_i(pos, k) > 0$, the impact of a term towards its neighborhood is always non-negative. We only consider the positive influence among query terms.*

2. *Continuous: $|f_i(pos, k) - f_i(pos, k+1)|$ is small, i.e., there is a slight difference between two neighboring positions.*

3. *Symmetric: $f_i(pos, -k) = f_i(pos, k)$, the term has the same impact towards two equal-distance positions.*

4. *Monotonic:* $f_i(pos, k) > f_i(pos, k+1)$*, where* $k > 0$*. The influence decreases with the increase of* $|k|$*.*

5. *Identity:* $f_i(pos, 0) = 1$*, set one as standard influence.*

The Non-negative property ensures that the closely occurred query terms only boost each other. At the current stage, we do not take into account the negative impact, although some words may have the negative influence, such as "no" and "without". In this thesis, we will not discuss this issue, which could be our future work. The Continuous property is intuitive that the impact changes gradually. In practice, the domain is the cartesian product of natural numbers and integers, i.e., *pos* is a natural number and k is an integer. The Symmetric property is a compromise for the lack of prior knowledge towards the query terms. The semantic meaning of words could be very complicated to simulate. Therefore, we simplify the individual meanings by assuming their impact functions are identical and symmetric. The Monotonic property and the Symmetric property ensure that a closer position always has higher impact. The identity property ensures that the impact function is normalized to the range of 0 to 1.

Gaussian Kernel, Circle Kernel, and Triangle Kernel are widely used kernel functions and satisfy all of the above properties (Lv and Zhai 2009). We will discuss more about various kernel functions in Section 3.3. Without previous domain knowl-

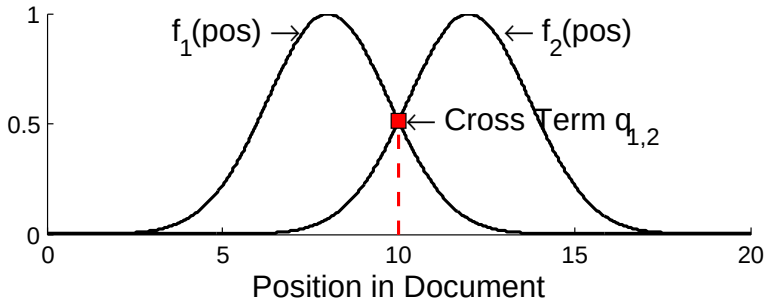


Figure 3.1: An example of bigram Cross Term

edge of a query and a collection, we assume that the query terms are identically distributed, i.e., the query terms have the same impact shape functions. When two query terms are close enough, their impact shape functions will join. According to the properties above, we can see that the point of intersection will have a higher value when two query terms occur closer, and the value of the intersection will be lower when two query terms are farther away. The point of intersection of the impact functions reflects the association between these two query terms. We create a new concept of bigram Cross Term to quantify the association between two query terms. This bigram Cross Term will be generated when two query terms' impact functions have an intersection.

Definition 1. *Given two query terms q_i and q_j, if there exists at least one point of intersection for impact functions $f_i(pos_1, k_1)$ and $f_j(pos_2, k_2)$, we say that a* ***bigram***

Cross Term *occurs, denoted as* $q_{i,j}$*. We call* q_i *and* q_j ***Generating Terms*** *of* $q_{i,j}$*.*

Definition 2. *When a bigram Cross Term* $q_{i,j}$ *occurs, the* ***bigram Cross Term's value*** *is the impact function's value at the intersection.*

A bigram Cross Term is always located in the middle of its generating query terms, and has a higher value when the query terms are closer to each other. Figure 3.1 shows an example of a bigram Cross Term. Two query terms, q_1 and q_2, locate at the 8th position and the 12th position in the document. More intuitively, we adopt the Gaussian Kernel as the impact shape function. Their impact's shape functions are located at $f_1(pos_1, 2)$ and $f_2(pos_2, -2)$, respectively in the document. We can see that two shape functions cross over each other, and there exists an intersection at the 10th position. There is no threshold incorporated into the definitions of bigram Cross Term and its value. For a continuous kernel function, there would always be an intersection if two terms occur in one document. For some kernel functions, its value naturally becomes 0 when a position is far enough from the center point.

3.2 Estimations for Bigram Cross Term Variants

Based on the previous definition of a Bigram Cross Term, we can see that it is impossible to evaluate it in the same way as a regular query term. In this Section, we propose to re-estimate the variants of a Bigram Cross Term. Several term-dependent

variants for this pseudo term are defined accordingly.

3.2.1 Within-Document Frequency Estimation of Bigram Cross Term

Here we will define the counting method for the frequency. The within-document frequency is the rate at which a term occurs in a document. For a single query term, its frequency in document D equals the number of times it occurs in D. There are two reasons that we need to redefine the within-document frequency for a bigram Cross Term. First, to simply count the occurrences of a Cross Term can not show the degree of the association between two generating query terms. For example, if two query terms q_i and q_j's impact functions have one point of intersection in both D_1 and D_2 with values 0.1 and 0.9 respectively, the occurrences of $q_{i,j}$ in both D_1 and D_2 are 1. However, we can see that there is a stronger association between q_i and q_j in D_2 than that in D_1. Second, the influence of $q_{i,j}$ is overemphasized, if we use the value of occurrence as Cross Term within-document frequency. For example, if two query terms q_i and q_j's impact functions in D have 5 points of intersections with values of $\{0.1, 0.2, 0.1, 0.3, 0.1\}$. Then the occurrences of $q_{i,j}$ in D equals 5. However, the influence of each occurrence of $q_{i,j}$ is lower than the influence of a query term.

We introduce a new estimation of a bigram Cross Term's within-document fre-

quency. Naturally we adopt the bigram Cross Term's value in estimating its within-document frequency. Suppose the positions of q_i in a document are $\{pos_{1,i}, pos_{2,i}, ..., pos_{tf_i,i}\}$, where tf_i is the term frequency of q_i. Correspondingly, the positions of q_j in the document are $\{pos_{1,j}, pos_{2,j}, ..., pos_{tf_j,j}\}$, where tf_j is the term frequency of q_j. Then the within-document term frequency of $q_{i,j}$ is defined as follows.

Definition 3. *The frequency of $q_{i,j}$ in D is the accumulation of $q_{i,j}$'s value.*

$$tf\left(q_{i,j}, D\right) = \sum_{k_1=1}^{tf_i} \sum_{k_2=1}^{tf_j} Kernel\left(\frac{1}{2} dist(pos_{k_1,i}, pos_{k_2,j})\right) \tag{3.1}$$

where tf is the term frequency of $q_{i,j}$ in D, $Kernel(\cdot)$ is the kernel function adopted in query term's impact function, and $dist(\cdot)$ is the distance between two positions

$$dist(pos_{k_1,i}, pos_{k_2,j}) = |pos_{k_1,i} - pos_{k_2,j}| \tag{3.2}$$

Please note that the frequency of a bigram Cross Term might not be an integer. Various kernel functions will be studied in Section 3.3.

3.2.2 Document frequencies Estimation of Bigram Cross Term

To evaluate the number of documents containing a bigram Cross Term $q_{i,j}$, it is not reasonable to simply count the documents in which $q_{i,j}$ occurs. The contribution

from a query term and a bigram Cross Term is different. For a query term q_i, an occurrence means its frequency accumulates 1, and the number of documents containing q_i is

$$nd(q_i) = \sum_{D \in Index} \mathbf{1}_{\{qi \in D\}}$$

where $\mathbf{1}_{\{qi \in D\}}$ is an indicator function, which is equal to 1 if $q_i \in D$ and is equal to 0 otherwise. On the other hand, an occurrence of a bigram Cross Term adds a value less than 1 to its frequency. A bigram Cross Term's value could be varying, and ranges from 0 to 1. For example, if a bigram Cross Term $q_{i,j}$ occurs only once in D with a value of 0.03 at this occurrence, its term frequency equals 0.03 according to formula (3.1). In this case, there is a very small amount of the association between q_i and q_j in D. Therefore, D is more likely to contribute a value less than 1 to $nd(q_{i,j})$. We accumulate the average value of $q_{i,j}$ on each document in $nd(q_{i,j})$ as shown in Definition 4.

Definition 4. *The number of documents containing a bigram Cross Term $q_{i,j}$, is the sum of $q_{i,j}$'s average value on each document, shown as follows*

$$nd(q_{i,j}) = \sum_{D \in Index, Occur(q_{i,j}, D) \neq 0} \frac{tf(q_{i,j}, D)}{Occur(q_{i,j}, D)} \tag{3.3}$$

where Occur is the number of occurrences of $q_{i,j}$, which is

$$Occur(q_{i,j}, D) = \sum_{k_1=1}^{tf_i} \sum_{k_2=1}^{tf_j} \mathbf{1}_{\{Kernel(\frac{1}{2}dist(pos_{k_1,i}, pos_{k_2,j}))\neq 0\}}$$

3.2.3 Within-Query Frequency Estimation of Bigram Cross Term

To evaluate a bigram Cross Term's within-query term frequency, we can track each position of q_i and q_j the same way as within-document frequencies by the sum of all possible intersections. Moreover, different from in documents, query terms distribute densely in a query. So we can assume that query terms are adjacent to each other and let $dist(q_i, q_j)$ =1.

Definition 5. *The within-query frequency of $q_{i,j}$ is*

$$\begin{aligned} qtf(q_{i,j}) &= Kernel(\frac{1}{2} \cdot dist(q_i, q_j)) \cdot min\{(qtf(q_i), qtf(q_j)\} \\ &= Kernel(\frac{1}{2}) \cdot min\{(qtf(q_i), qtf(q_j)\} \end{aligned} \tag{3.4}$$

where $qtf(q_i)$ and $qtf(q_j)$ are within query term frequencies of q_i and q_j, and Kernel is the same kernel function utilized in $tf(q_{i,j}, D)$.

3.3 Kernel Functions

Density functions based on proximity have been adopted to characterize term influence propagation (de Kretser and Moffat 1999, Kise et al., Lv and Zhai 2009, Mercier and Beigbeder 2005, Petkova and Croft 2007). (de Kretser and Moffat 1999) is early work that proposed to propagate the $tf \cdot idf$ score of each query term to other positions, where triangle, cosine, circle, and arc contribution functions were discussed. The highest accumulated $tf \cdot idf$ score on all the positions was adopted as the document's score. (Kise et al.) used hanning (cosine) window function to characterize the density of terms. Lv and Zhai (Lv and Zhai 2009) proposed a positional language model that incorporates the term proximity in a model based approach using four term propagation functions: Gaussian, Triangle, Cosine, and Circle. We employ the above term influence propagation functions for the concept of Cross Term in measuring term association, and introduce more potential functions: Quartic, Epanechnikov and Triweight.

The impact of a query term's occurrence is characterized by a kernel function. Here we present seven kernel functions that satisfy the query term impact function's properties (Property 3.1.1). Kernel functions have been studied in IR to characterize term propagation (de Kretser and Moffat 1999). We first apply four kernel functions that have been brought into IR applications. Among them, the Gaussian kernel

is widely used in statistics and machine learning algorithms such as the Support Vector Machines. Moreover, the Triangle kernel, Circle Kernel, and Cosine Kernel come from basic genomic graphics, which are applied to estimate the proximity-based density distribution for the positional language model (Lv and Zhai 2009). In addition, since it is difficult to determine which kernel functions can better simulate the terms' impacts, we investigate more kernel functions in this thesis. Among all the other kernel functions that satisfy the query term impact function's properties (Property 3.1.1), we introduce three most commonly used kernel functions: Quartic Kernel, Epanechnikov Kernel and Triweight Kernel to estimate the proximity-based terms' impacts.

- Gaussian Kernel:

$$Kernel(u) = exp[\frac{-u^2}{2\sigma^2}] \tag{3.5}$$

- Triangle Kernel:

$$Kernel(u) = (1 - \frac{u}{\sigma}) \cdot \mathbf{1}_{\{u \leq \sigma\}} \tag{3.6}$$

- Circle Kernel:

$$Kernel(u) = \sqrt{1 - (\frac{u}{\sigma})^2} \cdot \mathbf{1}_{\{u \leq \sigma\}} \tag{3.7}$$

- Cosine Kernel:

$$Kernel(u) = \frac{1}{2}[1 + cos(\frac{u\pi}{\sigma})] \cdot \mathbf{1}_{\{u \leq \sigma\}} \tag{3.8}$$

- Quartic Kernel:

$$Kernel(u) = (1 - (\frac{u}{\sigma})^2)^2 \cdot \mathbf{1}_{\{u \leq \sigma\}} \tag{3.9}$$

- Epanechnikov Kernel:

$$Kernel(u) = (1 - (\frac{u}{\sigma})^2) \cdot \mathbf{1}_{\{u \leq \sigma\}} \tag{3.10}$$

- Triweight Kernel:

$$Kernel(u) = (1 - (\frac{u}{\sigma})^2)^3 \cdot \mathbf{1}_{\{u \leq \sigma\}} \tag{3.11}$$

where σ is a normalization parameter, and $\mathbf{1}_{\{u \leq \sigma\}}$ is indicator function, which equal 1 if $u \leq \sigma$ and equals 0 otherwise.

Figure 3.2 shows an example of the included kernel functions. The curves represent the impact of a term occurring at position 10 with various kernel functions. The distance between a position and position 10 is the input u, and the parameter σ equals 6 in this example. We can see that most of the kernel functions have similar shapes with different gradients. They have positive values in the interval $u \in (-\sigma, \sigma)$, and a value of 0 otherwise. The parameter σ determines the range of the positions

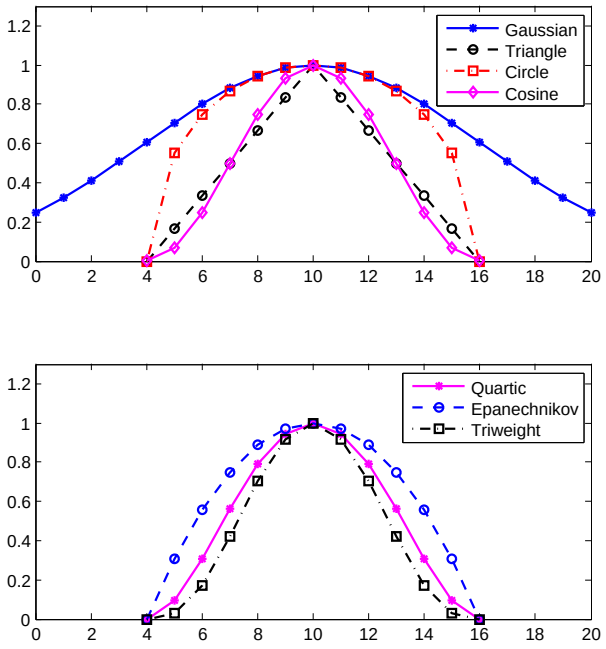


Figure 3.2: Kernel Functions

considered. A special kernel function is Gaussian kernel, which has positive values even when $u \notin (-\sigma, \sigma)$. With Gaussian kernel, the impact functions of the query terms will all meet in a document. When query terms are far away, the value at the intersection will be very small. The parameter σ for Gaussian kernel controls the extent of the impact considered. The performance of using all the kernel functions will be investigated in the experiments in Section 3.4.

3.4 Bigram CRoss-TErm Retrieval ($CRTER_2$) Model

We now propose Cross Term retrieval models based on the concept of Cross Term defined in Section 3.1. First, we adopt the BM25 probabilistic model as the basic weighting model in Section 3.4.1, and propose a bigram CRoss TErm Retrieval ($CRTER_2$) model as a basis case in Section 3.4.2. BM25 is a classical weighting function employed by the Okapi system (Robertson et al. 1996). As shown in previous TREC experiments, BM25 usually provides very effective retrieval performance on the TREC collections that are used in (Voorhees and Harman 2005, Ye et al. 2009).

The proposed approach can also extend the Language Model (Ponte and Croft 1998) in a similar way as BM25. Statistical Language Models have been used in many natural language processing applications. A language model is associated with a document, and the documents are ranked based on the probability of generating the query terms from the document's language model. We propose to extend the Language Model (LM) using Cross Terms and describe several corresponding well-known open source IR systems in Section 3.4.3.

3.4.1 Unigram Model: BM25

In BM25, the weight of a search term is assigned based on its within-document term frequency and query term frequency. The corresponding weighting function is as follows.

$$w(q_i, D) = \frac{(k_1 + 1) * tf(q_i, D)}{K + tf(q_i, D)} * \frac{(k_3 + 1) * qtf(q_i)}{k_3 + qtf(q_i)} * \log \frac{N - nd(q_i) + 0.5}{nd(q_i) + 0.5} \quad (3.12)$$

where $w(\cdot)$ is the weight of a query term q_i in a document. The variants in the above formula can be grouped into two categories as follows:

- The first category of variants is query independent. In this category, N is the number of indexed documents in the collection. k_1 and k_3 are tuning constants which depend on the dataset used and possibly on the nature of the queries. K equals $k_1 * ((1 - b) + b * dl/avdl)$, dl is the length of the document, and $avdl$ is the average document length.

- The values of the other group of variants change from term to term. In this category, $nd(q_i)$ is the number of documents containing a specific term. $tf(q_i, D)$ is the within-document term frequency. $qtf(q_i)$ is the within-query term frequency.

Finally, a document's weight for a query is given by the sum of its weights for all

terms in the query,

$$BM25(D) = \sum_{i=1}^{|Q|} w(q_i, D) \tag{3.13}$$

where w is the term weight obtained from Equation (5.8), and $|Q|$ is the length of the query Q.

3.4.2 BM25-based $CRTER_2$ Model

We propose a bigram Cross Term Retrieval ($CRTER_2$) model, utilizing the bigram Cross Term as a special term. In this model, a weighting score is calculated based on both the query terms and the bigram Cross Terms. The association among query terms is naturally imported into the weighting model by bigram Cross Terms. A new combined weighting model for a document is

$$CRTER_2(D) = (1 - \lambda_2) \cdot \sum_{1 \le i \le |Q|} w(q_i, D) + \lambda_2 \cdot \sum_{1 \le i < j \le |Q|} w_2(q_{i,j}, D) \tag{3.14}$$

where $w(\cdot)$ is the weighting function of query terms in the query Q, $w_2(\cdot)$ is the bigram Cross Term $q_{i,j}$'s weight, which has the following form.

$$w_2(q_{i,j}, D) = \frac{(k_1+1) * tf(q_{i,j}, D)}{K + tf(q_{i,j}, D)} * \frac{(k_3+1) * qtf(q_{i,j})}{k_3 + qtf(q_{i,j})} * \log \frac{N - nd(q_{i,j}) + 0.5}{nd(q_{i,j}) + 0.5} \tag{3.15}$$

where we replace the term dependent variants with the bigram Cross Term's variants defined in Equation (3.1), Equation (3.3), and Equation (3.4).

In $CRTER_2$ model, λ_2 is a parameter balancing the query terms and bigram Cross Terms. When λ_2 is equal to 0, the retrieval model uses query terms only, which is the standard BM25 weighting model. When λ_2 is equal to 1, the retrieval model uses bigram Cross Terms only. Since the weights of query terms and bigram Cross Terms are normalized independently, the value of λ_2 reflects the influence of using bigram Cross Terms.

3.4.3 Language Model based $CRTER_2^{LM}$ Model

In order to evaluate the effect of Cross Terms, we also compare with several state-of-the-art language models. To have a fair comparison, here we extend Cross Terms on the basis of the Language model. In a language model (Zhai and Lafferty 2001),

a document is ranked by

$$LM(D) = \sum_{1 \leq i \leq |Q|} log \frac{P(q_i|D)}{P(q_i|C)} \tag{3.16}$$

where $P(q_i|C) = \frac{cf(q_i)}{|C|}$ is the collection language model, C represents the collection, $|C|$ is the total number of terms in the collection, $cf(q_i)$ is the frequency of q_i over the collection, and $P(q_i|D)$ is the document language. Several approaches smooth $P(q_i|D)$ to solve the zero probability problem. Dirichlet smoothing is one of the most popular approaches and reported to have the best average precision in most of the cases (Zhai and Lafferty 2001, 2004). The conditional probability $P(q_i|D)$ is smoothed by

$$P(q_i|D) = \frac{tf(q_i, D) + \mu \frac{cf(q_i)}{|C|}}{|D| + \mu} \tag{3.17}$$

where μ is the Dirichlet smoothing parameter. μ is tuned to be optimal in our experiments.

A bigram language model is usually linearly combined to a unigram language model (Song and Croft 1999). The Language model with bigram Cross Term is

$$CRTER_2^{LM}(D) = (1 - \lambda_2^{LM}) \cdot LM(D) + \lambda_2^{LM} \cdot \sum_{1 \leq i < j \leq |Q|} log \frac{P(q_{i,j}|D)}{P(q_{i,j}|C)} \tag{3.18}$$

where $P(q_{i,j}|D) = \frac{tf(q_{i,j},D)+\mu\frac{cf(q_{i,j})}{|C|}}{|D|+\mu}$ and $P(q_{i,j}|C) = \frac{cf(q_{i,j})}{|C|}$. $tf(q_{i,j}, D)$ is defined in formula (3.1), and $cf(q_{i,j})$ is $q_{i,j}$'s frequency over the collection which can be derived by accumulating $tf(q_{i,j})$ over the collection

$$cf(q_{i,j}) = \sum_{D \in C} tf(q_{i,j}, D)$$

We naturally further extend the language model with n-gram Cross Terms. The weighting function is recursively defined as

$$CRTER_n^{LM}(D) = (1 - \lambda_n^{LM}) \cdot CRTER_{n-1}^{LM}(D) + \lambda_n^{LM} \cdot \sum_{1 \le i_1 < i_2 < \ldots < i_n \le K} \frac{P(q_{i_1,i_2,\ldots,i_n}|D)}{P(q_{i_1,i_2,\ldots,i_n}|C)} \tag{3.19}$$

where $P(q_{i_1,i_2,\ldots,i_n}|D) = \frac{tf(q_{i_1,i_2,\ldots,i_n},D)+\mu\frac{cf(q_{i_1,i_2,\ldots,i_n})}{|C|}}{|D|+\mu}$, $P(q_{i_1,i_2,\ldots,i_n}|C) = \frac{cf(q_{i_1,i_2,\ldots,i_n})}{|C|}$, $tf(q_{i_1,i_2,\ldots,i_n}, D)$ is defined in formula (4.1), and $cf(q_{i_1,i_2,\ldots,i_n})$ is $q_{i_1,i_2,\ldots,i_n}$'s frequency over the collection which can be derived by accumulating $tf(q_{i,j})$ over the collection

$$cf(q_{i_1,i_2,\ldots,i_n}) = \sum_{D \in C} tf(q_{i_1,i_2,\ldots,i_n}, D)$$

.

In the rest of this thesis, we use the notation $CRTER_2^{LM}$ to represent the Lan-

guage Model with bi-gram Cross Terms. For implementation, there are several well-known open source IR systems supporting the Language Model. For instance, the Lemur project (Lavrenko and Croft 2001, Strohman et al. 2005) develops the Lemur Toolkit and the Indri search engine which combines inference nets and language modelling in an architecture designed for large-scale applications. The Terrier search engine (Ounis et al. 2006a, 2007) implements indexing and retrieval functionalities, combines ideas from probabilistic theory, statistical analysis, and data compression techniques. $CRTER_2^{LM}$ can be implemented on any of these IR systems. The corresponding settings will be discussed in Section 3.6.5.

3.5 Experimental Settings

Here we first present the data sets and the related evaluation measures used in our experiments in Section 3.5.1. Second, we describe our Okapi experimental platform in Section 3.5.3. Finally, we describe four major probabilistic and language models for comparison in Section 3.5.4.

3.5.1 Data Sets and Evaluation Measures

3.5.1.1 TREC Data Sets

We present seven standard TREC collections used in our experiments, the statistics of which are shown in Table 3.1. These collections are diverse in both sizes and content, which facilitate a thorough evaluation of our proposed models. TREC8 contains newswire articles from various sources, such as Financial Times (FT), the Federal Register (FR) etc., which are usually considered as high-quality text data with little noise. Robust has the same data collection as TREC8, but Robust includes more topics (301-450 and 601-700). We use both Robust and TREC8 to test the retrieval models on the same collection with different topic sets in Section 3.7.3. AP88-89 contains articles published by Association Press from the year of 1988 to 1989. The WT2G collection is a 2G size crawl of Web documents. The WT10G collection is a medium size crawl of Web documents, which was used in the TREC9 and TREC10 Web tracks. It contains 10 Gigabytes of uncompressed data. The .GOV2 collection, which has 426 Gigabytes of uncompressed data, is a crawl from the .gov domain. This collection has been employed in the TREC14 (2004), TREC15 (2005) and TREC16 (2006) Terabyte tracks. The Blog06 collection includes 100,649 blog feeds collected over an 11 week period from December 2005 to February 2006. Following the official

TREC settings (Ounis et al. 2006b), we index only the permalinks, which are the blog posts and their associated comments. For all test collections used, each term is stemmed using Porter's English stemmer, and standard English stopwords are removed.

Collection Name	# of Docs	Topics	# of Topics
TREC8	528,155	401-450	50
Robust	528,155	301-450 & 601-700	250
AP88-89	164,597	51-100	50
WT2G	247,491	401-450	50
WT10G	1,692,096	451-550	100
.GOV2	25,178,548	701-850	150
Blog06	3,215,171	851-950 & 1000-1050	150

Table 3.1: Overview of the TREC collections used

3.5.1.2 Topic Processing and Cross Validation

A topic usually contains three topic fields, namely title, description and narrative. Figure 3.3 shows an example of a standard TREC topic. We only use the title topic field that contains very few keywords related to the topic. The title-only queries are usually short which is a realistic snapshot of real user queries in practice. On each collection, we evaluate our proposed model by a 10-fold cross-validation. The test topics associated to each collection are randomly split into ten equal subsets. In each fold, 9 subsets of the test topics are used for training, and the remaining subset is used for testing. The overall retrieval performance is averaged over all 10 test subsets

<top>
<num> Number: 725
<title> Low white blood cell count
<desc> Description:
What would cause a lowered white blood cell count?
<narr> Narrative:
A relevant document will describe a condition or disease that causes a lowered white blood cell count. Lowered white blood cell counts caused by HIV infection, bone marrow failure and chemotherapy are relevant. A low count caused by a treatment or medication would also be relevant.
</top>

Figure 3.3: An example of a standard TREC topic

of topics.

3.5.2 Evaluation Metrics

We use the TREC official evaluation measures in our experiments, namely the topical MAP on Blog06 (Ounis et al. 2006b), and the Mean Average Precision (MAP) on the other six collections (Voorhees and Harman 2005). To put emphasis on the top retrieved documents, we also include P@5 and P@20 in the evaluation measures. All statistical tests are based on two-tailed Wilcoxon Matched-pairs Signed-rank test with a significance level of 0.05.

3.5.3 The Okapi System

In our experiments, we use Okapi BSS (Basic Search System) (Robertson and Walker 1994) as our main search system and conduct our information retrieval experiments using the improved Okapi system (Fan et al. 2006, Huang et al. 2013, Huang and Hu 2009, Huang et al. 2006a,b, 2005, Miao et al. 2012, Yin et al. 2013). Okapi is an information retrieval system based on the probability model of Robertson and Sparck Jones (Beaulieu et al. 1997, Robertson and Walker 1994), which is one of the most established and best-performing systems in information retrieval research and experimentation (Voorhees and Harman 2005). The retrieval documents are ranked in the order of their probabilities of relevance to the query. A search term is assigned weight based on its within-document term frequency and query term frequency. The weighting function used is BM25 (see Section 3.4.1). In our experiments, the values of k_1, k_3 in formula (5.8) are set to the defaults 1.2 and 8, respectively, which is the recommended setting in (Robertson et al. 1996). The parameter b is set to 0.35, which is shown to be optimal in our preliminary experiments (see Figure 3.8).

3.5.4 Major Probabilistic and Language Models for Comparison

Here we briefly describe four proximity models used for comparison. We choose two probabilistic BM25 based proximity models, namely PPM (Song et al. 2011)

and BM25TP (Büttcher et al. 2006), as well as two language model (LM) based term dependency models, namely MRF (Metzler and Croft 2005) and PLM (Lv and Zhai 2009). PPM and PLM are based on the same assumptions as our proposed $CRTER$ model, and use kernel functions to simulate the term influence's propagation. BM25TP is a probabilistic BM25-based proximity model with good performance, and it is parameter free. MRF is a one of the most famous LM-based dependency models, and was widely investigated by many IR researchers. In experiments, we compare with these four models.

- Proximity Probabilistic Model (PPM)

  In PPM, a position dependent term count is related to both the number of occurrences of a term and the term counts propagated from other terms. Each query term has a pseudo term frequency, which is the original query term frequency added by accumulating term counts propagated from the nearest occurrences of the other terms.

- BM25TP

  A proximity accumulator is associated with each query term in BM25PT. In a document, once there is a posting of a query term, both the current query term's and the previous query term's accumulators are incremented (if they

are different query terms).

- Markov Random Fields (MRF)

  MRF uses Markov Random Fields to model the joint distribution over queries and documents. A graph is constructed to represent the query dependencies to the document, and a set of potential functions over the cliques of this graph are defined. The potential functions are expected to satisfy $\psi(q_{i_1}, .., q_{i_n}, D) > \psi(q'_{j_1}, .., q'_{j_n}, D)$ when $q_{i_1}, .., q_{i_n}$ are much more compatible with document D than the terms $q'_{j_1}, .., q'_{j_n}$. Three variants are considered and linearly combined in MRF: full independency (FI) is a smoothed unigram language model estimate; sequential dependence (SD) estimates all subphrases; full dependence (FD) is a proximity-based estimate which is implemented with the co-occurrences of two or more query terms within a window.

- Positional Language Model (PLM)

  PLM is also a kernel based approach, that a term at each position can propagate its occurrence at that position to other positions. The PLM at each position can then be estimated based on all the propagated counts of all the words. A language model is defined for each position of a document.

3.6 Experiments

Here we conduct a series of experiments for our proposed models $CRTER_2$ and $CRTER_n$ respectively. For $CRTER_2$, we implement it with either BM25 or Dirichlet LM as the basis model. In order to distinguish them, we denote $CRTER_2^{LM}$ as the LM based $CRTER_2$ model, and $CRTER_2$ as the BM25 based $CRTER_2$ model. We first present the results, investigate the impact of important parameters and analyze the robustness of the $CRTER_2$ model in Section 3.6.1, 3.6.2, and 3.6.3. Then we compare $CRTER_2$ with two state-of-the-art probabilistic proximity approaches in Section 3.6.4, and compare $CRTER_2^{LM}$ with two state-of-the-art LM proximity approaches in Section 3.6.5.

3.6.1 Experimental Results of $CRTER_2$

In our evaluation, we first investigate the performance of our proposed bigram Cross Term Retrieval ($CRTER_2$) model compared to the basic weighting model BM25. Specifically, we use BM25 with optimal settings as our baseline. The parameter b is set to 0.35, which is shown to be optimal in our preliminary experiments (see Figure 3.8). The related experimental results are presented in Table 3.2. Seven different kernel functions are applied to instantiate the $CRTER_2$ model, including: Gaussian, Triangle, Circle, Cosine, Quartic, Epanechnikov, and Triweight kernels. All

	Eval Metric	TREC8	AP88-89	WT2G	WT10G	.GOV2	Blog06
BM25	MAP	0.2561	0.2710	0.3156	0.2119	0.3039	0.3246
	P@5	0.4920	0.4360	0.5280	0.3800	0.6134	0.6400
	P@20	0.4000	0.3860	0.3930	0.2670	0.5426	0.5997
$CRTER_2$ Gaussian	MAP	0.2604 (+1.679%)	0.2787 (+2.841%)	0.3354* (+6.274%)	0.2213* (+4.436%)	**0.3342*** (+9.970%)	0.3505* (+7.979%)
	P@5	0.5040 (+2.439%)	0.4520 (+3.670%)	**0.5480** (+3.788%)	**0.4080*** (+7.368%)	**0.6550*** (+6.782%)	0.6560 (+2.500%)
	P@20	0.4190 (+4.750%)	0.3900 (+1.036%)	0.4070 (+3.562%)	0.2775 (+3.933%)	0.5809* (+7.059%)	0.6260* (+4.386%)
$CRTER_2$ Triangle	MAP	**0.2606** (+1.757%)	**0.2789** (+2.915%)	**0.3359*** (+ 6.432%)	0.2207* (+4.153%)	0.3339* (+9.871%)	0.3512* (+8.195%)
	P@5	0.5040 (+2.439%)	0.4520 (+3.670%)	**0.5480** (+3.788%)	**0.4080*** (+7.368%)	0.6537* (+6.570%)	**0.6733*** (+5.203%)
	P@20	0.4190 (+4.750%)	0.3890 (+0.777%)	**0.4100*** (+4.326%)	0.2775 (+3.933%)	0.5792* (+6.745%)	0.6340 (+5.720%)
$CRTER_2$ Circle	MAP	0.2599 (+1.484%)	0.2783 (+2.694%)	**0.3359*** (+ 6.432%)	**0.2227*** (+5.97%)	0.3331* (+9.608%)	0.3515* (+8.287%)
	P@5	0.5040 (+2.439%)	**0.4600*** (+5.505%)	0.5440 (+ 3.030%)	0.4060* (+6.842%)	0.6523* (+ 6.342%)	0.6653 (+3.953%)
	P@20	0.4190 (+4.750%)	0.3890 (+0.777%)	0.4080* (+ 3.817%)	**0.2785*** (+4.307%)	0.5782* (+6.561%)	**0.6357*** (+6.003%)
$CRTER_2$ Cosine	MAP	0.2599 (+1.484%)	**0.2789** (+2.915%)	0.3358* (+ 6.401%)	0.2216* (+4.578%)	0.3339* (+9.872%)	0.3515* (+8.287%)
	P@5	0.5040 (+2.439%)	0.4560 (+4.587%)	0.5440 (+3.030 %)	**0.4080*** (+7.368%)	0.6537* (+6.570%)	**0.6733*** (+5.203%)
	P@20	0.4190 (+4.750%)	**0.3910** (+1.295%)	0.4090 (+ 4.071%)	0.2765 (+3.558%)	**0.5812*** (+7.114%)	0.6343* (+5.770%)
$CRTER_2$ Quartic	MAP	0.2599 (+1.484%)	0.2787 (+2.841%)	0.3352* (+6.210%)	0.2212* (+4.389%)	0.3338* (+9.839%)	**0.3517*** (+8.349%)
	P@5	0.5040 (+2.439%)	0.4560 (+4.587%)	0.5440 (+ 3.030%)	**0.4080*** (+7.368%)	0.6537* (+6.570%)	**0.6733*** (+5.203%)
	P@20	0.4170 (+4.250%)	0.3920 (+1.554%)	**0.4100*** (+ 4.326%)	0.2780* (+4.120%)	0.5805* (+6.985%)	0.6343* (+5.770%)
$CRTER_2$ Epanechnikov	MAP	0.2602 (+1.601%)	0.2787 (+2.841%)	0.3343* (+ 5.925%)	0.2217* (+4.625%)	0.3330* (+9.576%)	0.3512* (+8.195%)
	P@5	**0.5080** (+3.252%)	0.4520 (+3.670%)	0.5440 (+ 3.030%)	**0.4080*** (+7.368%)	0.6523* (+6.342%)	0.6720* (+5.000%)
	P@20	**0.4200*** (+5.000%)	0.3890 (+0.777%)	0.4070 (+ 3.562%)	0.2785* (+4.307%)	0.5802* (+6.930%)	0.6340* (+5.720%)
$CRTER_2$ Triweight	MAP	0.2601 (+1.562%)	0.2788 (+2.878%)	0.3356* (+ 6.337%)	0.2225* (+5.002%)	0.3341* (+9.937%)	**0.3517*** (+8.349%)
	P@5	**0.5080** (+3.252%)	**0.4600*** (+5.505%)	0.5440 (+ 3.030%)	0.4060* (+6.842%)	**0.6550*** (+6.782%)	0.6720* (+5.000%)
	P@20	0.4180 (+4.500%)	**0.3910** (+1.295%)	0.4070 (+ 3.562%)	0.2780* (+4.120%)	0.5809* (+7.059%)	0.6350* (+5.886%)

Table 3.2: Comparison between BM25 baseline and $CRTER_2$ with different kernel functions: BM25 parameter b is initialized to be 0.35. All the results are evaluated by MAP, P@5, and P@20. $CRTER_2$ outperforms BM25 on all collections. "*" means the improvements over the BM25 are statistically significant ($p<0.05$ with Wilcoxon Matched-pairs Signed-rank test).

the results are evaluated by MAP, P@5, and P@20. The percentage of how much $CRTER_2$ outperforms BM25 is also listed. The best result obtained on each collection is marked bold. The experiments on .Gov2 and Blog06 are more comprehensive than in (Zhao et al. 2011). As shown by the results, our proposed $CRTER_2$ model outperforms BM25 on all the six collections used. The advantage of $CRTER_2$ over BM25 is especially evident on the relatively larger collections, WT10G, .GOV2 and Blog06, where statistically significant improvements are observed with all 7 kernel functions used. Moreover, according to the results in Table 3.2, each kernel function has its advantage on some aspects. There is no single kernel function that can outperform all the other kernel functions on all the datasets. Figure 3.4 shows the winnings of the kernel functions on different datasets under different evaluation metrics, which further illustrates Table 3.2. We can see that the triangle kernel performs the best (wins 7 times) in terms of winning times.

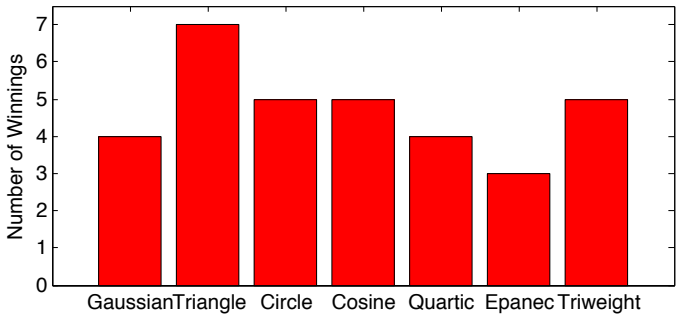
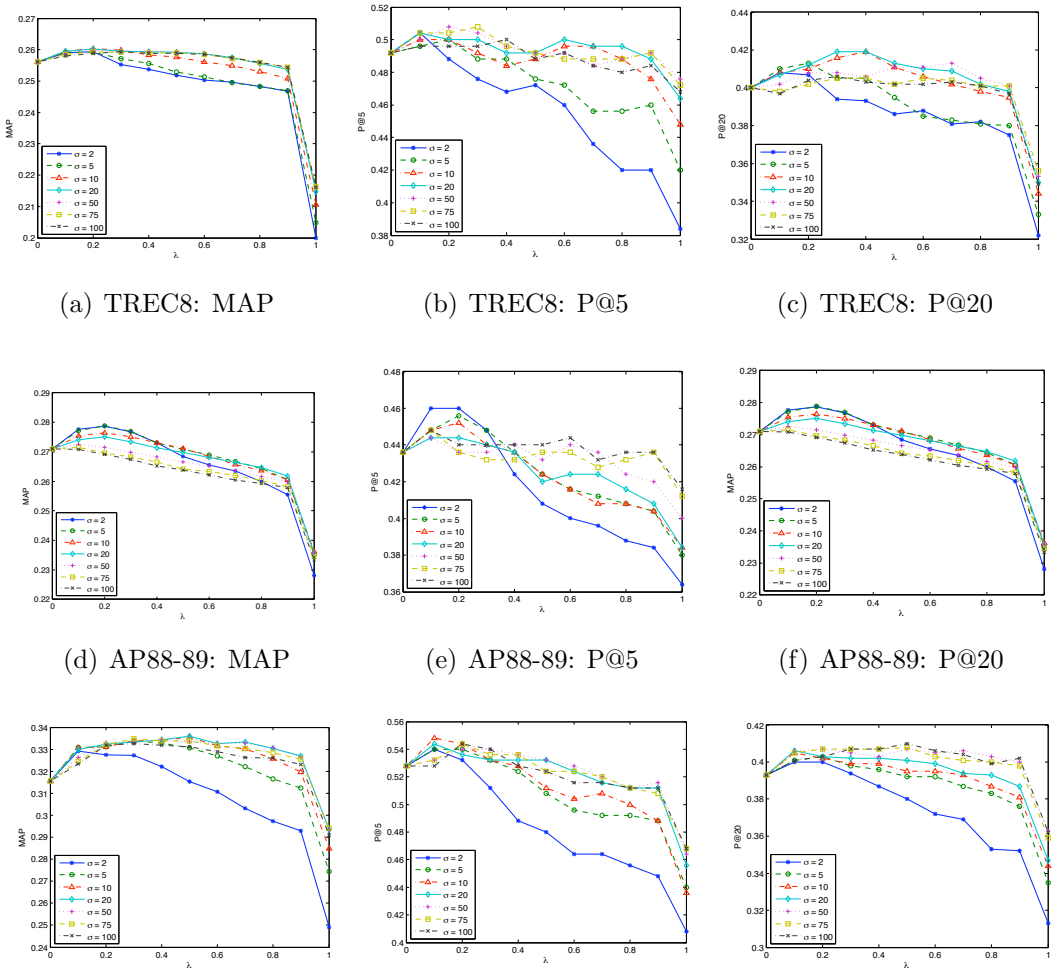


Figure 3.4: Comparison among kernel functions



(a) TREC8: MAP (b) TREC8: P@5 (c) TREC8: P@20

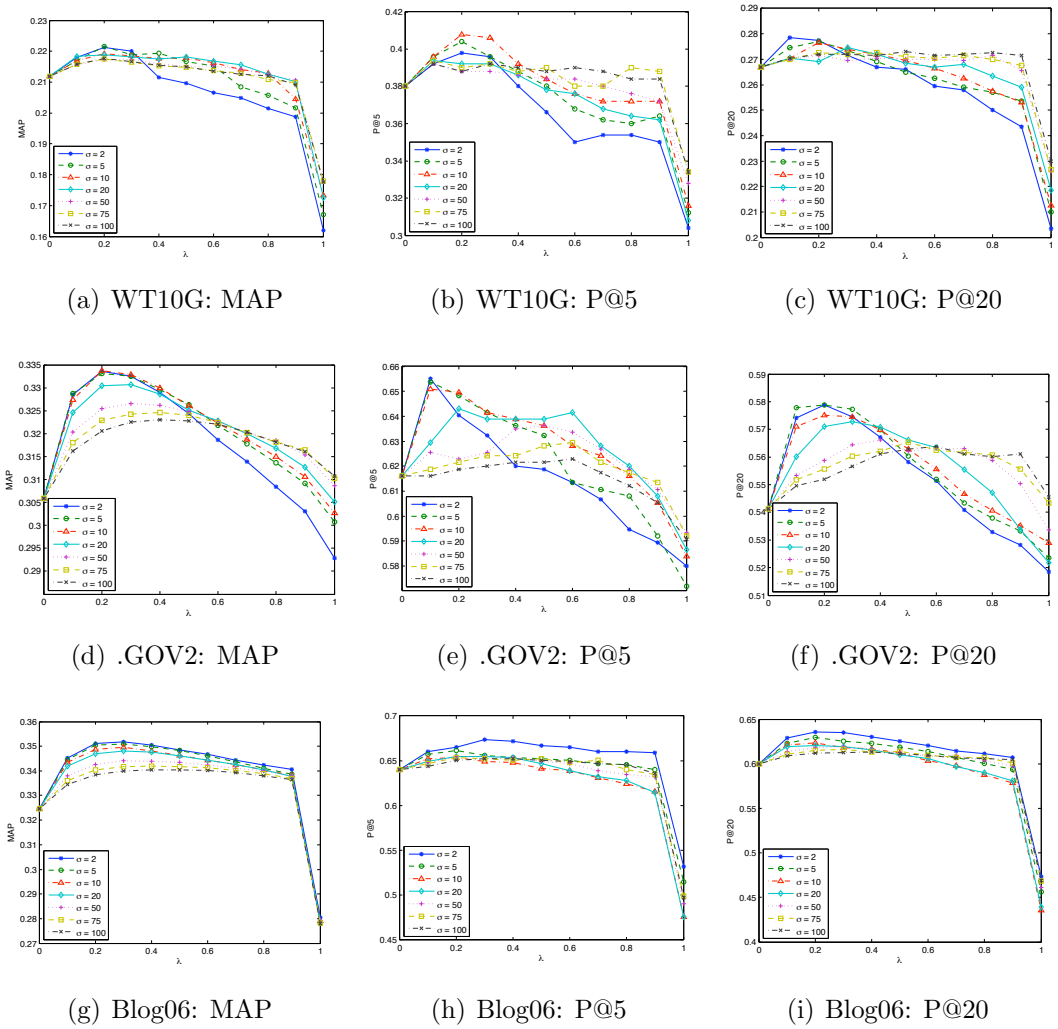
(d) AP88-89: MAP (e) AP88-89: P@5 (f) AP88-89: P@20

(g) WT2G: MAP (h) WT2G: P@5 (i) WT2G: P@20

Figure 3.5: Sensitivity to $CRTER_2$ parameter λ with different kernel parameters on TREC8, AP88-89, and WT2G

3.6.2 Parameter Sensitivity

An important issue that may affect the robustness of the $CRTER_2$ model is the sensitivity of its parameters λ (in Equation 4.11) and σ (in Equation 4-10) to retrieval



(a) WT10G: MAP (b) WT10G: P@5 (c) WT10G: P@20

(d) .GOV2: MAP (e) .GOV2: P@5 (f) .GOV2: P@20

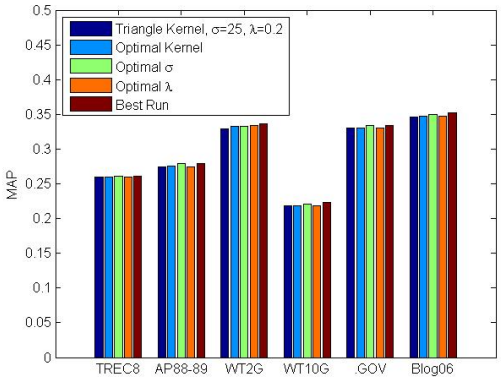
(g) Blog06: MAP (h) Blog06: P@5 (i) Blog06: P@20

Figure 3.6: Sensitivity to $CRTER_2$ parameter λ with different kernel parameters on WT10G, .GOV, and Blog06

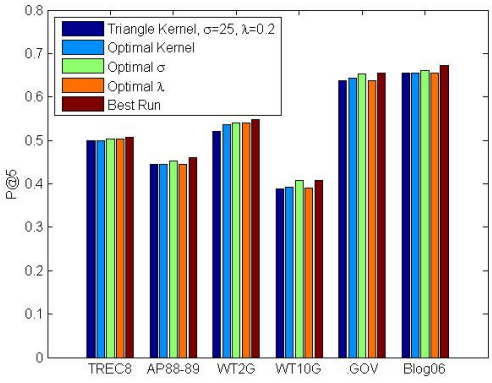
performance. The parameter λ balances the influence of the query terms and the bigram Cross Terms. When λ is equal to 0, the retrieval model uses query terms only, which is the standard BM25 weighting model. When λ is equal to 1, the retrieval model uses bigram Cross Terms only. Since the weights of query terms and bigram Cross Terms are normalized independently, the value of λ reflects the influence of using bigram Cross Terms. The kernel parameter σ controls the range of a query term's impact. When σ is small, a bigram Cross Term occurs only if its generating term is very close. When σ is large, query terms far away from each other can generate a bigram Cross Term. But the bigram Cross Term's value will be different according to the distance between query terms.

Figure 3.5 and Figure 3.6 plot the evaluation metrics MAP, P@5, and P@20 obtained by $CRTER_2$ over λ values ranging from 0 to 1 on all the data sets. In addition, a group of different settings of σ are applied, namely $\sigma = 2, 5, 10, 20, 50, 75, 100$. The general tendency on each evaluation metric is similar. As we can see from the above two figures, $CRTER_2$'s retrieval performance decreases with large λ values. $CRTER_2$ generally performs well over different datasets when λ falls between 0 and 0.2. Overall, a λ value between 0 and 0.2 is recommended as it is shown to be reliable. In addition, σ is recommended to be between 3 and 25, considering both effectiveness and efficiency. The number of terms between 3 and 25 is consistent with

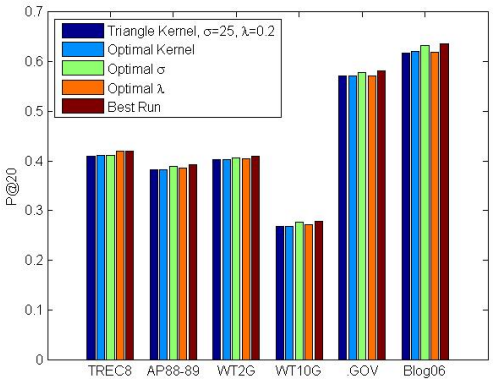
the length of a sentence.



(a) MAP



(b) P@5



(c) P@20

Figure 3.7: Robustness of $CRTER_2$: Compare $CRTER_2$'s retrieval performance provided by an empirical setting, namely Triangle Kernel, $\sigma = 25$, and $\lambda = 0.2$ with the following optimization strategies. First, optimize the kernel function, σ or λ individually while setting the other parameters to the empirical values. Second, optimize all the three parameters: kernel functions, σ and λ together.

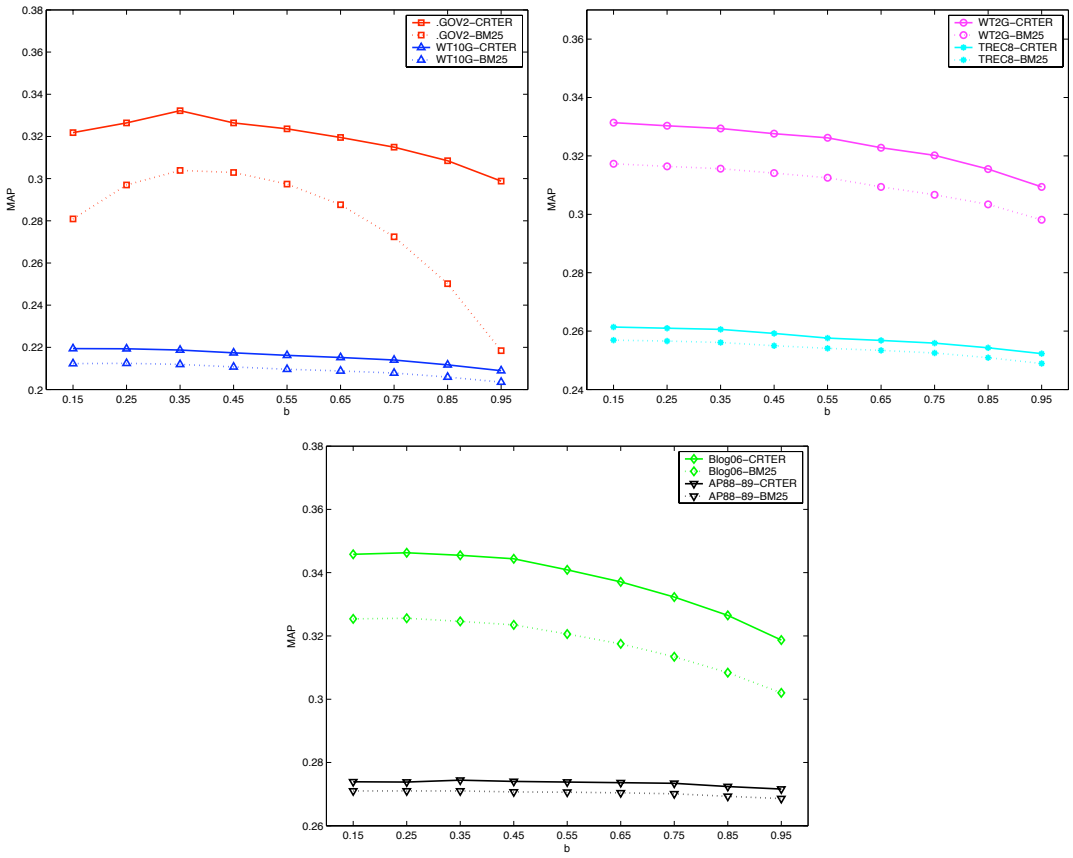


Figure 3.8: Generalized Performance of $CRTER_2$: Compare MAP between $CRTER_2$ and BM25 with the change of b ($CRTER_2$ uses fixed parameter: Triangle Kernel, $\sigma = 25$, $\lambda = 0.2$)

3.6.3 Robustness of $CRTER_2$

The proposed $CRTER_2$ model's robustness is important to its applications in practice. Ideally, we would like to have reliable retrieval performance using $CRTER_2$ on various datasets with its parameters within a stable safe range. This issue is

particularly crucial for a given new dataset without training data.

We first fix BM25's parameter b = 0.35, and evaluate the robustness of $CRTER_2$ provided by an empirical setting, obtained from previous observations, namely Triangle Kernel, $\sigma = 25$, and $\lambda = 0.2$. We compare this empirical setting with the following optimization strategies. First, optimize the kernel function, σ or λ individually while setting the other parameters to the empirical values. Second, optimize all the three parameters: kernel functions, σ and λ together. From Figure 3.7, we can see that the performance obtained by the empirical setting is comparable to the retrieval performance obtained by the optimized parameter settings on all datasets used.

We also test $CRTER_2$'s performance under the same empirical setting with BM25 over different BM25 parameters on all the six collections. In BM25, b ranges from 0.15 to 0.95. In Figure 3.8, MAP of BM25 changes over different b values, and $CRTER_2$ can boost BM25 under all b's settings and over all the collections. More specifically, for .GOV2 dataset, we can see that BM25's performance is very sensitive to b, its MAP gradually increases with the increment of b at the beginning, and sharply decreases later. $CRTER_2$, on the other hand, boosts basic BM25 over different b values, and tends to stabilize the retrieval performance.

In general, $CRTER_2$ performs robustly and has strong generalized performance.

Without much knowledge of a new dataset, a group of parameters are recommended for $CRTER_2$: Triangle Kernel, $\sigma = 25$, and $\lambda = 0.2$.

3.6.4 Comparison with Major Probabilistic Proximity Models

	Eval Metric	TREC8	AP88-89	WT2G	WT10G	.GOV2	Blog06
BM25	MAP	0.2561	0.2710	0.3156	0.2119	0.3039	0.3246
	P@5	0.4920	0.4360	0.5280	0.3800	0.6134	0.6400
	P@20	0.4000	0.3860	0.3930	0.2670	0.5426	0.5997
PPM	MAP	0.2604* (+1.679%)	0.2763 (+1.956%)	0.3230 (+2.345%)	0.2213* (+4.436%)	0.3166* (+4.179%)	0.3403* (+4.837%)
	P@5	**0.5200** (+5.691%)	0.4360 (+0.000%)	0.5280 (+0.000%)	0.3880 (+2.105%)	0.6188 (+0.880%)	0.6400 (+0.000%)
	P@20	0.4070 (+1.750%)	0.3860 (+0.000%)	0.3990 (+1.527%)	0.2755 (+3.184%)	0.5584* (+2.912%)	0.6133* (+2.268%)
$BM25TP$	MAP	0.2582 (+0.820%)	0.2781 (+2.620%)	0.3276 (+3.802%)	0.2161 (+1.982%)	**0.3346*** (+10.102%)	0.3419 (+5.330%)
	P@5	0.4920 (+0.000%)	0.4440 (+1.835%)	0.5120 (-3.030%)	0.3700 (-2.632%)	0.6443 (+5.037%)	0.6507 (+1.672%)
	P@20	0.3960 (-1.000%)	0.3870 (+0.259%)	0.3960 (+0.763%)	0.2750 (+2.996%)	**0.5852*** (+7.851%)	0.6147 (+2.501%)
$CRTER_2$	MAP	**0.2606** (+1.757%)	**0.2789** (+2.915%)	**0.3359*** (+6.432%)	**0.2227*** (+5.097%)	0.3342*† (+9.970%)	**0.3517***†‡ (+8.349%)
	P@5	0.5080 (+3.252%)	**0.4600*** (+5.505%)	**0.5480**‡ (+3.788%)	**0.4080***†‡ (+7.368%)	**0.6550***† (+6.782%)	**0.6733***† (+5.203%)
	P@20	**0.4200***‡ (+5.000%)	**0.3910** (+1.295%)	**0.4100*** (+4.326%)	**0.2785*** (+4.307%)	0.5812*† (+7.114%)	**0.6357***†‡ (+6.003%)

Table 3.3: Comparison among three BM25 based proximity models: PPM, $BM25TP$ and $CRTER_2$. BM25 parameter b is initialized to be 0.35. All the results are evaluated in terms of MAP, P@5, and P@20. "*" means the improvements over BM25 are statistically significant; "†" means the improvement over PPM is significant; and "‡" means the improvement over $BM25TP$ is significant ($p<0.05$)

In order to further evaluate our proposed model, we study how the proposed $CRTER_2$ model performs compared to the state-of-the-art probabilistic proximity approaches. We first compare the proposed $CRTER_2$ model with two probabilistic BM25-based proximity models, namely the Proximity Probabilistic Model (PPM)

(Song et al. 2011) and $BM25TP$ (Büttcher et al. 2006) respectively.

In order to establish a fair comparison, we implement PPM and $BM25TP$ with the same pre-processing techniques and the same optimal BM25 parameter settings. For PPM, we use the same kernel functions in (Song et al. 2011) and tune parameters to be optimal by using cross validation. For $BM25TP$, we follow the same experimental procedures reported in (Büttcher et al. 2006). We test $CRTER_2$, PPM and $BM25TP$ on the same collections with the same queries. The experimental results are shown in Table 3.3. In general, we can see that our proposed $CRTER_2$ shows improvements on all six data collections, and it is at least comparable to, if not better than, PPM and $BM25TP$. To illustrate the results in Table 3.3 graphically, we re-plot the improvement rates of these three models over BM25 in Figure 3.9, in which the x-axis shows the collections and the y-axis is the improvement rate. 0% means there is no improvement over BM25. We can find that all three models improve BM25 in terms of evaluation metrics MAP and P@20. From Figure 3.9, we can see that $CRTER_2$ provides stable performance for all the evaluation metrics. In most of the cases, $CRTER_2$ outperforms PPM and $BM25TP$, especially on the top ranked documents (e.g. top 5 and 20). It could be explained by $CRTER_2$'s capability of integrating more information into the retrieval process, and considering the collection information (e.g. the number of documents containing a Cross Term).

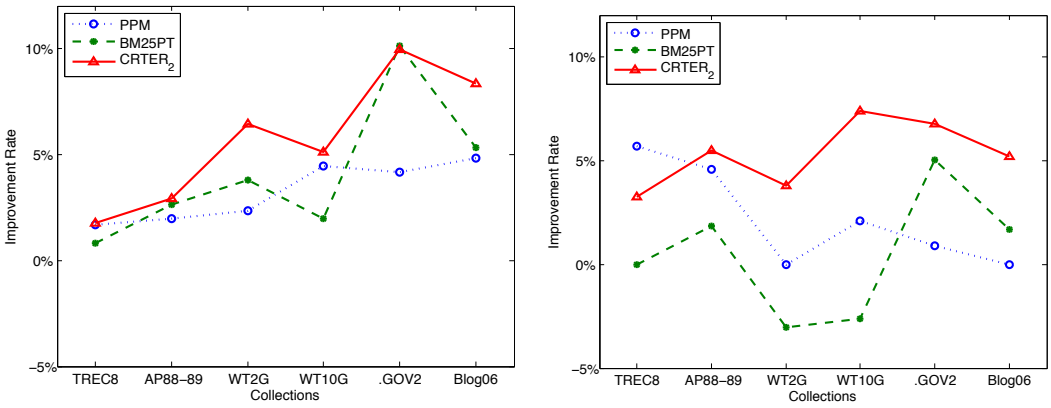
	TREC8	AP88-89	WT2G	WT10G	.GOV2	Blog06
μ	800	600	1850	650	800	1350

Table 3.4: Parameter μ for Dirichlet LM

In PPM, a query term occurrence is only propagated to other query terms' nearest occurrences. In BM25TP, an occurrence of a query term contributes to the query terms at its previous posting and its following posting. Suppose we have an example query $\{q_1, q_2\}$, and the following document

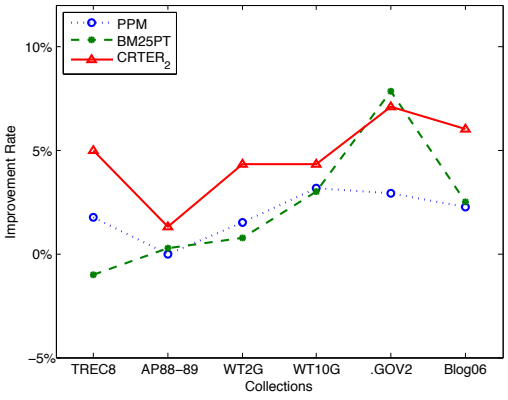
$$\{...q_1^{(1)}q_2^{(1)}...q_2^{(2)}...q_1^{(2)}...\}$$

where $q_1^{(1)}$ and $q_1^{(2)}$ are the first and second occurrences of q_1, and $q_2^{(1)}$ and $q_2^{(2)}$ are the first and second occurrences of q_2. Both PPM and $BM25TP$ consider the associations: $< q_1^{(1)}, q_2^{(1)} >$ and $< q_2^{(2)}, q_1^{(2)} >$. On the other hand, $CRTER_2$ considers the associations: $< q_1^{(1)}, q_2^{(1)} >$, $< q_1^{(1)}, q_2^{(2)} >$, $< q_1^{(2)}, q_2^{(1)} >$, and $< q_1^{(2)}, q_2^{(2)} >$. This could give us an explanation why our $CRTER_2$ model performs better in most of the cases. Therefore, the retrieval performance could be boosted via integrating the associations between all query term postings properly.



(a) MAP

(b) P@5



(c) P@20

Figure 3.9: Improvement Rates over BM25: Compare PPM, $BM25PT$ and $CRTER_2$ over six collections

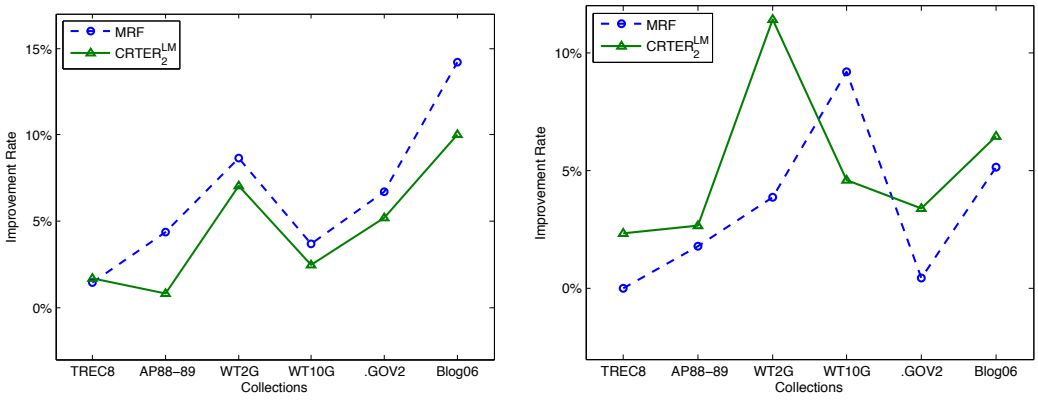
		TREC8	AP88-89	WT2G	WT10G	.GOV2	Blog06
Dirichlet LM	MAP	0.2552	0.2763	0.3060	0.2126	0.2983	0.3160
	P@5	0.5080	0.4531	0.5160	0.3480	0.5906	0.6187
	P@20	0.4020	0.4020	0.3810	0.2680	0.5268	0.5933
MRF	MAP	0.2589 (+1.450%)	**0.2884*** (+4.380%)	**0.3325*** (+8.660%)	**0.2204*** (+3.669%)	**0.3183*** (+6.705%)	**0.3609*** (+14.21%)
	P@5	0.5080 (+0.000%)	0.4612 (+1.788%)	0.5360 (+3.876%)	**0.3800*** (+9.195%)	0.5933 (+0.457%)	0.6507* (+5.172%)
	P@20	0.4020 (+0.000%)	**0.4112** (+2.289%)	0.3940 (+3.412%)	0.2715 (+1.306%)	**0.5510*** (+4.593%)	0.6140* (+3.489%)
Dirichlet $CRTER_2^{LM}$	MAP	**0.2595** (+1.684%)	0.2785 (+0.796%)	0.3275* (+7.026%)	0.2179 (+2.493%)	0.3138* (+5.196%)	0.3477* (+10.03%)
	P@5	**0.5200** (+2.362%)	**0.4653** (+2.693%)	**0.5750*** (+11.43%)	0.3640* (+4.598%)	**0.6107*** (+3.403%)	**0.6587*** (+6.465%)
	P@20	**0.4120** (+2.487%)	0.4092 (+1.791%)	**0.4025** (+5.643%)	**0.2740** (+2.238%)	0.5455 (+3.314%)	**0.6263*** (+5.562%)

Table 3.5: Comparison between two Language Model based proximity models: $CRTER_2^{LM}$ with MRF. All the results are evaluated in terms of MAP, P@5, and P@20. "*" means the improvements over the Dirichlet LM are statistically significant ($p<0.05$ with Wilcoxon Matched-pairs Signed-rank test).

3.6.5 Comparison with Major Proximity Language Models

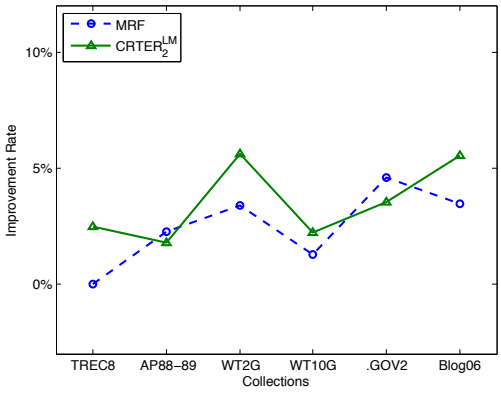
As we have discussed previously in Section 3.4.3, $CRTER_2$ can be extended on the basis of Language Models (LM). We use Dirichlet LM as the baseline model, which is language model with Dirichlet smoothing technique (Zhai and Lafferty 2001). We test the performance of $CRTER_2^{LM}$ by comparing Dirichlet $CRTER_2^{LM}$ with Markov Random Field (MRF) (Metzler and Croft 2005) and Positional Language Model (PLM) (Lv and Zhai 2009), which are described in Section 3.5.4.

First, we implement MRF under the same experimental environment and the same data sets as Dirichlet $CRTER_2^{LM}$. Both of these two models uses Dirichlet LM as the basic model, and the parameter μ is tuned to be optimal on each data set, as



(a) MAP

(b) P@5



(c) P@20

Figure 3.10: Improvement Rates over Dirichlet LM: Compare MRF and $CRTER_2^{LM}$ over six collections

shown in Table 3.4. The corresponding results are shown in Table 3.5. The other parameters are optimized by hill climbing as in (Metzler and Croft 2005). There is only one data collection and the corresponding query set exactly the same as the one in (Metzler and Croft 2005), which is WT10G. Our implemented MRF has a very

similar MAP on WT10G as the one reported in (Metzler and Croft 2005). We can see from Table 3.5 that $CRTER_2^{LM}$ and MRF have very comparable results. To illustrate the results in Table 3.5 graphically, we re-plot the improvement rates of MRF and $CRTER_2^{LM}$ over Dirichlet LM in Figure 3.10. Under the MAP evaluation metric, MRF slightly outperforms $CRTER_2^{LM}$ on most collections. For the top 5 and top 20 documents, $CRTER_2^{LM}$ has better performance than MRF on most of the data collections. The difference between MRF and $CRTER_2^{LM}$ is not significant. Therefore, it is reasonable to state that $CRTER_2^{LM}$ favours top ranked documents (e.g. top 5 and 20) compared to MRF. The experimental results show that $CRTER_2^{LM}$'s retrieval performance is comparable to MRF. One of the possible reasons could be that similar features are incorporated in $CRTER_2^{LM}$ and MRF. On the other hand, the modeling ideas of $CRTER_2^{LM}$ and MRF are quite different so that $CRTER_2^{LM}$ favours top ranked documents. There are some cases that $CRTER_2^{LM}$ can handle term dependency more specifically than MRF. For example, if each of q_1, q_2, q_3, q_4 occurs once in a document and $1 < dist_{q_1,q_2} < dist_{q_3,q_4} < N_{window}$, where N_{window} is the fixed window size, then MRF weights the dependency of $\{q_1, q_2, D\}$ and the dependency of $\{q_3, q_4, D\}$ the same. However, $CRTER_2^{LM}$ weights the dependency of $\{q_1, q_2, D\}$ higher than the dependency of $\{q_3, q_4, D\}$, because the distance between q_1 and q_2 is shorter than the distance between q_3 and q_4.

Further, the effectiveness of $CRTER_2^{LM}$ is evaluated by a cross-comparison with PLM (Lv and Zhai 2009). PLM estimates a language model for every single position in a document. Among various kernel functions tested, the PLM model is shown to provide the best retrieval performance with the Gaussian kernel. We directly use the results reported in (Lv and Zhai 2009). For PLM, the baseline is the fixed-length arbitrary passage retrieval method under the language modeling framework (e.g. Passage LM in Table 3.6). Since we are using different LM baselines, we compare PLM with LM-based $CRTER_2^{LM}$ and BM25-based $CRTER_2$ by the relative improvements in terms of percentage over the basic retrieval models for fair comparison. The comparison is conducted on the datasets used in (Lv and Zhai 2009), including AP88-89, WT2G and TREC8. As illustrated in Table 3.6, PLM improves the LM baseline by 1.3%, 2.0%, and 1.1% on TREC8, AP88-89 and WT2G, respectively. The $CRTER_2^{LM}$ improves Dirichlet LM by 1.7%, 0.8% and 7.0% on TREC8, AP88-89 and WT2G, and the corresponding improvement over BM25 by $CRTER_2$ is 1.8%, 2.9% and 6.4%, respectively. Overall, PLM, $CRTER_2$ and $CRTER_2^{LM}$ all boost their baselines. In particular, compared with PLM, $CRTER_2^{LM}$ has more improvement on TREC8 and WT2G, and less improvement on AP88-89. $CRTER_2$'s improvement rates over BM25 are all higher than the improvement rates of PLM over passage LM. We can conclude that $CRTER_2^{LM}$ and $CRTER_2$ have comparable

	TREC8	AP88-89	WT2G
Passage LM	0.2518	0.2154	0.3249
PLM	0.2550 (+1.3%)	0.2198 (+2.0%)	0.3285 (+1.1%)
Dirichlet LM	0.2552	0.2763	0.3060
$CRTER_2^{LM}$	0.2595(+1.7%)	0.2785(+0.8%)	0.3275(+7.0%)
BM25	0.2561	0.2710	0.3156
$CRTER_2$	0.2606 (+1.8%)	0.2789 (+2.9%)	0.3359 (+6.4%)

Table 3.6: Direct MAP Comparison with PLM

(if not better) performance compared to PLM.

3.7 Further Analysis and Discussions

In this section, we first investigate the possible reasons why the proposed $CRTER$ models perform well with a case study. Then in Section 3.7.2, we present examples for bigram Cross Terms for analyzing the proposed models. In Section 3.7.3, we test the stability of the proposed models by retrieving different topic sets on the same collection.

3.7.1 A Case Study on the Manually Judged Documents

We conduct a case study to investigate the possible reasons why the proposed bigram models are effective. In particular, out of the 550 test topics used in our experiments, we show a case study on topic 931 of the Blog06 collection. The title of topic 931

is “Fort McMurray”, which refers to the Canadian oil boom town named Fort McMurray in Alberta. When we read each word individually, “Fort” usually represents a permanent army post, and “McMurray” is a family name. Therefore, the term association between ‘Fort” and “McMurray” would be very useful in finding the documents that are relevant to the topic “Fort McMurray”. In detail, using the judged documents, namely the golden standard provided by TREC, we explore how these models could differentiate relevant documents from non-relevant ones. In Table 3.7, we list our proposed bigram models and the proximity models that are presented in Section 3.6.4 and 3.6.5, namely PPM, $CRTER_2$, $BM25TP$, MRF and $CRTER_2^{LM}$. We explore how these models could distinguish the relevant documents from non-relevant documents. For each model, we extract the part that is directly related to term association. For PPM, new term frequency with proximity is utilized in BM25 instead of the raw term frequency. A new term frequency is related to both the number of occurrences of a term and the term counts propagated from other terms. We also extract the bigram Cross Term frequency in $CRTER_2$ for comparison with PPM. For both $BM25TP$ and MRF, additional proximity weights are linearly combined into the original weighting functions. We also extract the Cross Terms' weights in $CRTER_2$ and $CRTER_2^{LM}$ for comparison. Table 3.7 shows the average values on all the relevant documents and average values on non-relevant documents.

We also calculate the ratios of the results on relevant documents over the results on non-relevant documents. A model with a higher ratio would be more effective in identifying the relevant documents from the non-relevant ones for this topic.

Model	Description	Relevant	non-relevant	$Ratio(\frac{Relevant}{non-relevant})$
PPM	Position dependent Frequency	3.01	6.24	0.48
$CRTER_2$	Bigram Cross Term Frequency	2.96	0.42	7.11
$BM25TP$	Weight of term proximity	3.56	1.15	3.08
$CRTER_2$	Weight of bigram Cross Terms	140.23	24.23	5.79
MRF	Weight of phrases	23.21	12.64	1.84
$CRTER_2^{LM}$	Weight of bigram Cross Terms	11.89	3.17	3.75

Table 3.7: A Case Study: Distribution of terms on relevant and non-relevant documents (Topic 931 "Fort McMurray" on the Blog06 Collection). In the first row, PPM uses position dependent frequency with proximity in BM25 instead of the raw term frequency. For $CRTER_2$, bigram Cross Term frequency is extracted to compare with PPM. In the second and third rows, we extract the additional proximity weight of each model for comparison.

From Table 3.7, we can see that $CRTER_2$'s ratio is much higher than PPM's ratio in terms of frequencies. $CRTER_2$'s ratio is also higher than $BM25TP$'s ratio in terms of the additional weight. In addition, $CRTER_2^{LM}$'s ratio is higher than MRF's ratio in terms of their additional weights. For this topic, $CRTER_2$ and $CRTER_2^{LM}$ can better distinguish the relevant documents from the non-relevant documents compared with PPM, $BM25TP$ and MRF accordingly.

Similar patterns can be found on most of the topics. In Table 3.8, we show the comparisons of the proximity models on all the topics of Blog06 collection. The

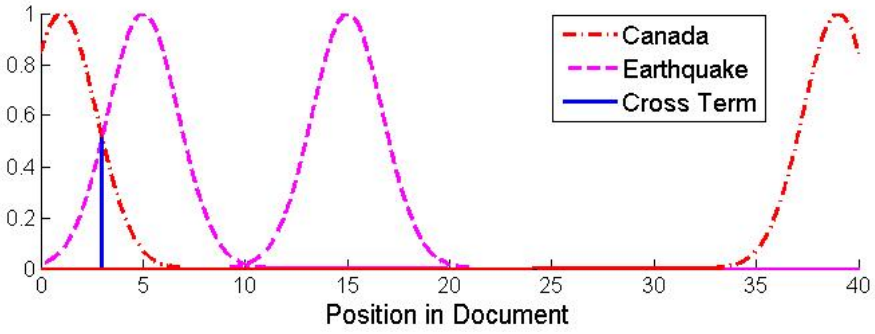
# of Topics on Blog06 Collection	150
# of Topics with More Than One Term	101
# of Topics that $CRTER_2$ distinguishes better than PPM	66
# of Topics that $CRTER_2$ distinguishes better than $BM25TP$	73
# of Topics that $CRTER_2^{LM}$ distinguishes better than MRF	83

Table 3.8: Comparisons on the judged documents over the Blog06 collection

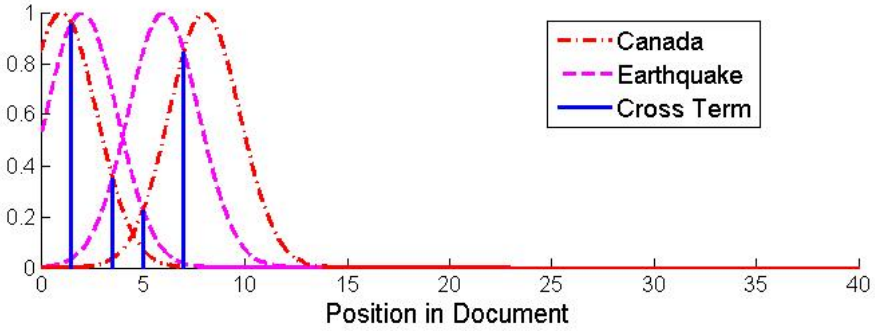
Blog06 collection includes 150 topics, where 49 topics contain one term. Here we analyze the rest of the 101 topics that contain two or more terms. We found that $CRTER_2$ better distinguishes relevant documents from non-relevant documents than PPM on 66 out of 101 topics; $CRTER_2$ distinguishes better than $BM25TP$ on 73 topics; and $CRTER_2^{LM}$ distinguishes better than MRF on 83 topics. From this case study, we can see that our proposed $CRTER$ models have better performance in distinguishing the relevant documents from the non-relevant documents on most of the topics. This could give us a possible explanation why $CRTER$ models can perform better than other models in retrieving.

3.7.2 An Analysis on the Bigram Cross Term Example

Figure 3.11 gives an example of a bigram Cross Term by query terms "Canada" and "Earthquake", which corresponds to the example documents in Figure 1.4. "Canada" and "Earthquake" occur twice in both the relevant document and the non-relevant



(a) A non-relevant document



(b) A relevant document

Figure 3.11: A case study: example of bigram Cross Term

document. A traditional retrieval model will assign weights to both documents similarly. Here we visualize the bigram Cross Term generated by "Canada" and "Earthquake" in these two documents. In Figure 3.11, we plot the query terms' impact shape functions and the corresponding bigram Cross Terms' values over the whole document. We use Gaussian kernel with a small σ for a better view of the Cross Terms. Figure 3.11(a) shows the non-relevant document and Figure 3.11(b) is related the relevant document.

We can see that the bigram Cross Term occurs in the non-relevant document (Figure 3.11(a)) only once, while it occurs in the relevant document (Figure 3.11(b)) four times. In particular, supposing we only have these two documents in the collection, we calculate the bigram Cross Term's values, frequency and weight in each document. These definitions are introduced in Section 3.1, 3.2 and 3.4.2. In Table 3.9, we can see that the bigram Cross Term has a higher within-document frequency in the relevant document than that in the non-relevant document. This is because the bigram Cross Term has higher value when two query terms occur closer and the bigram Cross Term' value is lower when two query terms are farther away. Eventually, the Cross Term's weight in the relevant document is higher than that in the non-relevant document. Since query term weights of these two documents are similar, this relevant document's overall weight is higher than the non-relevant docu-

ment's overall weight. This example demonstrates that bigram Cross Term is a good measure for distinguishing relevant documents from non-relevant documents.

	Non-relevant	Relevant
Cross Term occurrence	1	4
Cross Term values	0.5121	0.9567, 0.3519, 0.2225, 0.8442
Cross Term within-document frequency	0.5121	2.3753
Cross Term document frequency	1.1059	
Cross Term within-query frequency	0.9567	
Cross Term weight	0.2634	0.5723

Table 3.9: The values corresponding to the non-relevant and the relevant documents in the case study

3.7.3 The Stability of Using Different Topic Sets

	TREC8			Robust		
	MAP	P@5	P@20	MAP	P@5	P@20
BM25	0.2561	0.4920	0.4000	0.2420	0.3458	0.4562
$CRTER_2$	0.2606	0.5080	0.4200	0.2452	0.3494	0.4586
	(+1.757%)	(+3.252%)	(+5.000%)	(+1.322%)	(+1.041%)	(+0.526%)
Dirichlet LM	0.2552	0.5080	0.4020	0.2562	0.4739	0.3653
MRF	0.2589	0.5080	0.4020	0.2651	0.4867	0.3691
	(+1.450%)	(0.000%)	(0.000%)	(+3.474%)	(+2.701%)	(+1.040%)
$CRTER_2^{LM}$	0.2595	0.5200	0.4120	0.2666	0.4827	0.3741
	(+1.684%)	(+2.362%)	(+2.487%)	(+4.059%)	(+1.857%)	(+2.409%)

Table 3.10: Experimental results on robust track

Finally, we test our models on the same collection with different topic sets. Some studies in the past few years conducted experiments on the TREC Robust 2004

(Bendersky et al. 2010, Wang and Zhu 2009). The Robust and TREC8 have the same data collection containing newswire articles. TREC8 includes topics 401-450, and Robust includes more topics (301-450) and 601-700. The experimental results are shown in Table 3.10. We can see that the tendencies of the performance on TREC8 and Robust are very similar. $CRTER_2$ has better retrieval precision compared to BM25. $CRTER_2^{LM}$ and MRF outperform Dirichlet LM, and $CRTER_2^{LM}$'s retrieval performance is better than MRF in most of the cases. These results further confirm the effectiveness and stability of $CRTER_2$ and $CRTER_2^{LM}$.

4 N-gram Cross Term Retrieval Model

In the previous chapter, we have defined the concept of bigram Cross Term, which is generated from two query terms. In this chapter, we further discuss the association among n query terms ($n > 2$). Section 4.1 presents the concept of n-gram Cross Term. Section 4.2 discusses the estimations for n-gram Cross Term variants, as well as defining several new distance metrics for n terms. Then we recursively define a n-gram CRoss TErm Retrieval ($CRTER_n$) model for $n > 2$ in Section 4.3. Further, we present a detailed analysis on the algorithms and complexities in Section 4.4. We test the performance of the proposed $CRTER_n$ model in Section 4.5.

4.1 A New Pseudo Term: N-gram Cross Term

When there are n query terms, there doesn't exist an intersection where all the term impact functions cross each other. Figure 4.1 shows an example when there are multiple query terms that occur in one document. f_1, f_2, and f_3 represent the impact

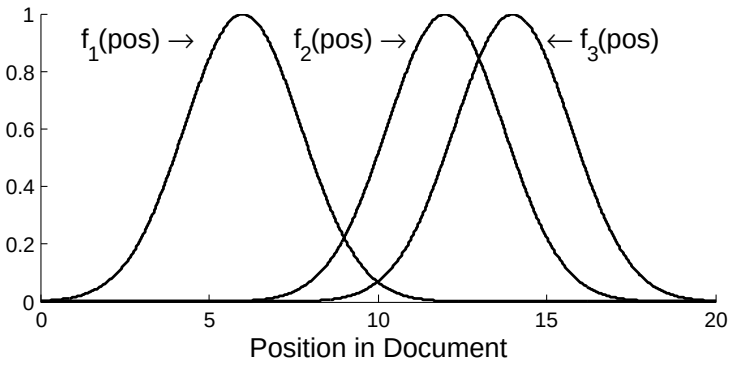


Figure 4.1: An example of Cross Term by Multiple Query Terms

functions for the occurrences of three query terms. f_1 and f_2 have one intersection. f_2 and f_3 have another intersection. However, we can see that there is no non-zero point that three query terms' impact functions all pass through no matter where these query terms occur in the document. Therefore, we have to find an alternative way to define n-gram Cross Term. As we can see from the previous chapter, a basic component in all the definitions is the distance between two query terms. In the higher dimensional case, we utilize the distance for n query term occurrences to define the n-gram Cross Term. A formal definition of n-gram Cross Term is shown in Definition 6.

Definition 6. *Given n query terms* $q_{i_1}, q_{i_2}, ..., q_{i_n}$ *occurring at positions* $pos_{k_1,i_1}, pos_{k_2,i_2}, ..., pos_{k_n,i_n}$, *a* ***n-gram Cross Term*** *occurs if its value* $Kernel\left(\frac{1}{2}dist(pos_{k_1,i_1}, pos_{k_2,i_2}, ..., pos_{k_n,i_n})\right)$

is larger than 0. We call a n-gram Cross Term ***nCT****, denoted as* $q_{i_1,i_2,...,i_n}$.

In Definition 6, $dist(\cdot)$ is a distance metric for n query term occurring positions which will be described in Section 4.2.1-4.2.3, and $Kernel(\cdot)$ is the kernel function adopted in query term's impact function which will be presented in Section 3.3.

4.2 Estimations for N-gram Cross Term Variants

We have defined how to do the counting for bigram Cross Term in Sections 3.2.1, 3.2.2 and 3.2.3. One identical part in formulas (3.1), (3.3) and (3.4) is the distance between the co-occurring positions of two query terms. We redefine the variants for an nCT by integrating the distance among the co-occurring positions of n query terms.

Definition 7. *Given a nCT,* $q_{i_1,i_2,...,i_n}$*, its variants are defined as follows.*

- *The within-document nCT frequency*

$$tf\left(q_{i_1,i_2,...,i_n}, D\right) = \sum_{k_1=1}^{tf_1} \sum_{k_2=1}^{tf_2} ... \sum_{k_n=1}^{tf_n} Kernel\left(\frac{1}{2} dist(pos_{k_1,i_1}, pos_{k_2,i_2}, ..., pos_{k_n,i_n})\right) \tag{4.1}$$

  where $tf\left(q_{i_1,i_2,...,i_n}, D\right)$ *is the term frequency of* $q_{i_1,i_2,...,i_n}$ *in* D*, and* tf_j *is the term frequency of query term* q_{i_j}.

- *The number of documents containing* $q_{i_1,i_2,...,i_n}$

$$nd(q_{i_1,i_2,...,i_n}) = \sum_{D \in Index, Occur(q_{i_1,i_2,...,i_n}, D) \neq 0} \frac{tf(q_{i_1,i_2,...,i_n}, D)}{Occur(q_{i_1,i_2,...,i_n}, D)} \tag{4.2}$$

where $Occur(q_{i_1,i_2,...,i_n}, D)$ *is the number of occurrence of* $q_{i_1,i_2,...,i_n}$*, which is*

$$Occur(q_{i_1,i_2,...,i_n}, D) = \sum_{k_1=1}^{tf_1} \sum_{k_2=1}^{tf_2} ... \sum_{k_n=1}^{tf_n} \mathbf{1}_{\{Kernel(\frac{1}{2} dist(pos_{k_1,1}, pos_{k_2,1}, ..., pos_{k_n,n})) \neq 0\}}$$

- *The within-query nCT frequency*

$$\begin{aligned} qtf(q_{i_1,i_2,...,i_n}) &= Kernel(\frac{1}{2} \cdot dist(q_{i_1}, q_{i_2}, ..., q_{i_n})) \cdot min\{qtf(q_{i_1}), qtf(q_{i_2}), ..., qtf(q_{i_n})\} \\ &= Kernel(\frac{1}{2}) \cdot min\{qtf(q_{i_1}), qtf(q_{i_2}), ..., qtf(q_{i_n})\} \end{aligned} \tag{4.3}$$

where $qtf(q_{i_1,i_2,...,i_n})$ *is the within query term frequency of* $q_{i_1,i_2,...,i_n}$*, and* $qtf(q_{i_j})$ *is the query term frequency of query term* q_{i_j}*. Here we assume that query terms are adjacent to each other and let* $dist(q_{i_1}, q_{i_2}, ..., q_{i_n})$ *=1.*

4.2.1 Lp-Norm Distances for N Terms

Traditional distance definitions evaluate how far away two points are. We newly define several distances for multi-terms, in order to consider the association among n query terms. Suppose we have n query terms q_{i_1}, q_{i_2}, q_{i_3}, ..., q_{i_n} that occur in a document D with the frequency of tf_1, tf_2, tf_3, ..., tf_n. For a set of co-occurring query terms, we record the positions of their occurrences in D as $p_1, p_2, p_3, ..., p_n$. The distance metrics for a set of query term occurrences are defined as follows.

One way to naturally define multi-term distance is to learn from the distance concepts in geometry. In the Euclidean Space $\Re^n$, the p-norm distance evaluates how far apart two points are. For a vector $(x_1, x_2, ..., x_n)$, and a vector $(y_1, y_2, ..., y_n)$, a generalized form of Lp-Norm distance is $||x - y||_p = (\sum_{i=1}^{n} |x_i - y_i|^p)^{1/p}$. The Lp-Norm is commonly used for values 1, 2, and infinity. We redefine L1-Norm, L2-Norm and L∞-Norm distances for query term co-occurrences as follows.

Definition 8. *L1-Norm Based Distance*

$$dist_{L1}(p_1, p_2, p_3, ..., p_N) = \sum_{i \neq j} |p_i - p_j| \tag{4.4}$$

Definition 9. *L2-Norm Based Distance*

$$dist_{L2}(p_1, p_2, p_3, ..., p_N) = \sqrt{\sum_{i \neq j} (p_i - p_j)^2} \tag{4.5}$$

Definition 10. *L∞-Norm Based Distance*

$$dist_{L\infty}(p_1, p_2, p_3, ..., p_N) = \max_{i \neq j} |p_i - p_j| \tag{4.6}$$

Please note that the above definitions are different from the distance definitions in geometry, since we are not evaluating the distance between two vectors. L1-Norm Based Distance is the average distance of all possible different position pairs. L2-Norm Based Distance is the Root Sum Squares (RSS) of the difference of all possible position pairs. L∞-Norm Based Distance is the maximum distance of all possible different position pairs. When n=2, all the above distance metrics have the same form as the distance for bigram Cross Term in (3.2).

4.2.2 Pairwise Distance

We also redefine two previously utilized distance metrics in IR. In (Tao and Zhai 2007), the minimum pair distance and the maximum pair distance are defined as the smallest/largest distance value of all pairs of unique matched query terms among

all the occurrences of two query terms . We modify the concept in this thesis to characterize the pair distance for n co-occurring query terms as follows.

Definition 11. *Pairwise Minimum Distance*

$$dist_{min}(p_1, p_2, p_3, ..., p_N) = \min_{i \neq j} (p_i - p_j) \tag{4.7}$$

Definition 12. *Pairwise Maximum Distance*

$$dist_{max}(p_1, p_2, p_3, ..., p_N) = \max_{i \neq j} (p_i - p_j) \tag{4.8}$$

4.2.3 Altitude and Hypotenuse Based Distance

The goal of a distance metric for n query terms is to have a lower value when the query terms are close and to have a higher value when the query terms are far apart. We naturally build two more metrics based on the altitude and hypotenuse of triangles defined as follows.

Definition 13. *Altitude Based Distance*

$$dist_{alti}(p_1, p_2, p_3, ..., p_n) = \sqrt{\prod_{i=1,2,...,n-1} (p'_{i+1} - p'_i)} \tag{4.9}$$

where $p'_1, p'_2, p'_3, ..., p'_n$ *is a ranked list generated from* $p_1, p_2, p_3, ..., p_n$*, such that* $p'_1 \leq p'_2 \leq p'_3 \leq ... \leq p'_n$ *and* $p'_i \in \{p_1, p_2, p_3, ..., p_n\}$ *for all* i*.*

Definition 14. *Hypotenuse Based Distance*

$$dist_{hypo}(p_1, p_2, p_3, ..., p_n) = \sqrt{\sum_{i=1,2,...,n-1} (p'_{i+1} - p'_i)^2} \tag{4.10}$$

Theorem 4.2.1. *Altitude Based Distance and Hypotenuse Based Distance have the following same properties.*

1. *Non-negative:* $dist_{alti}(p_1, p_2, p_3, ..., p_n) \geq 0$*;* $dist_{hypo}(p_1, p_2, p_3, ..., p_n) \geq 0$*.*

2. *Identity:* $dist_{alti}(p_1, p_2, p_3, ..., p_N) = 0$*;* $dist_{hypo}(p_1, p_2, p_3, ..., p_n) = 0$*, when* $p_1 = p_2 = ... = p_n$*.*

3. *Monotonic to* p_1*: when* $p'_1 < p_1$*,* $dist_{alti}(p'_1, p_2, p_3, ..., p_n) > dist_{alti}(p_1, p_2, p_3, ..., p_n)$ *and* $dist_{hypo}(p'_1, p_2, p_3, ..., p_n) > dist_{hypo}(p_1, p_2, p_3, ..., p_n)$*.*

4. *Monotonic to p_N: when $p'_n > p_n$, $dist_{alti}(p_1, p_2, p_3, ..., p'_n) > dist_{alti}(p_1, p_2, p_3, ..., p_n)$ and $dist_{hypo}(p_1, p_2, p_3, ..., p'_N) > dist_{hypo}(p_1, p_2, p_3, ..., p_n)$.*

The above properties are the same for both Altitude Based Distance and Hypotenuse Based Distance. It is easy to prove that both distances are larger than or equal to zero. When $p_1 = p_2 = ... = p_N$, Altitude Based Distance and Hypotenuse Based Distance are equal to zero. Due to the speciality of multiple inputs, it would be impossible to verify the triangle inequality property and symmetry property. The monotonic property means the distance values are larger when the span of the positions is wider, which is intuitively reasonable for a distance function in IR.

Theorem 4.2.2. *Besides the above properties, Altitude Based Distance and Hypotenuse Based Distance have the following difference. Suppose the span $p_N - p_1$ is fixed, the median positions affect Altitude Based Distance and Hypotenuse Based Distance differently: Altitude Based Distance has a higher value, while Hypotenuse Based Distance has a lower value, if the median positions tend to spread more evenly, i.e., $p_2 - p_1 = p_3 - p_2 = ... = p_N - p_{N-1}$.*

For the proof of the theorems, more details can be found in Appendix A. In this section, we define seven different distance measures from Formula (4.4) to (4.10) for multi-terms, which could also be applied in other models and domains. These measures are so-called distance, and they are norms rather than distances. In this thesis,

they are called "distance" for coherence. Furthermore, possible distance measures not listed as above may also be applicable in our proposed models.

4.3 The Recursive N-gram CRoss-TErm Retrieval ($CRTER_n$) Model

We further extend the BM25 model by considering the association among multiple query terms, namely n-gram Cross Term. An n-gram CRoss TErm Retrieval ($CRTER_n$) model is built recursively based on $CRTER_{n-1}$ and n-gram Cross Terms $q_{i_1,\ldots,i_n}$ as follows:

$$CRTER_n(D) = (1-\lambda_n)\cdot CRTER_{n-1}(D) + \lambda_n \cdot \sum_{1\le i_1<i_2<\ldots<i_n\le K} w_n(q_{i_1,\ldots,i_n}, D) \quad (4.11)$$

where w_n is the weight of an n-gram Cross Term $q_{k_1,\ldots,k_n}$.

$$\begin{aligned} w_n(q_{i_1,i_2,\ldots,i_n}, D) &= \frac{(k_1+1)*tf(q_{i_1,\ldots,i_n}, D)}{K + tf(q_{i_1,\ldots,i_n}, D)} \\ &\quad * \frac{(k_3+1)*qtf(q_{i_1,\ldots,i_n})}{k_3 + qtf(q_{i_1,\ldots,i_n})} * \log \frac{N - nd(q_{i_1,\ldots,i_n}) + 0.5}{nd(q_{i_1,\ldots,i_n}) + 0.5} \end{aligned}$$

where we replace the term dependent variants with the n-gram Cross Term's variants defined in Equation (4.1), Equation (4.2), and Equation (4.3).

In $CRTER_n$ model, λ_n is a parameter balancing the influence of n-gram Cross Term. When λ_n is equal to 0, the association among n query terms is not considered, and $CRTER_n$ becomes $CRTER_{n-1}$. When λ_n is equal to 1, $CRTER_n$ only accumulates the weights of n-gram Cross Terms. This recursive model has $n-1$ parameters in total: $\lambda_2, ..., \lambda_n$. To simplify the parameter optimization process, we only optimize λ_n in $CRTER_n$, and use the $\lambda_2, ..., \lambda_{n-1}$ obtained from $CRTER_{n-1}$. The n-gram model for $CRTER_2^{LM}$ can be similarly defined. In this chapter, we only focus on BM25-based n-gram model.

4.4 Algorithm and Time Analysis

In this section, we present the $CRTER_2$ and $CRTER_n$ algorithms with analyses on their time complexities. In order to shorten the extra retrieval time for Cross Terms, we build an offline vocabulary that stores Cross Term information. All the possible Cross Terms in the vocabulary are generated from the queries. We first calculate and permanently store the Cross Terms' within-document term frequency and the number of documents containing the Cross Term. Meanwhile, the variants for query terms are stored in the index as well. It is a common technique used in many existing

INPUT: A query Q, a document D, the location-based $Index$, $nd(q_{i,j})$ and $tf(q_{i,j})$ for $q_{i,j}$ in Q

for all $q_i \in Q$ **do**
  Compute $qtf(q_i)$
  Get $nd(q_i)$
end for
for all $q_i, q_j \in Q$ **do**
  Compute $qtf(q_{i,j})$
  Get $nd(q_{i,j})$
end for
$w(D) = 0$
$w_2(D) = 0$
for all $D \in Index$ **do**
  for all $q_i \in Q$ **do**
    Get $tf(q_i, D)$
    Compute $w(q_i, D)$
    $w(D) = w(D) + w(q_i, D)$
  end for
  for all $q_i, q_j \in Q$ **do**
    Get $tf(q_{i,j}, D)$
    Compute $w_2(q_{i,j}, D)$
    $w_2(D) = w_2(D) + w_2(q_{i,j}, D)$
  end for
end for
Normalize $w(D)$
Normalize $w_2(D)$
for all $D \in Index$ **do**
  Compute $CRTER_2(D) = (1 - \lambda_2) * w(D) + \lambda_2 * w_2(D)$
end for
OUTPUT: $CRTER_2(D)$

Figure 4.2: Algorithm for $CRTER_2$

IR experiments.

In Figure 4.2, we show the algorithm for $CRTER_2$ model. Suppose the number of query terms in a query Q is $|Q|$, and the total number of documents in an

INPUT: A query Q, a document D, the location-based $Index$,
$nd(q_{i_1,i_2})$ and $tf(q_{i_1,i_2})$ for q_{i_1,i_2} in Q,
$nd(q_{i_1,i_2,i_3})$ and $tf(q_{i_1,i_2,i_3})$ for q_{i_1,i_2,i_3} in Q

1: **for all** $q_i \in Q$ **do**
2: Compute $qtf(q_i)$
3: Get $nd(q_i)$
4: **end for**
5: **for all** $q_{i_1}, q_{i_2} \in Q$ **do**
6: Compute $qtf(q_{i_1,i_2})$
7: Get $nd(q_{i_1,i_2})$
8: **end for**
9: **for all** $q_{i_1}, q_{i_2}, q_{i_3} \in Q$ **do**
10: Compute $qtf(q_{i_1,i_2,i_3})$
11: Get $nd(q_{i_1,i_2,i_3})$
12: **end for**
13: $w(D) = 0$
14: $w_2(D) = 0$
15: $w_3(D) = 0$
16: **for all** $D \in Index$ **do**
17: **for all** $q_i \in Q$ **do**
18: Get $tf(q_i, D)$
19: Compute $w(q_i, D)$
20: $w(D) = w(D) + w(q_i, D)$
21: **end for**
22: **for all** $q_{i_1}, q_{i_2} \in Q$ **do**
23: Get $tf(q_{i_1,i_2}, D)$
24: Compute $w_2(q_{i_1,i_2}, D)$
25: $w_2(D) = w_2(D) + w_2(q_{i_1,i_2}, D)$
26: **end for**
27: **for all** $q_{i_1}, q_{i_2}, q_{i_3} \in Q$ **do**
28: Get $tf(q_{i_1,i_2,i_3}, D)$
29: Compute $w_3(q_{i_1,i_2,i_3}, D)$
30: $w_3(D) = w_3(D) + w_3(q_{i_1,i_2,i_3}, D)$
31: **end for**
32: **end for**
33: Normalize $w(D)$
34: Normalize $w_2(D)$
35: Normalize $w_3(D)$
36: **for all** $D \in Index$ **do**
37: Compute $CRTER_3(D) = (1 - \lambda_3) * ((1 - \lambda_2) * w(D) + \lambda_2 * w_2(D)) + \lambda_3 * w_3(D)$
38: **end for**

OUTPUT: $CRTER_3(D)$

Figure 4.3: Algorithm for $CRTER_3$

index is $|Index|$. The first 8 steps process the query, where the within-query term frequency is computed for both query terms and bigram Cross Terms, which are $qtf(q_i)$ and $qtf(q_{i,j})$. Meanwhile, the number of documents containing the term $nd(q_i)$ and the number of documents containing the Cross Term $nd(q_{i,j})$ are obtained from the index. The time complexity of this part is $O(|Q| + |Q|^2)$. The remaining steps are document processing through the index. Query terms and bigram Cross Terms are computed separately and then combined as shown in Formula (3.14). In Steps 9 to 22, $w(D)$ represents the score from the basic weighting function, and $w_2(D)$ is the weighting score fore the Cross Terms as in Formula (3.15). For each document, we read the within-document query term frequency $tf(q_i)$ and the within-document bigram Cross Term frequency $tf(q_{i,j})$ from the hard disk, and compute the corresponding term weights. The time complexity of this part is $O(|Index| \cdot (|Q| + |Q|^2))$. In Steps 23 to 27, the weights are normalized and linearly combined, where the time complexity is $O(|Index|)$. This algorithm for $CRTER_2$ model has the time complexity as follows

$$O(|Q| + |Q|^2 + |Index| \cdot (|Q| + |Q|^2) + |Index|) \tag{4.12}$$

We can see from formula (4.12) that the processing time for the query in $CRTER_2$

is $O(|Q| + |Q|^2)$ within both the documents and the queries, while a traditional weighting model's time complexity for processing the query is $O(|Q|)$. In most of the applications, the queries are usually short. Therefore, the time increase is not very significant.

To analyze the time complexity of $CRTER_n$ model, we first show the algorithm for $CRTER_3$ model as a special case in Figure 4.3. In $CRTER_3$, the main procedures are the same as $CRTER_2$ with a few modifications. In steps 9 to 12, the within-query nCT frequency $qtf(q_{i_1,i_2,i_3})$ is calculated, and the number of documents containing nCT $nd(q_{i_1,i_2,i_3})$ is obtained from the index. In steps 27 to 31, we read the nCT term frequency $tf(q_{i_1,i_2,i_3}, D)$ from index and compute the weighing score for nCT $w_3(D)$. Then, $w_3(D)$ is normalized in step 35. Finally in steps 36 to 38, the weight of $CRTER_3$ is calculated based on $w(D)$, $w_2(D)$, and $w_3(D)$ as shown in Formula (4.11). For $CRTER_n$ model, we add the corresponding nCT processing procedures similarly. The time complexity increases for query processing within the query and the documents. Therefore, the generalized time complexity for $CRTER_n$ is

$$O(\sum_{l=1}^{l\leq n} |Q|^l + |Index| \cdot \sum_{l=1}^{l\leq n} |Q|^l + |Index|) \tag{4.13}$$

The time complexity increases exponentially with the number of grams considered. When n is fixed, formula (4.13) grows polynomially with the increase of the query

size $|Q|$. Therefore, it is a trade-off between considering a greater number of grams and spending more time for retrieval.

INPUT: A query Q, the location-based $Index$
for all $q_i, q_j \in Q$ **do**
  $nd(q_{i,j}) = 0$
  for all $D \in Index$ **do**
  $tf(q_{i,j}, D) = 0$
  $Occur(q_{i,j}, D) = 0$
    for $k_1 < tf(q_i, D)\&k_2 < tf(q_j, D)$
      $Kernel_{temp} = Kernel(\frac{1}{2}|pos_{k_1,i} - pos_{k_2,j}|)$
      if $Kernel_{temp} \neq 0$
        $tf(q_{i,j}, D) += Kernel_{temp}$
        $Occur(q_{i,j}, D) += 1$
      end if
    end for
    $nd(q_{i,j}) += \frac{tf(q_{i,j},D)}{Occur(q_{i,j},D)}$;
  end for
end for
OUTPUT: $tf(q_{i,j})$ and $nd(q_{i,j})$ for all $q_{i,j}$ in Q

Figure 4.4: Algorithm for Computing $tf(q_{i,j}, D)$ and $nd(q_{i,j})$

For the queries that have not been processed and the corresponding Cross Terms have not been stored in the index, we need to perform preliminary steps to process the Cross Terms. For $CRTER_2$, we show the algorithm to compute $tf(q_{i,j}, D)$ and $nd(q_{i,j})$ in Figure 4.4. Here the term positions are stored in the index. For each pair of query terms, the corresponding Cross Terms' term frequencies in each document $tf(q_{i,j})$ are calculated as in Formula (3.1), and the average term frequencies are added up to be the number of documents containing the Cross Term $nd(q_{i,j})$ as

INPUT: A query Q, the location-based $Index$

for all $q_{i_1}, q_{i_2}, ...q_{i_n} \in Q$ **do**
  $nd(q_{i_1,...,i_n}) = 0$
  for all $D \in Index$ **do**
  $tf(q_{i_1,...,i_n}, D) = 0$
  $Occur(q_{i_1,...,i_n}, D) = 0$
    for $k_1 < tf(q_{i_1}, D)$ **&** $k_2 < tf(q_{i_2}, D)$ **&** ... **&** $k_n < tf(q_{i_n}, D)$
      $Kernel_{temp} = Kernel(\frac{1}{2}dist(pos_{k_1,i_1}, pos_{k_2,i_2}, ..., pos_{k_n,i_n}))$
      if $Kernel_{temp} \neq 0$
        $tf(q_{i_1,...,i_n}, D) += Kernel_{temp}$
        $Occur(q_{i_1,...,i_n} += 1$
      end if
    end for
    $nd(q_{i_1,...,i_n}) += \frac{tf(q_{i_1,...,i_n},D)}{Occur(q_{i_1,...,i_n},D)}$;
  end for
end for

OUTPUT: $tf(q_{i_1,...,i_n})$ and $nd(q_{i_1,...,i_n})$ for all $q_{i_1,...,i_n}$ in Q

Figure 4.5: Algorithm for Computing $tf(q_{i_1,...,i_n}, D)$ and $nd(q_{i_1,...,i_n})$

in Formula (3.3). $Kernel(\cdot)$ is the kernel function used to simulate the term impact, and it is calculated by the formulas in Section 3.3. The total time spent in this process is $O(|Q|^2 \cdot |Index| \cdot \overline{tf}^2)$, where $\overline{tf}$ represents the within-document term frequency. For $CRTER_n$, Figure 4.5 shows the algorithm for computing $tf(q_{i_1,...,i_n}, D)$ and $nd(q_{i_1,...,i_n})$ as shown in Formula (4.1-4.2). The time spent in this process is $O(|Q|^n \cdot |Index| \cdot \overline{tf}^n)$. When n is fixed, the time complexity is a polynomial function to $|Q| \cdot \overline{tf}$ in a real-life setting since Cross Terms need to be extracted from the queries dynamically.

Reranking (Cormack et al. 2011, Kurland and Lee 2004) is an approach to re-

duce the computational complexity and reduce the running time in Information Retrieval. We rerank the top $k_{reranking}$ documents instead of the whole collection in our experiments. Then the additional time spent on $CRTER_2$ is $O(k_{reranking} \cdot |Q|^2 \cdot \overline{tf}^2)$ for processing Cross Terms, and $O(|Q| + |Q|^2 + k_{reranking} \cdot (|Q| + |Q|^2) + k_{reranking})$ for reranking respectively. Similarly, the overall additional time spent on the $CRTER_n$ is $O(\sum_{l=1}^{l \leq n} k_{reranking} \cdot |Q|^2 \cdot \overline{tf}^n)$ for processing all grams of Cross Terms, and $O(\sum_{l=1}^{l \leq n} |Q|^l + k_{reranking} \cdot \sum_{l=1}^{l \leq n} |Q|^l + k_{reranking})$ for reranking respectively. With a fixed constant $k_{reranking}$, the time spent on larger collections will be reduced significantly. We use $k_{reranking} = 2000$ in our experiments. In the future, we will study how to further reduce the computational complexity for the proposed algorithms in Figure 4.2 - 4.5.

4.5 Experiments

In previous subsections, we have illustrated the effectiveness and robustness of $CRTER_2$. Now we further conduct experiments to evaluate the effectiveness of $CRTER_n$ on the same collections over different settings. Further, we compare and analyze the runtime of BM25, $CRTER_2$ and $CRTER_n$. The difference between $CRTER_n$ and $CRTER_2$ is that $CRTER_n$ imports the nCT, which represents the association among more than two query terms. To compare $CRTER_n$ with $CRTER_2$, only the queries

Collection Name	# of Topics	# of Topics with ≥ 3 terms
TREC8	50	28
AP88-89	50	36
WT2G	50	24
WT10G	100	59
.GOV2	150	109
Blog06	150	35

Table 4.1: Overview of the TREC collections with more than 3 terms

with at least three words make sense. Therefore, we remove the queries that are shorter than three query terms. Table 4.1 shows the query information in this section. For the documents, we use the original documents and process them in the same way as that of in Section 3.5.1.

As we can see from Table 4.1, around half of the utilized standard TREC queries contain three or more terms on most of the datasets. Further, as we have discussed in Section 4.4, the computational complexity of the n-gram model grows exponentially with the increase of n. Therefore, we focus on discussing the case when $n = 3$ in this section. We have come to the conclusion from Section 3.6.3 that the recommended parameters of $CRTER_2$ are as follows: using Triangle Kernel as the kernel function, kernel parameter $\sigma = 25$, and balancing parameter $\lambda = 0.2$. The results of $CRTER_3$ with different distance metrics are compared with $CRTER_2$ using these recommended parameters.

	Eval Metric	TREC8	AP88-89	WT2G	WT10G	.GOV2	Blog06
$CRTER_2$	MAP	**0.2270**	**0.2286**	0.3338	0.2196	0.3322	0.2907
	P@5	**0.4286**	0.3944	**0.5500**	0.4475	0.6495	0.5543
	P@20	0.3554	**0.3222**	0.3687	0.3017	0.5830	0.5143
$CRTER_3$ with Pairwise-Min Distance	MAP	**0.2270**	**0.2286**	0.3338	0.2202	0.3322	0.2907
	P@5	**0.4286**	0.4000	**0.5500**	0.4508	0.6514	0.5657
	P@20	0.3589	**0.3222**	0.3687	0.3017	0.5853	0.5186
$CRTER_3$ with Pairwise-Max Distance	MAP	**0.2270**	**0.2286**	0.3338	0.2204	0.3322	0.2907
	P@5	**0.4286**	0.3944	**0.5500**	0.4508	0.6495	**0.5714***
	P@20	0.3607	**0.3222**	0.3687	0.3025	0.5858	0.5186
$CRTER_3$ with L_1 Distance	MAP	**0.2270**	**0.2286**	0.3381	0.2198	0.3329	0.2964
	P@5	**0.4286**	0.3944	**0.5500**	**0.4576**	0.6514	0.5657
	P@20	0.3643	**0.3222**	0.3687	0.3085	0.5835	0.5186
$CRTER_3$ with L_2 Distance	MAP	**0.2270**	**0.2286**	0.3388	0.2199	0.3332	0.2960
	P@5	**0.4286**	0.3944	**0.5500**	**0.4576**	0.6532	0.5657*
	P@20	0.3661	**0.3222**	0.3687	**0.3102**	0.5849	0.5143
$CRTER_3$ with L_∞ Distance	MAP	**0.2270**	**0.2286**	0.3395	0.2201	0.3330	0.2963
	P@5	**0.4286**	0.3944	**0.5500**	0.4508	**0.6569**	0.5657
	P@20	**0.3679**	**0.3222**	0.3687	0.3042	0.5853	0.5171
$CRTER_3$ with Altitude Distance	MAP	**0.2270**	**0.2286**	0.3440	**0.2211**	**0.3338**	**0.2994**
	P@5	**0.4286**	0.4000	**0.5500**	0.4542*	0.6550	0.5657
	P@20	0.3661	**0.3222**	0.3688	0.3042	0.5858	**0.5229**
$CRTER_3$ with Hypotenuse Distance	MAP	**0.2270**	**0.2286**	**0.3435**	0.2198	0.3332	0.2979
	P@5	**0.4286**	**0.4056**	**0.5500**	0.4508	0.6550	0.5657
	P@20	0.3607	**0.3222**	**0.3729**	0.3051	**0.5885**	**0.5229**

Table 4.2: Comparison between $CRTER_2$ and $CRTER_3$ with different distance metrics: Both use fixed parameters: Triangle Kernel, $\sigma = 25$, $\lambda = 0.2$. "*" means the improvements over $CRTER_2$ are statistically significant ($p<0.05$ with Wilcoxon Matched-pairs Signed-rank test).

4.5.1 Performance of $CRTER_3$

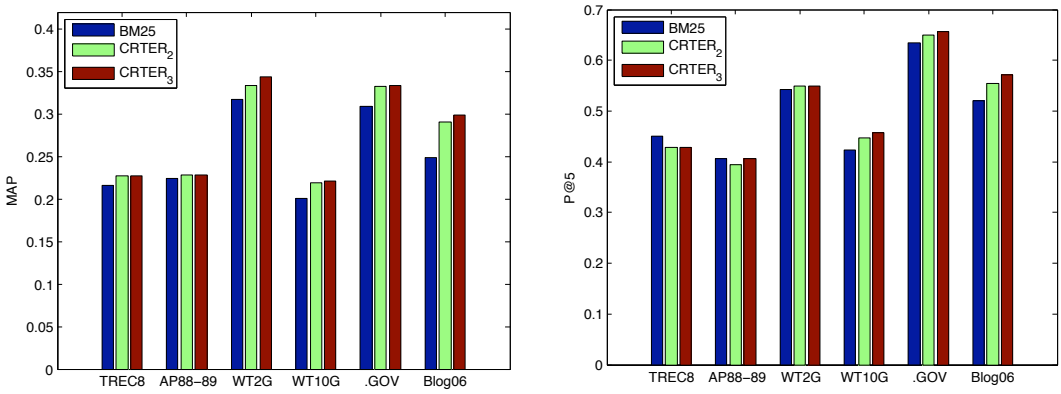
Seven distance metrics are used for $CRTER_3$, which are the Pairwise-Min Distance, Pairwise-Max Distance, L_1-Norm Based Distance, L_2-Norm Based Distance, L_∞-Norm Based Distance, Altitude Based Distance, and Hypotenuse Based Distance. The experimental results are shown in Table 4.2. Although only two cases are observed where $CRTER_3$ outperforms $CRTER_2$ significantly, there are still some small improvements of $CRTER_3$ over $CRTER_2$ in most of the cases. Comparing different distance metrics, more improvement is observed on $CRTER_3$ using Hypotenuse Based Distance.

4.5.2 Runtime Analysis

	TREC8	AP88-89	WT2G	WT10G	.GOV2	Blog06
BM25	1.527	0.999	0.893	8.381	652.449	22.340
$CRTER_2$	1.935	1.942	1.270	9.411	670.452	23.094
$CRTER_3$	2.633	2.349	1.886	12.105	730.141	28.714

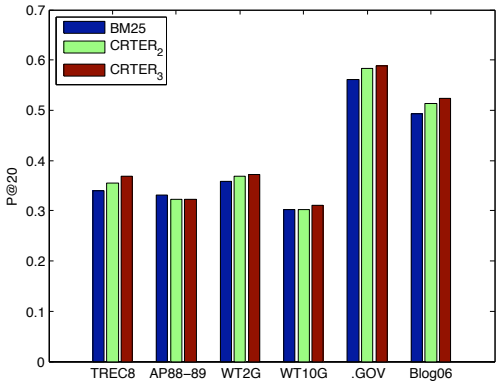
Table 4.3: Runtime (in Seconds)

Figure 4.6 shows the comparison between uni-gram model BM25, bi-gram model $CRTER_2$ and tri-gram model $CRTER_3$ on all the data sets using queries with equals to or more than three terms. All the experimental results are obtained by



(a) MAP

(b) P@5



(c) P@20

Figure 4.6: Overview performance of uni-gram model BM25, bi-gram model $CRTER_2$ and tri-gram model $CRTER_3$

using a fixed set of parameters. In most of the cases, $CRTER_2$ improves BM25, and $CRTER_3$ improves $CRTER_2$. In general, considering the association among three terms in $CRTER_3$ is a useful addition to using two terms in $CRTER_2$ if the time and space conditions allow, although $CRTER_2$ is sufficient to provide a reliable

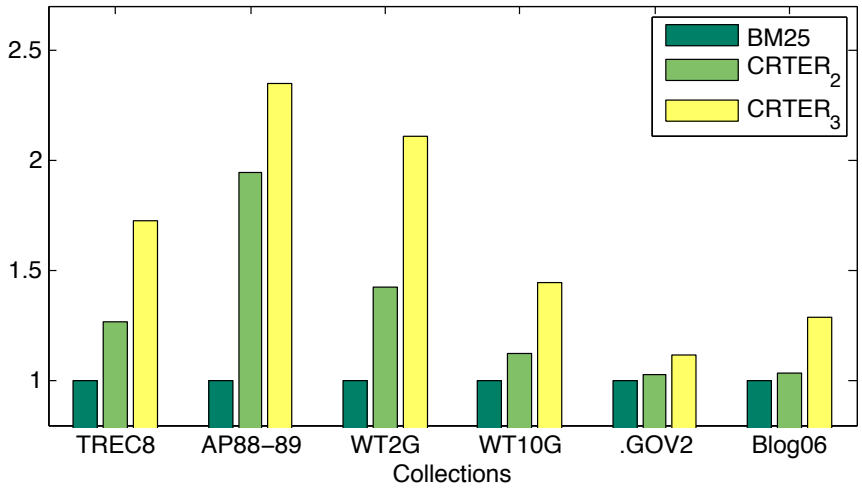


Figure 4.7: Runtime of BM25, $CRTER_2$ and $CRTER_3$ (divided by BM25's runtime)

retrieval performance on the collections used.

In Table 4.3, we present the runtime of BM25, $CRTER_2$ and $CRTER_3$ on each collection with the corresponding topic set. As a case study, the reported experiments are conducted on a computer with Intel 2.53GHz CPU and 8G memory. For a faster implementation, we rerank the top 2,000 documents instead of the whole collection. On each data set, we report the average runtime of 10 runs. To have an intuitive view over all the runtime, we further plot the runtime of BM25, $CRTER_2$ and $CRTER_3$ divided by BM25's runtime in Figure 4.7. In general, $CRTER_2$ consumes more time than BM25, and $CRTER_3$ consumes more time than $CRTER_2$. There is a balance between time complexity and effectiveness. The improvements of $CRTER_2$ over $BM25$ are higher than the improvements of $CRTER_3$ over $CRTER_2$. $CRTER_3$

	Online Retrieval (in Seconds)	Offline Retrieval (in Seconds)	MAP	P@5	P@20
Collection	7520.00	38.74	0.3540	0.6627	0.6277
Top 2000	940.39	28.37	0.3477	0.6587	0.6257

Table 4.4: A case study: trade-off between effectiveness and efficiency for reranking top documents on Blog06

is still an option to further boost the performance, if a higher accuracy is required. However, we also have to take into account the additional complexity of $CRTER_3$. Since we are using an offline vocabulary and the top 2,000 reranking technique, the relative increments on larger data sets are smaller.

We further use Blog06 as a case study to show the trade-off between the effectiveness and the efficiency for reranking the top documents in Table 4.4. We also show the time spent on both the online and offline retrieval discussed in Section 4.4. In the online retrieval, the Cross Terms have not been processed before retrieval. In the offline retrieval, the information of Cross Terms generated by given queries is stored in the index. In real applications where the queries are not known, building the offline index for all possible cross terms on the whole collection would have a very high space complexity. Further feature selection techniques are required in such implementations and we do not discuss them in detail in this thesis, which could be our future work. First of all, we can see that the online retrieval takes much more time than the offline retrieval. Secondly, reranking the top 2000 documents rather than the whole collection would further reduce the retrieval time. Finally, we are

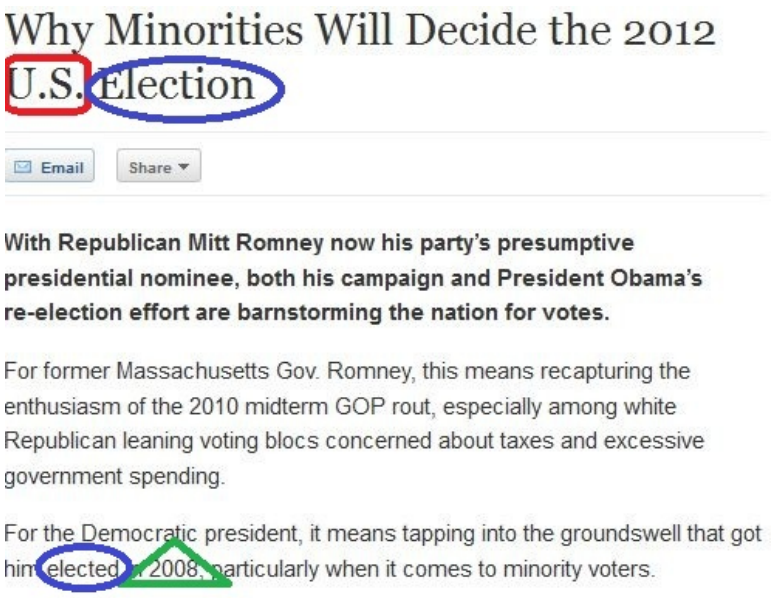
able to reduce the retrieval time from 7520.00 seconds to 28.37 seconds. On the other hand, reranking the top 2000 documents sacrifices minor accuracy. Therefore, we need to balance the trade-off between the effectiveness and the efficiency in real applications.

4.5.3 The Usefulness of N-gram Cross Terms

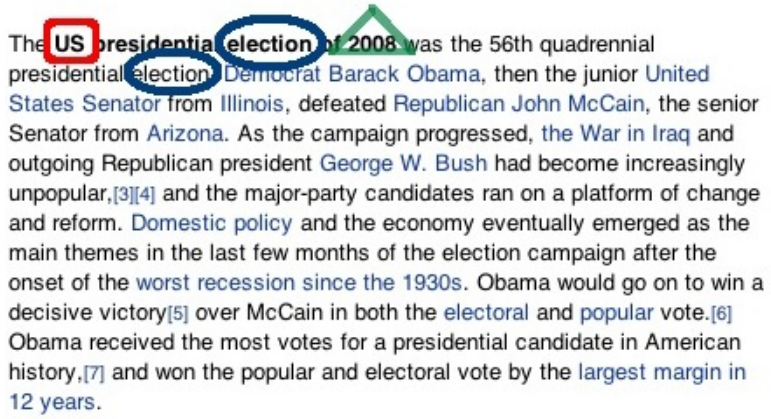
	Non-relevant	Relevant
Tri-gram Cross Term occurrences	2	2
Generating query term positions	(7, 8, 61), (7, 60, 61)	(2, 4, 6), (2, 12, 6)
Distances for query terms	53.009, 53.009	2.828, 7.211
Tri-gram Cross Term values	0.2099, 0.2099	0.9956, 0.9715
Tri-gram Cross Term within-document frequency	0.4198	1.9671

Table 4.5: The values corresponding to the non-relevant and the relevant documents in the tri-gram example

Figure 4.8 gives an example showing the usefulness of tri-gram dependency. In this figure, we show you an example of a three-term query namely "the U.S election of 2008". After stop words are removed, it becomes " U.S election 2008". Figure 4.8 (a) shows a non-relevant document and Figure 4.8 (b) shows a relevant document. We can see that the query terms have the same frequencies in both documents, while three query terms occur more closely in the relevant document. In Figure 4.8 (a), "US" is close to "election" at the beginning, and "elected" is close to "2008" at the



(a) A non-relevant document



(b) A relevant document

Figure 4.8: An Example of term association for a tri-gram: "the U.S election of 2008"

end. However, three query terms do not occur together as in Figure 4.8 (b). Trigram Cross Term represents the association among all three query terms.

We calculate the tri-gram Cross Term's values and it's frequencies in the non-

relevant and relevant documents, as shown in Table 4.5. The corresponding definitions for the tri-gram Cross Term and its variants are presented in Section 4.1 and 4.2. Here we use Gaussian kernel with parameter $\sigma = 15$. The tri-gram Cross Term occurs twice in each document, and has higher values in the relevant document than that in the non-relevant document. Therefore, the tri-gram Cross Term within-document frequency in the relevant document is higher than that in the non-relevant document. Since the other variants of these two documents are very similar, the document in which three query terms occur closely is boosted by employing tri-gram Cross Term in retrieval. This example illustrates that $CRTER_3$ could possibly further boost $CRTER_2$.

5 An Enhanced Context-sensitive Proximity Model

Term pair proximity approaches in Information Retrieval (IR) distinguish the term pairs in the documents by rewarding the documents where the query terms occurring closer to each other, such as $CRTER_2$ proposed in Chapter 3. In this chapter, I propose that there is also a need to distinguish the term pairs in the queries. Contextual relevance of term proximity is defined to represent how much a term pair is related to the search intent of the query with respect to term proximity. In the experiments, the new context-sensitive approaches can further enhance $CRTER_2$. These context-sensitive approaches can also be applied in other proximity models.

Section 5.1 shows the motivation to incorporate the context-sensitivity in proximity models. In Section 5.2, we present the concept for Contextual Relevance of Term Proximity, and introduce four measures for estimation. In Section 5.3, we propose an enhanced context-sensitive proximity model using the proposed measures. Section

5.4 presents the experimental results and parameter sensitivities.

5.1 Motivation

The study of how to integrate the context information of queries and documents into retrieval process draw a lot of attention in recent years (Croft et al. 2010). More specifically, many term proximity approaches (Büttcher et al. 2006, Tao and Zhai 2007, Zhao et al. 2011, 2014), which reward the documents where the query terms occurring closer to each other, show significant improvements over basic Information Retrieval models. In these proximity-based approaches, all the query term pairs are usually treated equally and the difference among various query pairs are not considered, although how much the proximity of a term pair is related to the search intent of the query should be considered. For example, given a query "mickey mouse movie", there is a stronger association between "mickey" and "mouse" than the association between "mickey" and "movie", or "mouse" and "movie".

To evaluate such relevance, we assume that the distance between the occurrences of a pair of terms in the relevant documents could reflect the contextual relevance of term proximity. For example, Figure 5.1 shows two top ranked documents from google for the query "mickey mouse movie". We can see that "mickey" and "mouse" occur closer in these documents than "mickey" and "movie", or "mouse"

Figure 5.1: An Example: Top ranked documents returned for the query "mickey mouse movie"

and "movie". It indicates that the distance of terms in the relevant documents can distinguish the importance of term pairs. Further, we assume that the top ranked documents from a good ranking model can be regarded as relevant documents (usually recognized as a truth in relevance feedback).

In this chapter, we focus on the problem of differentiating the influence of associated query term pairs. We propose a proximity enhancement approach to integrate the contextual relevance of term proximity into the retrieval process. We also propose four measures for estimating the contextual relevance of term proximity, and reward the query term pairs according to both the proximity and the contextual relevance of proximity. There are several studies that boost proximity retrieval models. A machine learning method is proposed to determine the "goodness" of a span in (Svore

et al. 2010). (Bendersky et al. 2010) learns the concept importance from several sources (e.g. google n-gram corpus, query logs and wikipedia titles). SVM is used to learn different weights for various term dependencies in (Shi and Nie 2010). The importance of the global statistics is examined for proximity weighting (Macdonald and Ounis 2010). Phrases are treated as providing a context for the component query terms in (Song et al. 2008). The contribution of this chapter is that we propose the contextual relevance of term proximity, which represents to what extent the corresponding term pair should be related to the topic of the query. The contextual relevancy of term proximity is combined with the value of term proximity to characterize how much a document should be boosted.

5.2 Contextual Relevance of Term Proximity

In this section, we propose how to estimate the contextual relevance of term proximity. The contextual relevance of term proximity is defined as how much the corresponding term pair should be related to the topic of the query in the context. We measure the contextual relevance of term proximity based on the assumption that distributions of q_i and q_j in a relevant documents can represent the association between q_i and q_j. If q_i and q_j occur closely in relevant documents, the contextual relevance of term proximity between q_i and q_j is high. On the contrary, if q_i and q_j

do not co-occur or occur far away to each other, the contextual relevance of term proximity between q_i and q_j is low. Therefore we propose the following four methods for estimating the contextual relevance of term proximity between q_i and q_j in a relevant document D. For the extreme case when q_i and q_j do not co-occur in D, we consider the contextual relevance of term proximity equals 0. Otherwise, we define the following measures to generate a positive value for the contextual relevance of term proximity.

Definition 15. *$Rel_{CoOccur}(q_i, q_j, D)$ is defined to be 1, if q_i and q_j both occur in D.*

$$Rel_{CoOccur}(q_i, q_j, D) = \mathbf{1}_{\{q_i \in D \wedge q_j \in D\}} \tag{5.1}$$

Definition 16. *The $Rel_{SqRecip}$ is defined as the sum of squared reciprocal distances between q_i and q_j.*

$$Rel_{SqRecip}(q_i, q_j, D) = \sum_{k_1=1}^{tf(q_i,D)} \sum_{k_2=1}^{tf(q_j,D)} \frac{1}{dist(pos_{k_1,i}, pos_{k_2,j})^2} \tag{5.2}$$

Definition 17. *The $Rel_{MinDist}$ is a defined as the following function of the minimum distance between q_i and q_j.*

$$Rel_{MinDist}(q_i, q_j, D) = ln(\alpha + e^{-MinDist(q_i,q_j,D)}) \tag{5.3}$$

where α is a parameter, and $MinDist(q_i, q_j, D)$ is the minimum distance between all co-occurring q_i and q_j in D.

$$MinDist(q_i, q_j, D) = min_{k_1 \in \{1..tf(q_i,D)\}, k_2 \in \{1..tf(q_j,D)\}}(dist(pos_{k_1,i}, pos_{k_2,j}))$$

Definition 18. *The Rel_{Kernel} is defined as the sum of the kernel functions of distances between q_i and q_j.*

$$Rel_{Kernel}(q_i, q_j, D) = \sum_{k_1=1}^{tf(q_i,D)} \sum_{k_2=1}^{tf(q_j,D)} Kernel(\frac{1}{2} dist(pos_{k_1,i}, pos_{k_2,j})) \tag{5.4}$$

where $Kernel(\cdot)$ is kernel function. Here we use the triangle kernel function.

$$Kernel(u) = (1 - \frac{u}{\sigma}) \cdot \mathbf{1}_{\{u \leq \sigma\}} \tag{5.5}$$

where u is an input value, and σ is the kernel parameter.

These functions measure the contextual relevance from different perspectives. $Rel_{CoOccur}$ measures whether q_i and q_j are co-occurring in D. $Rel_{SqRecip}$, $Rel_{MinDist}$ and Rel_{Kernel} considers the positions of q_i and q_j in D. In $Rel_{SqRecip}$, we generate a squared reciprocal function for the distances between all the occurrences of q_i and q_j, and accumulate the values over D. Then the query term pairs with terms occurring

	q_1, q_2	q_1, q_3	q_1, q_4	q_1, q_5	q_2, q_3	q_2, q_4	q_2, q_5	q_3, q_4	q_3, q_5	q_4, q_5
$Rel_{CoOccur}$	1.0000	1.0000	1.0000	0.0000	1.0000	1.0000	0.0000	1.0000	0.0000	0.0000
$Rel_{SqRecip}$	0.2331	0.0408	0.0434	0.0000	0.0523	0.0434	0.0000	1.0000	0.0000	0.0000
$Rel_{MinDist}$	0.0486	0.0009	0.0025	0.0000	0.0067	0.0025	0.0000	0.3133	0.0000	0.0000
Rel_{Kernel}	1.3200	0.7200	0.7200	0.0000	0.7200	0.7200	0.0000	0.4800	0.0000	0.0000

Table 5.1: An example of the contextual relevance of term proximity

closer to each other and/or occurring more frequently will have higher contextual relevance. $Rel_{MinDist}$ is modified from (Tao and Zhai 2007), where the minimum distance is shown to be more effective than the other distance-based and span-based proximity approaches. Rel_{Kernel} utilizes the term proximity approach proposed in (Zhao et al. 2011), where a query term is simulated by the kernel function, where the triangle kernel function is recognized to be the most effective. Different types of information are incorporated in these measures.

To better analyze the contextual relevance measurements defined above, we present an example for a given query $Q = \{q_1, q_2, q_3, q_4, q_5\}$ and a relevant document D.

$$D = \{xq_1xq_2xxxxq_3q_4xxxxxq_1xq_2\}$$

where x represents a non-query term. By observing the query and the document, we find that the term q_5 does not present in D, which means it does not related to D, and therefore do not have an association with other query terms. Since q_3 and q_4 are adjacent to each other and far apart from other query terms, q_3 and q_4 are more

likely to have a stronger association than the combination of q_2 and q_4. We calculate the contextual relevance of term proximity between q_2 and q_4 as an instance, and the procedure will be the same for the rest of the query term pairs. The term frequency of terms q_2 and q_4 are $tf(q_2, D) = 2$ and $tf(q_4, D) = 1$. The positions of q_2 and q_4 in D are $pos(q_2, D) = \{4, 18\}$ and $pos(q_4, D) = \{10\}$ correspondingly. Therefore there are 2 co-occurrences of q_2 and q_4, and the corresponding distances between these co-occurrences are $\{6, 8\}$. Then we can calculate $Rel(q_2, q_4, D)$ with these distances by formulae (5.1-5.4). Table 5.1 shows the values of the contextual relevance of term proximity in this example.

We can see that the contextual relevance measures defined above show different characteristics. $Rel_{CoOccur}$ detects term pairs with or without an association. For example, q_5 and other query terms do not have an association. The term pairs containing q_5 are distinguished by $Rel_{CoOccur}$. On the other hand, $Rel_{CoOccur}$ does not consider the term distributions in D. $Rel_{MinDist}$ takes into account the closest occurrences between a pair of quay terms, and does not consider the frequency of occurrences. $Rel_{MinDist}$ and Rel_{Kernel} accumulates over all the occurrences of two query terms, with different functions and therefore generates different values.

5.3 A Context-Sensitive Proximity Model

In this section, we propose a context-sensitive proximity retrieval model, by integrating the proposed measures for contextual relevance of term proximity into retrieval process. Naturally we treat the values of contextual relevance as weights to reward the query term pairs with higher contextual relevance and to penalize the query term pairs with lower contextual relevance. In practice, we assume the top ranked documents returned by a basic retrieval model (for example, BM25) are more relevant than the rest of the documents. The averaged contextual relevance of term proximity over the top ranked documents is multiplied by the proximity part in the weighting function. A general form of the context-sensitive proximity model is

$$RelProx(D') = (1-\delta) * \Sigma_{q_i} w(q_i, D') + \delta * \Sigma_{q_i,q_j} AR(q_i, q_j, topDoc) * Prox(q_i, q_j, D') \tag{5.6}$$

where D' is a given document, $w(q_i, D')$ is the weight of q_i in D' by a basic probabilistic weighting function, $Prox(q_i, q_j, D')$ is a bigram proximity weighting approach, λ is a balancing parameter, $topDoc$ is the number of top ranked documents, $AR(q_i, q_j, topDoc)$ is the average contextual relevance value of term proximity between q_i and q_j over the top ranked documents

$$AR(q_i, q_j, topDoc) = \frac{1}{topDoc} \sum_{D} Rel(q_i, q_j, D) \tag{5.7}$$

where $Rel(q_i, q_j, D)$ is one of the measures defined in Section 2. Please note that $AR(q_i, q_j, topDoc)$, $Prox(q_i, q_j, D')$ and $w(q_i, D')$ need to be normalized to the same scale.

In formula (5.6), we use the probabilistic BM25 (Robertson et al. 1995) as the basic weighting function. We adopt the proximity approach used in $CRTER_2$. The BM25 weighting function has the following form.

$$w(q_i, D') = \frac{(k_1 + 1) * tf(q_i, D')}{K + tf(q_i, D')} * \frac{(k_3 + 1) * qtf(q_i)}{k_3 + qtf(q_i)} * \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \tag{5.8}$$

where N is the number of documents in the collection, $n(q_i)$ is the number of documents containing q_i, $qtf(q_i)$ is the within-query term frequency, $dl(D')$ is the length D', $avdl$ is the average document length, the k_is are tuning constants, K equals $k_1 * ((1 - b) + b * dl(D)/avdl)$.

The proximity part of $CRTER_2$ can be written as follows.

$$Pro_(q_i, q_j, D') = w(q_{i,j}, D') \tag{5.9}$$

where $q_{i,j}$ represents the association between query terms q_i and q_j, $w(q_{i,j}, D')$ is the

BM25 weighting function with the following features of $q_{i,j}$

$$tf(q_{i,j}, D') = \sum_{k_1=1}^{tf(q_i,D')} \sum_{k_2=1}^{tf(q_j,D')} Kernel(\frac{1}{2} dist(pos_{k_1,i}, pos_{k_2,j}))$$

$$qtf(q_{i,j}) = Kernel(\frac{1}{2}) \cdot min(qtf(q_i), qtf(q_j))$$

$$n(q_{i,j}) = \sum_{D'} \frac{tf(q_{i,j}, D')}{Occur(q_{i,j}, D')}$$

where $Kernel(\cdot)$ is a kernel function, and $Occur(q_{i,j}, D')$ is estimated as

$$Occur(q_{i,j}, D') = \sum_{k_1=1}^{tf_i} \sum_{k_2=1}^{tf_j} \mathbf{1}_{\{Kernel(\frac{1}{2} dist(pos_{k_1,i}, pos_{k_2,j}))\neq 0\}}$$

.

5.4 Experiments

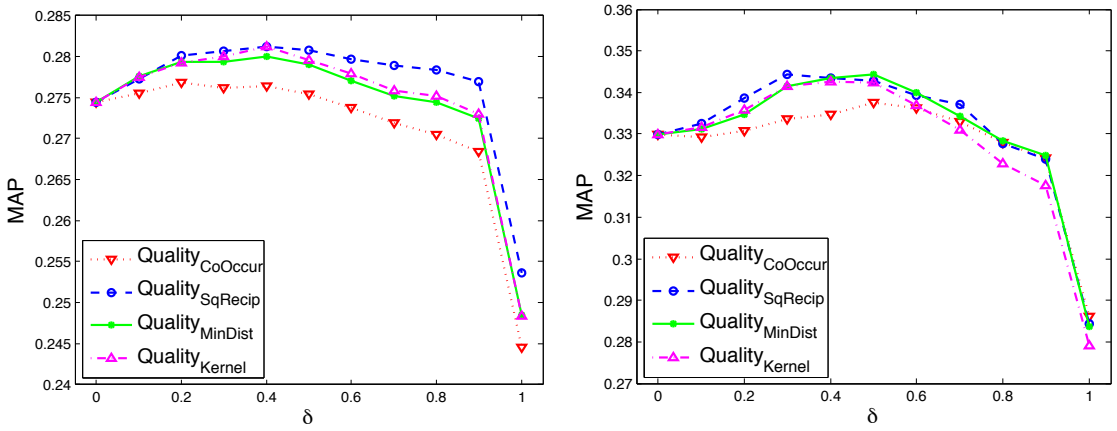
We evaluate the proposed approach on three standard TREC data sets. They are AP88-89 with topics 51-100, Web2G with topics 401-450, and TREC8 with topics 401-450. AP88-89 contains articles published by Association Press from the year of 1988 to 1989. For all the data sets used, each term is stemmed using Porter's English stemmer, and standard English stopwords are removed. We have three baseline models, BM25, Dirichlet Language Model (LM) and $CRTER_2$. The best parameters are chosen in the baseline models for fair comparisons. In BM25, the values of k_1, k_2, k_3 and b are set to be 1.2, 0, 8 and 0.35 respectively, since they are

recognized with a good performance. In $CRTER_2$ model, we use the recommended settings (Zhao et al. 2011)., which are $\sigma = 25$, $\lambda = 0.2$, and triangle kernel function. In our proposed context-sensitive proximity model, we use the same parameters in the basic weighting model part (e.g. BM25) and the proximity part (e.g. $CRTER_2$). In Rel_{Kernel}, we set the kernel parameter $\sigma = 25$. In $Rel_{MinDist}$, we set $\alpha = 1$, which has the best performance in (Tao and Zhai 2007). We normalize $AR(q_i, q_j, topDoc)$, $Prox(q_i, q_j, D')$ and $w(q_i, D')$ in formula (5.6) to the scale of [0,1]. We use the Mean Average Precision (MAP) as our evaluation metric.

Data Sets	AP88-89	Web2G	TREC8
BM25	0.2708	0.3136	0.2467
Dirichlet LM	0.2763	0.3060	0.2552
$CRTER_2$	0.2744	0.3298*	0.2606 *
Improvement over BM25	1.329%	5.166%	5.634%
$Rel_{CoOccur}$	0.2768	0.3375*	0.2622*
Improvement over BM25	2.216%	7.621%	6.283%
Improvement over $CRTER_2$	0.875%	2.335%	0.614%
$Rel_{SqRecip}$	**0.2812***‡	**0.3444*** ‡	**0.2633***
Improvement over BM25	3.840%	9.821%	6.729%
Improvement over $CRTER_2$	2.478%	4.427%	1.036%
$Rel_{MinDist}$	0.2800	**0.3444*** ‡	0.2615*
Improvement over BM25	3.397%	9.821%	5.999%
Improvement over $CRTER_2$	2.041%	4.427%	0.345%
Rel_{Kernel}	**0.2812***‡	0.3425*‡	0.2625*
Improvement over BM25	3.840%	9.216%	6.405%
Improvement over $CRTER_2$	2.478%	3.851%	0.729%

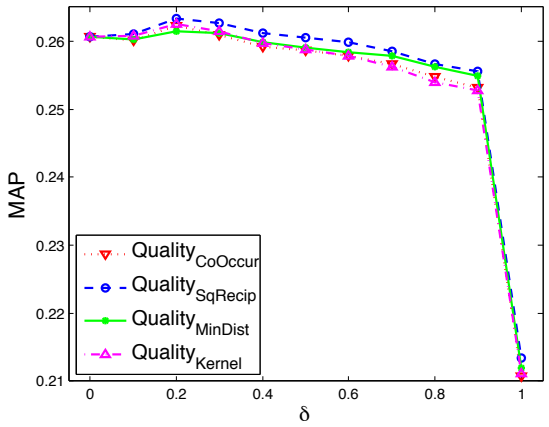
Table 5.2: Overall MAP Performance ("*" indicates significant improvement over BM25, and "‡" indicates significant improvement over $CRTER_2$)

Table 5.2 shows the overall MAP performance. The proposed context-sensitive proximity model outperforms BM25, Language Model (LM) and $CRTER_2$ with all



(a) AP88-89

(b) Web2G



(c) TREC8

Figure 5.2: Sensitivity of δ

		5	10	20	30	40	50	60	70	80
AP88-89	$Rel_{CoOccur}$	0.2757	0.2766	**0.2768**	0.2767	0.2762	0.2759	0.2758	0.2755	0.2753
	$Rel_{SqRecip}$	0.281	**0.2812**	0.2801	0.2801	0.2789	0.2781	0.278	0.2777	0.2774
	$Rel_{MinDist}$	**0.2800**	0.2794	**0.2800**	0.2796	0.2797	0.2784	0.279	0.2787	0.2786
	Rel_{Kernel}	**0.2812**	0.2796	0.2801	0.2801	0.2789	0.2781	0.2783	0.2781	0.278
Web2G	$Rel_{CoOccur}$	0.3348	0.3359	0.3354	0.3367	**0.3375**	0.3369	0.3367	0.3374	0.3363
	$Rel_{SqRecip}$	0.3401	0.342	0.3409	0.3401	**0.3444**	0.3435	0.3433	0.3424	0.3419
	$Rel_{MinDist}$	0.3351	**0.3444**	0.3421	0.3382	0.3406	0.3414	0.3427	0.3434	0.3433
	Rel_{Kernel}	0.3358	0.3409	**0.3425**	0.3406	0.3423	0.3418	0.3421	0.3414	0.3399
TREC8	$Rel_{CoOccur}$	**0.2622**	0.2612	0.2614	0.2611	0.2611	0.261	0.2608	0.2601	0.2599
	$Rel_{SqRecip}$	**0.2633**	0.263	0.2627	0.2623	0.2617	0.2615	0.2613	0.2614	0.2612
	$Rel_{MinDist}$	0.2612	0.2614	**0.2615**	0.2612	0.2606	0.2605	0.2606	0.2607	0.2605
	Rel_{Kernel}	**0.2625**	0.2622	0.2619	0.2615	0.2607	0.2604	0.2605	0.2605	0.2606

Table 5.3: Performance over the change of *topDoc*

of the contextual relevance measuring approaches on all the data sets. For the space limitation, we only include these comparisons. It shows that using the contextual relevance of term proximity can further boost the retrieval performance. We can see that the $Rel_{CoOccur}$, which measures whether two query terms are co-occurring in the relevant documents, reaches the lowest MAP among the contextual relevance measures, which indicates the necessity of considering the term location information in the term pair contextual relevance definition. $Rel_{SqRecip}$ has the highest MAP over the other approaches on all the data sets. In general, considering both the closeness and frequency of two query terms in the contextual relevance definition benefits the contextual relevance estimation.

In Table 5.3, we investigate how the number of top relevant documents affects the retrieval performance. We take the *topDoc* = 5, 10, 20, 30, 40, 50 60, 70, and

80 documents as relevant, and calculate the average contextual relevance obtained from these documents. The bolded values are the best performance among different *topDoc* values. We can see that the best *topDoc* will be around 5 to 40. It means that selecting too many top documents as relevant will introduce noises to the model.

Figure 5.2 shows the sensitivity of δ on all the data sets. We can see that with the growth of δ, MAP first increases and then decreases. Please note that when $\delta = 0$, there is no proximity utilized, which is our baseline BM25. When $\delta = 1$, only term proximity and the contextual relevance of term proximity are considered. In $CRTER_2$, the recommended setting for the balancing parameter is 0.2. After introducing the contextual relevance of term proximity, we can see that the balancing parameter δ with a value of around 0.3 or 0.4 is better. The reason is that the contextual relevance is normalized to [0,1]. The value for the second part of formula (5.6) becomes smaller, therefore it requires a larger balancing parameter.

The proposed model (Zhao et al. 2014) provides a way to further enhance the proposed CRTER models, and it can also be applied for other proximity-based IR models.

6 Summaries

6.1 Practical Impact of the Proposed Approaches

In this section, we discuss the practical impact of the proposed approaches, and briefly introduce how other various domains can benefit from the proposed models. The exact models used might need to adapt to the specific problems in these domains.

The proposed concept of Cross Term can be useful in other text analysis realms. It is natural to study term associations in document clustering, document summarization, sentiment analysis, etc. With properly defined Cross Terms, we can visualize and quantify these term associations for more precise results. For example, document clustering aims to combine a set of documents into clusters so that intra cluster documents are more similar to each other than inter cluster documents (Agrawal and Phatak 2013). Thus we can measure the similarity between documents by matching the real terms and the Cross Terms. As another example, document summarization can represent the document with a short piece of text covering the main topics (He

et al. 2012). We can extract the most important Cross Terms in the original document and try to keep these Cross Terms in the summarization. Cross Terms can also be employed in sentiment analysis, where a review can be considered as a document generated by a number of hidden sentiment factors (Yu et al. 2012). We can use the idea of Cross Terms to represent the hidden sentiment information in reviews. We can also apply Cross Terms in the document distribution study in IR (Zhao et al. 2014).

The generalized concept of Cross Term can be applied in more domains such the medical data mining and image processing. Medical examinations have associations among each other (Zhao et al. 2013, 2012). For instance, if a patient is suspected to have diabetes, usually the doctor will assign both a Hemoglobin test and a Glucose Fasting test for this patient. There exists an association between Hemoglobin and Glucose Fasting with respect to some hidden information, diabetes in this case. We can use the idea of Cross Terms to simulate this hidden information among medical examinations/treatments. In image retrieval, it is usually difficult to match the actions in the images and researchers often integrate other text features as alternatives (Ye et al. 2010). The object boundaries and shapes can be detected by the approaches proposed in (Ferrari et al. 2007). The image objects at the form level and the content level can be represented by a logic-based model (Meghini et al. 1997). We

can use the idea of Cross Terms and kernel functions to investigate the correlations among the image objects and to identify the object actions. We can also investigate the associations among patterns discovered from the data (An et al. 2011, Rohian et al. 2009), and the associations among the chemical compounds (Lupu et al. 2011).

6.2 Summary of Using Cross Terms

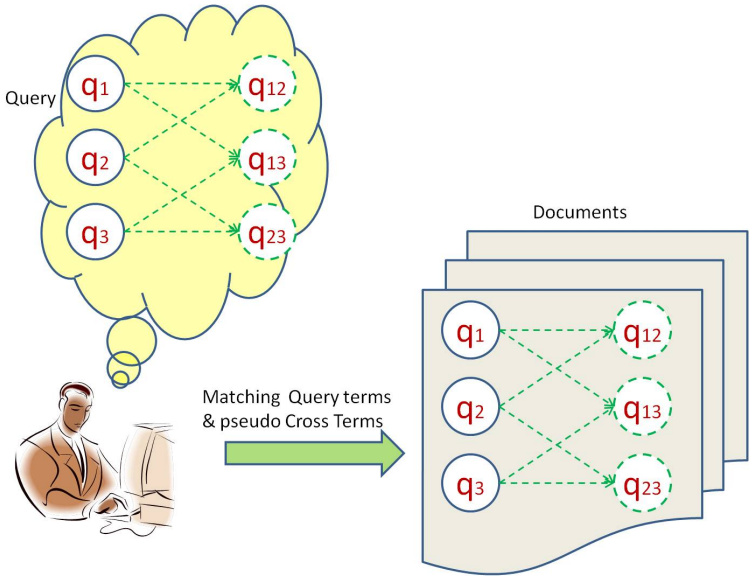


Figure 6.1: Hidden information of using Cross Terms

In this thesis, we focus on integrating proximity into the information retrieval process, which is similar to the other proximity models to some extent. However, the details vary significantly. We work on this problem from a new and novel perspective

that is unique and different from previous methods. In detail, we abstract the concept of term association into a pseudo term, Cross Term, which does not exist but is very important for modelling term associations. In a user's query, there exists some hidden information underlying the query terms. The proposed concept of Cross Terms provides a promising avenue for modeling and mining associations among terms in information retrieval. During retrieval, this model matches the query terms in the topic with the terms in the documents, and matches the Cross Terms in the topic with the Cross Terms in the documents, as shown in Figure 6.1. As we know, the most fundamental elements in IR are terms. Therefore, proposing the Cross Term concept as a pseudo term can make us easily simulate and manipulate term associations in retrieval models. Term associations are embedded in $CRTER$ models in a natural and global way. Compared to some other probabilistic models with proximity, we integrate proximity not only in the term level (changing some variables related to the individual terms), but also in the collection level (controlled by the number of documents containing a Cross Term, see Definition 2.5 and 2.8). We implement the Cross Terms using proximity, which defines and estimates the Cross Terms properly.

What is more, the Cross Term concept could be applied to other domains. These include: document clustering, document summarization, sentiment analysis and the

medical data mining. The exact models used might have to be modified to accommodate the specific characteristics of the target domains.

7 Conclusions and Future Work

In this thesis, I propose several models to integrate the associations among multiple query terms in retrieval models. They are bigram Cross TErm Retrieval model ($CRTER_2$) as the basis model for two query terms, n-gram Cross TErm Retrieval model ($CRTER_n$) for n query terms where n > 2, a Language model based model$CRTER_2^{LM}$, and an enhanced context-sensitive proximity model. A Cross Term in this thesis is a newly proposed pseudo term that is generated by two or more query terms occurring close to each other. Through extensive experiments on a relatively large set of representative TREC collections with various kernel functions, we show that the proposed models significantly outperform BM25 and Dirichlet LM baselines. By comparison, we conclude that the proposed $CRTER_2$ model is at least comparable to, if not better than, the state-of-the-art proximity models, BM25TP, MRF, PLM and PPM with optimal parameter settings. In more detail, we discuss the parameter settings and the proposed models are shown to be robust. A group of optimal parameters demonstrate the robustness of the $CRTER_2$ model on all collec-

tions used. We also experiment with the n-gram setting ($CRTER_n$), and find that $CRTER_n$ is a useful addition to $CRTER_2$. Considering the context information of Cross Terms can also boost the proposed $CRTER$ models.

As we know, the most fundamental elements in IR are terms. Therefore, proposing the concept of Cross Term as a pseudo term can allow us to simulate and manipulate term associations in any retrieval model. The implementation of Cross Terms using kernel functions provides a proper way to define and estimate Cross Terms.

The implementation for Cross Terms can be extend with Natural Language Processing (NLP) techniques in the future in the following ways. (1) The use of domain ontologies can be applied for some specific types of term associations to generate some specific Cross Terms. (2) We can also do sentence boundary disambiguation, when defining the Cross Terms' variants, i.e. put constraints on the term impact function. (3) The word sense disambiguation techniques can be implemented for Cross Terms. For example, if the Cross Term as multiple meanings, then select the meaning which makes the most sense in the context. (4) We can also co-index different Cross Terms with the same references.

Considering that the learning to rank approaches for term associations treat term association scores as an attribute and learn a model based on multiple attributes, $CRTER$ models are more straightforward that the term associations are naturally

adapted into the existing IR models. In the future, our approaches can be integrated with the learning to rank approaches by treating Cross Term scores as an attribute in the learning to rank process.

The topic modelling approaches are usually based on co-occurrence, and do not show the term distributions in the documents. The topic modelling approaches have advantages in building a latent topic space that the documents have topic representations and the topics are generated by terms. In the future, we can implement Cross Terms with topic modelling approaches by regarding that the latent topics as Cross Terms. We can also further improve the topic modelling approaches by considering the term distribution information for the latent topics.

The concept of the Cross Term provides more possible future research directions. We can further study the Cross Term's distribution by examining other solutions to implement the Cross Term. It is also interesting to explore its usefulness in the process of query expansion when the association among the original query terms and the candidate expansion term is considered. Moreover, by further reducing the computational complexity of the model implementation, the proposed model would be more applicable for real applications.

Bibliography

[1] Agrawal, R. and Phatak, M. (2013). A novel algorithm for automatic document clustering. In *Advance Computing Conference (IACC), 2013 IEEE 3rd International*, pages 877–882. IEEE.

[2] Ahmed, F. and Nurnberger, A. (2009). Evaluation of n-gram conflation approaches for Arabic text retrieval. *Journal of the American Society for Information Science and Technology*, 60(7):1448–1465.

[3] Allan, J., Ballesteros, L., Callan, J., Croft, W., and Lu, Z. (1995). Recent experiments with inquery. In *Proceedings of the 4th Text Retrieval Conference*, pages 49–64.

[4] Alvarez, C., Langlais, P., and Nie, J. (2004). Word pairs in language modeling for information retrieval. In *Proceedings 7th International Conference on Computer Assisted Information Retrieval*, pages 686–705.

[5] An, A. Q., Wan, Zhao, J. and Huang X. J.(2009). Diverging Patterns: Discover-

ing Significant Frequency Change Dissimilarities in Large Databases. In *Proceedings of the 18th ACM International Conference on Information and Knowledge Management*, pages 1473-1476. ACM, 2009.

[6] Baeza-Yates, R., Ribeiro-Neto, B., et al. (1999). *Modern information retrieval*, volume 463. ACM press New York.

[7] Beaulieu, M., Gatford, M., Huang, X., Robertson, S., Walker, S., and Williams, P. (1997). Okapi at TREC-5. *In Proceedings of the 5th Text REtrieval Conference. NIST SPECIAL PUBLICATION SP*, pages 143–166.

[8] Beigbeder, M. and Mercier, A. (2005). An information retrieval model using the fuzzy proximity degree of term occurences. In *Proceedings of the 2005 ACM Symposium on Applied Computing*, pages 1018–1022. ACM New York, NY, USA.

[9] Bendersky, M. and Croft, W. B. (2008). Discovering key concepts in verbose queries. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 491–498. ACM.

[10] Bendersky, M. and Croft, W. B. (2012). Modeling higher-order term dependencies in information retrieval using query hypergraphs. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 941–950, New York, NY, USA. ACM.

[11] Bendersky, M., Croft, W. B., and Smith, D. A. (2009). Two-stage query segmentation for information retrieval. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 810–811. ACM.

[12] Bendersky, M., Metzler, D., and Croft, W. B. (2010). Learning concept importance using a weighted dependence model. In *Proc. of WSDM*, pages 31–40. ACM.

[13] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

[14] Broschart, A. and Schenkel, R. (2008). Proximity-aware scoring for XML retrieval. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 845–846. ACM New York, NY, USA.

[15] Büttcher, S., Clarke, C., and Lushman, B. (2006). Term proximity scoring for ad-hoc retrieval on very large text collections. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 621–622. ACM New York, NY, USA.

[16] Callan, J., Croft, W., and Harding, S. (1992). The INQUERY retrieval system. In

Proceedings of the 3rd International Conference on Database and Expert Systems Applications, pages 78–83.

[17] Cormack, G. V., Smucker, M. D., and Clarke, C. L. (2011). Efficient and effective spam filtering and re-ranking for large web datasets. *Information retrieval*, 14(5):441–465.

[18] Croft, W., Metzler, D., and Strohman, T. (2010). *Search engines: Information retrieval in practice*. Addison-Wesley.

[19] de Kretser, O. and Moffat, A. (1999). Effective document presentation with a locality-based similarity heuristic. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 113–120. ACM.

[20] Fan, Y., Huang, X., and An, A. (2006). York University at TREC 2006: Enterprise Email Discussion Search. In *Proceedings of the 5th Text REtrieval Conference*, volume Special Publication 500-272.

[21] Ferrari, V., Jurie, F., and Schmid, C. (2007). Accurate object detection with deformable shape models learnt from images. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE.

[22] Fuhr, N. (1992). Probabilistic models in information retrieval. *The Computer Journal*, 35(3):243–255.

[23] Gao, J., Nie, J., Wu, G., and Cao, G. (2004). Dependence language model for information retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 170–177. ACM New York, NY, USA.

[24] Goker, A. and Davies, J. (2009). *Information retrieval: searching in the 21st century*. Wiley. com.

[25] Guo, J., Xu, G., Li, H., and Cheng, X. (2008). A unified and discriminative model for query refinement. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 379–386. ACM.

[26] Rohian, H., An, A., Zhao, J., and Huang, J. X. (2009). A bayesian-based prediction model for personalized medical health care. In *Bioinformatics and Biomedicine (BIBM), 2009 IEEE International Conference on*, pages 419-423. IEEE.

[27] Hawking, D. and Thistlewaite, P. (1995). Proximity operators - So near and yet so far. In *Proceedings of the 4th Text Retrieval Conference*, pages 131–143.

[28] He, B., Huang, J. X., and Zhou, X. (2011). Modeling term proximity for probabilistic information retrieval models. *Information Sciences Journal*, 181(14):3017–3031.

[29] He, Z., Chen, C., Bu, J., Wang, C., Zhang, L., Cai, D., and He, X. (2012). Document summarization based on data reconstruction. In *AAAI Conference on Artificial Intelligence*, pages 620–626. IEEE.

[30] Hoffman, M., Bach, F. R., and Blei, D. M. (2010). Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pages 856–864.

[31] Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM.

[32] Hou, Y., Zhao, X., Song, D., and Li, W. (2013). Mining pure high-order word associations via information geometry for information retrieval. *ACM Transactions on Information Systems (TOIS)*, 31.

[33] Hsiao, D. and Harary, F. (1970). A formal system for information retrieval from files. *Communications of the ACM*, 13(2):67–73.

[34] Hu, Q., Huang, J., and Hu, X. (2012). Modeling and mining term association for improving biomedical information retrieval performance. *BMC Bioinformatics*, 13(9):1–18.

[35] Huang, J. X., Miao, J., and He, B. (2013). High performance query expansion using adaptive co-training. *Information Processing & Management*, 49(2):441–453.

[36] Huang, X. and Hu, Q. (2009). A bayesian learning approach to promoting diversity in ranking for biomedical information retrieval. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieva*, pages 307–314. ACM.

[37] Huang, X., Huang, Y. R., Wen, M., An, A., Liu, Y., and Poon, J. (2006a). Applying data mining to pseudo-relevance feedback for high performance text retrieval. In *Proceedings of the 6th IEEE International Conference on Data Mining*, pages 295–306. IEEE Computer Society.

[38] Huang, X., Wen, M., An, A., and Huang, Y. R. (2006b). A platform for okapi-based contextual information retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 728. ACM.

[39] Huang, X., Zhong, M., and Si, L. (2005). York University at TREC 2005: Genomics Track. In *Proceedings of the 14th Text REtrieval Conference*.

[40] Jones, K. S. and Jackson, D. M. (1970). The use of automatically-obtained keyword

classifications for information retrieval. *Information Storage and Retrieval*, 5(4):175–201.

[41] Kise, K., Junker, M., Dengel, A., and Matsumoto, K. (2004) Passage retrieval based on density distributions of terms and its applications to document retrieval and question answering. *Reading and Learning, Adaptive Content Recognition 2004. Lecture Notes in Computer Science*, 2956:306–327.

[42] Kurland, O. and Lee, L. (2004). Corpus structure, language models, and ad hoc information retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 194–201. ACM.

[43] Lavrenko, V. and Croft, W. B. (2001). Relevance based language models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 120–127. ACM.

[44] Lupu, M., Zhao, J., Huang, J., Gurulingappa, H., Fluck, J., Zimmermann, M., Filippovm, V. I. and Tait, J. (2011). Overview of the TREC 2011 Chemical IR Track. In *Proceedings of the 20th Text Retrieval Conference*.

[45] Lv, Y. and Zhai, C. (2009). Positional language models for information retrieval.

In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 299–306. ACM New York, NY, USA.

[46] Macdonald, C. and Ounis, I. (2010). Global statistics in proximity weighting models. In *Web N-gram Workshop*, pages 30–37.

[47] Maron, M. E. and Kuhns, J. L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the ACM (JACM)*, 7(3):216–244.

[48] Mayfield, J. and McNamee, P. (2003). Single n-gram stemming. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 415–416. ACM New York, NY, USA.

[49] McNamee, P. and Mayfield, J. (2004). Character n-gram tokenization for European language text retrieval. *Information Retrieval*, 7(1):73–97.

[50] Meghini, C., Sebastiani, F., and Straccia, U. (1997). The terminological image retrieval model. In *Image Analysis and Processing*, pages 156–163. Springer.

[51] Mercier, A. and Beigbeder, M. (2005). Fuzzy proximity ranking with boolean queries. In *Proceedings of the 5th Text REtrieval Conference*.

[52] Metzler, D. and Croft, W. (2005). A Markov random field model for term dependencies. In *Proceedings of the 28th Annual International ACM SIGIR Conference*

on Research and Development in Information Retrieval, pages 472–479. ACM New York, NY, USA.

[53] Miao, J., Huang, J. X., and Ye, Z. (2012). Proximity-based rocchio's model for pseudo relevance. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 535–544.

[54] Nassar, M. O., Al Mashagba, F. F., and Al Mashagba, E. F. (2013). Investigating genetic algorithms to optimize the user query in the vector space model. *Australian Journal of Basic and Applied Sciences*, 7(2):47–53.

[55] Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., and Lioma, C. (2006a). Terrier: A high performance and scalable information retrieval platform. In *Proceedings of the OSIR Workshop*, pages 18–25.

[56] Ounis, I., De Rijke, M., Macdonald, C., Mishne, G., and Soboroff, I. (2006b). Overview of the TREC-2006 Blog track. In *Proceedings of the 15th Text REtrieval Conference*.

[57] Ounis, I., Lioma, C., Macdonald, C., and Plachouras, V. (2007). Research directions in terrier: a search engine for advanced retrieval on the web. *CEPIS Upgrade Journal*, 8(1).

[58] Park, J. H., Croft, W. B., and Smith, D. A. (2011). A quasi-synchronous dependence model for information retrieval. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 17–26. ACM.

[59] Petkova, D. and Croft, W. (2007). Proximity-based document representation for named entity retrieval. In *Proceedings of the sixteenth ACM Conference on Information and Knowledge Management*, pages 731–740. ACM New York, NY, USA.

[60] Pickens, J. (2000). A comparison of language modeling and probabilistic text information retrieval approaches to monophonic music retrieval. In *International Symposium on Music Information Retrieval*, page 11 pages. Plymouth, Massachusetts, USA.

[61] Ponte, J. M. and Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 275–281. ACM.

[62] Qiu, Y. and Frei, H.-P. (1993). Concept based query expansion. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 160–169. ACM.

[63] Rasolofo, Y. and Savoy, J. (2003). Term proximity scoring for keyword-based retrieval systems. *Lecture Notes in Computer Science*, pages 207–218.

[64] Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M., and Gatford, M. (1996). Okapi at TREC-4. In *Proceedings of the 4th Text Retrieval Conference*, pages 73–97.

[65] Robertson, S. E. and Walker, S. (1994). Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 232–241, New York, NY, USA.

[66] Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., Gatford, M., et al. (1995). Okapi at TREC-3. In *Proc. of TREC*, pages 109–126. NIST.

[67] Salton, G., Wong, A., and Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

[68] Salton, G. and Yang, C.-S. (1973). On the specification of term values in automatic indexing. *Journal of documentation*, 29(4):351–372.

[69] Shi, L. and Nie, J.-Y. (2010). Using various term dependencies according to their utilities. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 1493–1496. ACM.

[70] Song, F. and Croft, W. (1999). A general language model for information retrieval. In *Proceedings of the eighth International Conference on Information and Knowledge Management*, pages 316–321. ACM New York, NY, USA.

[71] Song, R., Taylor, M. J., Wen, J.-R., Hon, H.-W., and Yu, Y. (2008). Viewing term proximity from a different perspective. In *Proc. of ECIR*, pages 346–357. Springer.

[72] Song, R., Yu, L., Wen, J.-R., and Hon, H.-W. (2011). A proximity probabilistic model for information retrieval. In *Microsoft Technical Report*, page 8 pages. Microsoft Research.

[73] Srikanth, M. and Srihari, R. (2002). Biterm language models for document retrieval. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 425–426. ACM New York, NY, USA.

[74] Strohman, T., Metzler, D., Turtle, H., and Croft, W. B. (2005). Indri: A language model-based search engine for complex queries. In *Proceedings of the International Conference on Intelligent Analysis*, volume 2, pages 2–6.

[75] Svore, K. M., Kanani, P. H., and Khan, N. (2010). How good is a span of terms? exploiting proximity to improve web retrieval. In *Proc. of SIGIR*, pages 154–161. ACM.

[76] Tao, T. and Zhai, C. (2007). An exploration of proximity measures in information retrieval. In *Proceedings of the 30th Annual International ACM SIGIR Conference*

on Research and Development in Information Retrieval, pages 295–302. ACM New York, NY, USA.

[77] Voorhees, E. and Harman, D. (2005). *TREC: Experiment and evaluation in information retrieval*, volume 32. MIT Press.

[78] Wang, J. and Zhu, J. (2009). Portfolio theory of information retrieval. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 115–122. ACM New York, NY, USA.

[79] Wang, M. and Si, L. (2008). Discriminative probabilistic models for passage based retrieval. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 419–426. ACM.

[80] Wei, X. and Croft, W. B. (2007). *Modeling term associations for ad-hoc retrieval performance within language modeling framework*. Springer.

[81] Wong, S. K. M. and Yao, Y. Y. (1993). A probabilistic method for computing term-by-term relationships. *Journal of the American Society for Information Science*, 44(8):431–439.

[82] Ye, Z., Huang, J. X., and Miao, J. (2012). A hybrid model for ad-hoc information retrieval. In *The 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1025–1026.

[83] Ye, Z., Huang, X., He, B., and Lin, H. (2009). York University at TREC 2009: Relevance Feedback Track. In *Proceedings of the 18th Text REtrieval Conference*.

[84] Ye, Z., Huang, X., Hu, Q., and Lin, H. (2010). An integrated approach for medical image retrieval through combining textual and visual features. In *Multilingual Information Access Evaluation II. Multimedia Experiments*, pages 195–202. Springer.

[85] Yin, X., Huang, J., Li, Z., and Zhou, X. (2013). A survival modeling approach to biomedical search result diversification using wikipedia. *IEEE Transactions on Knowledge and Data Engineering*, pages 1201–1212.

[86] Yu, X., Liu, Y., Huang, X., and An, A. (2012). Mining online reviews for predicting sales performance: A case study in the movie domain. *IEEE on Knowledge and Data Engineering*, 24(4):720–734.

[87] Zhai, C. and Lafferty, J. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 334–342. ACM New York, NY, USA.

[88] Zhai, C. and Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)*, 22(2):179–214.

[89] Zhao, J. and Huang, J. (2014). An Enhanced Context-sensitive Proximity Model for Probabilistic Information Retrieval. In *Proceeding of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1131-1134. ACM 2014.

[90] Zhao, J., Huang, J., and He, B. (2011). CRTER: using cross terms to enhance probabilistic information retrieval. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 155–164. ACM 2011.

[91] Zhao, J., Huang, J. X., and Hu, X. (2013). BPLT+: A Bayesian-based personalized recommendation model for health care. *BMC Genomics*, 14(Suppl 4):S6.

[92] Zhao, J., Huang, J. X., Hu, X., Kurian, J., and Melek, W. (2012). A bayesian-based prediction model for personalized medical health care. In *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*, pages 579–582. IEEE.

[93] Zhao, J., Huang, J., and Wu, S. (2012). Rewarding Term Lo- cation Information to Enhance Probabilistic Information Retrieval. In *Proceeding of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1137-1138. ACM 2012.

[94] Zhao, J., Huang, J. X., and Ye, Z. (2014). Modeling term associations for probabilistic information retrieval. *ACM Transactions on Information Systems (TOIS)*, 32(2):7.

[95] Zhao, J., Huang, X. and Ye, Z. (2010). York University at TREC 2010: Chemical track. In *Proceedings of the 19th Text Retrieval Conference*.

[96] Zhao, J., Huang, X., Ye, Z., and Zhu, J. (2009). York University at TREC 2009: Chemical track. In *Proceedings of the 18th Text Retrieval Conference*.

[97] Zhao, J. and Yun, Y. (2009). A proximity language model for information retrieval. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 291–298. ACM.

A Proof for Theorem 4.2.2

Theorem 4.2.2 *Altitude Based Distance and Hypotenuse Based Distance has the following difference. Suppose the span $p_N - p_1$ is fixed, the median positions affect Altitude Based Distance and Hypotenuse Based Distance differently: Altitude Based Distance has a higher value, while Hypotenuse Based Distance has lower value, if the median positions tend to spread evenly, i.e., $p_2 - p_1 = p_3 - p_2 = ... = p_N - p_{N-1}$.*

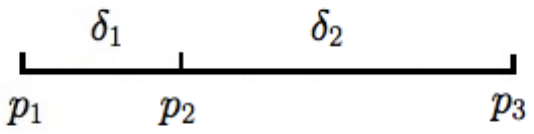


Figure A.1: Positions when $N = 3$

Proof. For the case that $N = 3$, we have three positions p_1, p_2, p_3 where $p_1 < p_2 < p_3$ as shown in Figure A.1. For convenience, we denote $|p_2 - p_1|$ as δ_1, $|p_3 - p_2|$ as δ_2, and the total span $|p_3 - p_1|$ as l_s. Suppose the spread p_1 and p_3 are fixed, and p_2 can move between p_1 and p_3. The Altitude distance is

$$
\begin{aligned}
dist_{alti}(p_1, p_2, p_3) &= \sqrt{\prod_{i=1,2} (p_{i+1} - p_i)} \\
&= \sqrt{\delta_1 \delta_2} \\
&= \sqrt{\delta_1 (l_s - \delta_1)} \\
&= \sqrt{-(\delta_1 - \frac{l_s}{2})^2 + \frac{l_s^2}{4}}
\end{aligned}
$$

Therefore, $dist_{alti}(p_1, p_2, p_3)$ has a higher value when δ_1 is closer to the $\frac{l_s}{2}$, i.e. p_2 positioned closer to the median point.

The Hypotenuse Based Distance is

$$
\begin{aligned}
dist_{alti}(p_1, p_2, p_3) &= \sqrt{\sum_{i=1,2} (p_{i+1} - p_i)^2} \\
&= \sqrt{\delta_1^2 + \delta_2^2} \\
&= \sqrt{\delta_1^2 + (l_s - \delta_1)^2} \\
&= \sqrt{2(\delta_1 - \frac{l_s}{2})^2 + \frac{l_s^2}{2}}
\end{aligned}
$$

Therefore, $dist_{hypo}(p_1, p_2, p_3)$ has a lower value when δ_1 is closer to the $\frac{l_s}{2}$, i.e. p_2 positioned closer to the median point.

For the case that $N \geq 4$, we have N positions $p_1, ..., p_N$ where $p_i < p_{i+1}$ for

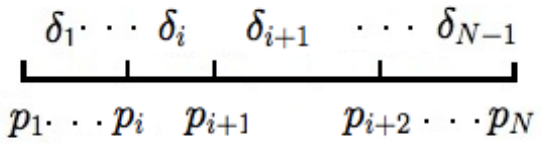


Figure A.2: Positions when $N \geq 4$

$1 \leq i \leq N-1$. We denote $p_{i+1} - p_i$ as δ_i for $1 \leq i \leq N-1$. We prove Altitude Based Distance has the maximum value when $\delta_1 = ... = \delta_{N-1}$ by contradiction. First we make an assumption that there exist δ_i and δ_{i+1} where $\delta_i \neq \delta_{i+1}$, such that $dist_{alti}(p_1, ..., p_N)$ has the maximum value. We denote $\delta_i + \delta_{i+1}$ as l'_s. In the following, we will prove that this assumption is false and the original statement is true.

$$
\begin{aligned}
dist_{alti}(p_1, ..., p_N) &= \sqrt{\prod_{i=1,...,N-1} (p_{i+1} - p_i)} \\
&= \sqrt{\delta_1 \delta_2 ... \delta_{N-1}} \\
&= \sqrt{\prod_{j \neq i, i+1} \delta_j \cdot \delta_i \delta_{i+1}} \\
&= \sqrt{\prod_{j \neq i, i+1} \delta_j \cdot \delta_i (l'_s - \delta_i)} \\
&= \sqrt{\prod_{j \neq i, i+1} \delta_j \cdot (-(\delta_i - \frac{l'_s}{2})^2 + \frac{l'^2_s}{4})}
\end{aligned}
$$

Then if we take new $\delta'_i = \delta'_{i+1} = \frac{l'_s}{2}$, the new distance $dist'_{alti}(p_1, ..., p_N)$ has a value

of $\sqrt{\prod_{j\neq i,i+1}\delta_j \cdot (\frac{l'^2_s}{4})}$. Since $\delta_i \neq \delta_{i+1} \neq \frac{l'_s}{2}$ then $dist'_{alti}(p_1, ..., p_N) > dist_{alti}(p_1, ..., p_N)$, which is contradictory to the assumption. Therefore Altitude Based Distance has the maximum value when $\delta_1 = ... = \delta_{N-1}$, i.e. $p_2 - p_1 = p_3 - p_2 = ... = p_N - p_{N-1}$.

We also prove Hypotenuse Based Distance has the minimum value when $\delta_1 = ... = \delta_{N-1}$ by contradiction. First we make an assumption that there exist δ_i and δ_{i+1} where $\delta_i \neq \delta_{i+1}$, such that $dist_{alti}(p_1, ..., p_N)$ has the minimum value.

$$\begin{aligned}
dist_{hypo}(p_1, ..., p_N) &= \sqrt{\sum_{i=1,...,N-1}(p'_{i+1} - p'_i)^2} \\
&= \sqrt{\delta_1^2 + \delta_2^2 ... + \delta_{N-1}^2} \\
&= \sqrt{(\sum_{j\neq i,i+1}\delta_j^2) + \delta_i^2 + \delta_{i+1}^2} \\
&= \sqrt{(\sum_{j\neq i,i+1}\delta_j^2) + \delta_i^2 + (l'_s - \delta_i)^2} \\
&= \sqrt{(\sum_{j\neq i,i+1}\delta_j^2) + (2(\delta_i - \frac{l'_s}{2})^2 + \frac{l'^2_s}{2})}
\end{aligned}$$

Then if we take new $\delta'_i = \delta'_{i+1} = \frac{l'_s}{2}$, the new distance $dist'_{hypo}(p_1, ..., p_N)$ has a value of $\sqrt{(\sum_{j\neq i,i+1}\delta_j^2) + \frac{l'^2_s}{2}}$. Since $\delta_i \neq \delta_{i+1} \neq \frac{l'_s}{2}$ then $dist'_{hypo}(p_1, ..., p_N) < dist_{alti}(p_1, ..., p_N)$, which is contradictory to the assumption. Therefore Hypotenuse Based Distance has the minimum value when $\delta_1 = ... = \delta_{N-1}$, i.e. $p_2 - p_1 = p_3 - p_2 =$

$... = p_N - p_{N-1}$.

Now we have proved the theorem for all cases. The proof for Altitude Based Distance can also be derived from the inequality of arithmetic and geometric means (AM-GM inequality).

□

B Topic Sets in Experiments

B.1 Appendix: Topics in TREC8

<top>
<num> Number: 401
<title> foreign minorities, Germany
<desc> Description: What language and cultural differences impede the integration of foreign minorities in Germany?
<narr> Narrative: A relevant document will focus on the causes of the lack of integration in a significant way; that is, the mere mention of immigration difficulties is not relevant. Documents that discuss immigration problems unrelated to Germany are also not relevant.
</top>
<top>
<num> Number: 402
<title> behavioral genetics
<desc> Description: What is happening in the field of behavioral genetics, the study of the relative influence of genetic and environmental factors on an individual's behavior or personality?
<narr> Narrative: Documents describing genetic or environmental factors relating to understanding and preventing substance abuse and addictions are relevant. Documents pertaining to attention deficit disorders tied in with genetics are also relevant, as are genetic disorders affecting hearing or muscles. The genome project is relevant when tied in with behavior disorders (i.e., mood disorders, Alzheimer's disease).
</top>
<top>
<num> Number: 403
<title> osteoporosis
<desc> Description: Find information on the effects of the dietary intakes of potassium, magnesium and fruits and vegetables as determinants of bone mineral density in elderly men and women thus preventing osteoporosis (bone decay).
<narr> Narrative: A relevant document may include one or more of the dietary intakes in the prevention of osteoporosis. Any discussion of the disturbance of nutrition and mineral metabolism that results in a decrease in bone mass is also relevant.
</top>
<top>
<num> Number: 404
<title> Ireland, peace talks
<desc> Description: How often were the peace talks in Ireland delayed or disrupted as a result of acts of violence?

<narr> Narrative: Any interruptions to the peace process not directly attributable to acts of violence are not relevant.
</top>
<top>
<num> Number: 405
<title> cosmic events
<desc> Description: What unexpected or unexplained cosmic events or celestial phenomena, such as radiation and supernova outbursts or new comets, have been detected?
<narr> Narrative: New theories or new interpretations concerning known celestial objects made as a result of new technology are not relevant.
</top>
<top>
<num> Number: 406
<title> Parkinson's disease
<desc> Description: What is being done to treat the symptoms of Parkinson's disease and keep the patient functional as long as possible?
<narr> Narrative: A relevant document identifies a drug or treatment program utilized in patient care and provides an indication of success or failure.
</top>
<top>
<num> Number: 407
<title> poaching, wildlife preserves
<desc> Description: What is the impact of poaching on the world's various wildlife preserves?
<narr> Narrative: A relevant document must discuss poaching in wildlife preserves, not in the wild itself. Also deemed relevant is evidence of preventive measures being taken by local authorities.
</top>
<top>
<num> Number: 408
<title> tropical storms
<desc> Description: What tropical storms (hurricanes and typhoons) have caused significant property damage and loss of life?
<narr> Narrative: The date of the storm, the area affected, and the extent of damage/casualties are all of interest. Documents that describe the damage caused by a tropical storm as "slight", "limited", or "small" are not relevant.
</top>
<top>
<num> Number: 409
<title> legal, Pan Am, 103
<desc> Description: What legal actions have resulted from the destruction of Pan Am Flight 103 over Lockerbie, Scotland, on December 21, 1988?
<narr> Narrative: Documents describing any charges, claims, or fines presented to or imposed by any court or tribunal are relevant, but documents that discuss charges made in diplomatic jousting are not relevant.
</top>
<top>
<num> Number: 410
<title> Schengen agreement
<desc> Description: Who is involved in the Schengen agreement to eliminate border controls in Western Europe and what do they hope to accomplish?
<narr> Narrative: Relevant documents will contain any information about the actions of signatories of the Schengen agreement such as: measures to eliminate border controls (removal of traffic obstacles, lifting of traffic restrictions); implementation of the information system data bank that contains unified visa issuance procedures; or strengthening

of border controls at the external borders of the treaty area in exchange for free movement at the internal borders. Discussions of border crossovers for business purposes are not relevant.
</top>
<top>
<num> Number: 411
<title> salvaging, shipwreck, treasure
<desc> Description: Find information on shipwreck salvaging: the recovery or attempted recovery of treasure from sunken ships.
<narr> Narrative: A relevant document will provide information on the actual locating and recovery of treasure; on the technology which makes possible the discovery, location and investigation of wreckages which contain or are suspected of containing treasure; or on the disposition of the recovered treasure.
</top>
<top>
<num> Number: 412
<title> airport security
<desc> Description: What security measures are in effect or are proposed to go into effect in airports?
<narr> Narrative: A relevant document could identify a specific airport and describe the security measures already in effect or proposed for use at that airport. Relevant items could also describe a failure of security that was cited as a contributing cause of a tragedy which came to pass or which was later averted. Comparisons between and among airports based on the effectiveness of the security of each are also relevant.
</top>
<top>
<num> Number: 413
<title> steel production
<desc> Description: What are new methods of producing steel?
<narr> Narrative: Relevant documents will discuss the processes adapted by entrepreneurs who have organized so-called "minimills" and are producing steel by methods which differ from the old blast furnace method of production. Documents that identify the new companies, the problems they have have encountered, and/or their successes or failures in the national and international markets are also relevant.
</top>
<top>
<num> Number: 414
<title> Cuba, sugar, exports
<desc> Description: How much sugar does Cuba export and which countries import it?
<narr> Narrative: A relevant document will provide information regarding Cuba's sugar trade. Sugar production statistics are not relevant unless exports are mentioned explicitly.
</top>
<top>
<num> Number: 415
<title> drugs, Golden Triangle
<desc> Description: What is known about drug trafficking in the "Golden Triangle", the area where Burma, Thailand and Laos meet?
<narr> Narrative: A relevant document will discuss drug trafficking in the Golden Triangle, including organizations that produce or distribute the drugs; international efforts to combat the traffic; or the quantities of drugs produced in the area.
</top>
<top>
<num> Number: 416
<title> Three Gorges Project
<desc> Description: What is the status of The Three Gorges Project?

<narr> Narrative: A relevant document will provide the projected date of completion of the project, its estimated total cost, or the estimated electrical output of the the finished project. Discussions of the social, political, or ecological impact of the project are not relevant.
</top>
<top>
<num> Number: 417
<title> creativity
<desc> Description: Find ways of measuring creativity.
<narr> Narrative: Relevant items include definitions of creativity, descriptions of characteristics associated with creativity, and factors linked to creativity.
</top>
<top>
<num> Number: 418
<title> quilts, income
<desc> Description: In what ways have quilts been used to generate income?
<narr> Narrative: Documents mentioning quilting books, quilting classes, quilted objects, and museum exhibits of quilts are all relevant. Documents that discuss AIDs quilts are irrelevant, unless there is specific mention that the quilts are being used for fundraising.
</top>
<top>
<num> Number: 419
<title> recycle, automobile tires
<desc> Description: What new uses have been developed for old automobile tires as a means of tire recycling?
<narr> Narrative: A relevant document must show advantageous uses of recycled tires, such as: destructive distillation of scrap rubber for valuable chemicals, reef building for fish habitats, filler or binder in asphalt roadway mixes, and burning in a controlled environment for heat generation.
</top>
<top>
<num> Number: 420
<title> carbon monoxide poisoning
<desc> Description: How widespread is carbon monoxide poisoning on a global scale?
<narr> Narrative: Relevant documents will contain data on what carbon monoxide poisoning is, symptoms, causes, and/or prevention. Advertisements for carbon monoxide protection products or services are not relevant. Discussions of auto emissions and air pollution are not relevant even though they can contain carbon monoxide.
</top>
<top>
<num> Number: 421
<title> industrial waste disposal
<desc> Description: How is the disposal of industrial waste being accomplished by industrial management throughout the world?
<narr> Narrative: Documents that discuss the disposal, storage, or management of industrial waste—both standard and hazardous—are relevant. However, documents that discuss disposal or storage of nuclear or radioactive waste, or the illegal shipment or dumping of waste to avoid legal disposal methods are not relevant.
</top>
<top>
<num> Number: 422
<title> art, stolen, forged
<desc> Description: What incidents have there been of stolen or forged art?

<narr> Narrative: Instances of stolen or forged art in any media are relevant. Stolen mass-produced things, even though they might be decorative, are not relevant (unless they are mass- produced art reproductions). Pirated software, music, movies, etc. are not relevant.
</top>
<top>
<num> Number: 423
<title> Milosevic, Mirjana Markovic
<desc> Description: Find references to Milosevic's wife, Mirjana Markovic.
<narr> Narrative: Any mention of the Serbian president's wife is relevant, even if she is not named. She may be referred to by her nickname, Mira. A general mention of his family, without specifying his wife, is not relevant.
</top>
<top>
<num> Number: 424
<title> suicides
<desc> Description: Give examples of alleged suicides that aroused suspicion of the death actually being murder.
<narr> Narrative: The intent of this query is to find criminal murders that are being disguised as suicide, but assisted suicides done out of compassion would be relevant if someone refers to them as murder.
</top>
<top>
<num> Number: 425
<title> counterfeiting money
<desc> Description: What counterfeiting of money is being done in modern times?
<narr> Narrative: Relevant documents must cite actual instances of counterfeiting. Anti-counterfeiting measures by themselves are not relevant.
</top>
<top>
<num> Number: 426
<title> law enforcement, dogs
<desc> Description: Provide information on the use of dogs worldwide for law enforcement purposes.
<narr> Narrative: Relevant items include specific information on the use of dogs during an operation. Training of dogs and their handlers are also relevant.
</top>
<top>
<num> Number: 427
<title> UV damage, eyes
<desc> Description: Find documents that discuss the damage ultraviolet (UV) light from the sun can do to eyes.
<narr> Narrative: A relevant document will discuss diseases that result from exposure of the eyes to UV light, treatments for the damage, and/or education programs that help prevent damage. Documents discussing treatment methods for cataracts and ocular melanoma are relevant even when a specific cause is not mentioned. However, documents that discuss radiation damage from nuclear sources or lasers are not relevant.
</top>
<top>
<num> Number: 428
<title> declining birth rates
<desc> Description: Do any countries other than the U.S. and China have a declining birth rate?
<narr> Narrative: To be relevant, a document will name a country other than the U.S. or China in which the birth rate fell from the rate of the previous year. The decline need not have occurred in more than the one preceding year.
</top>
<top>
<num> Number: 429

<title> Legionnaires' disease
<desc> Description: Identify outbreaks of Legionnaires' disease.
<narr> Narrative: To be relevant, a document must discuss a specific outbreak of Legionnaires' disease. Documents that address prevention of or cures for the disease without citing a specific case are not relevant.
</top>
<top>
<num> Number: 430
<title> killer bee attacks
<desc> Description: Identify instances of attacks on humans by Africanized (killer) bees.
<narr> Narrative: Relevant documents must cite a specific instance of a human attacked by killer bees. Documents that note migration patterns or report attacks on other animals are not relevant unless they also cite an attack on a human.
</top>
<top>
<num> Number: 431
<title> robotic technology
<desc> Description: What are the latest developments in robotic technology?
<narr> Narrative: A relevant document will contain information on current applications of robotic technology. Discussions of robotics research or simulations of robots are not relevant.
</top>
<top>
<num> Number: 432
<title> profiling, motorists, police
<desc> Description: Do police departments use "profiling" to stop motorists?
<narr> Narrative: A relevant document will report or discuss police department criteria for identifying motorists considered likely to be carrying contraband. Documents discussing the detention of individuals by foreign security forces are not relevant.
</top>
<top>
<num> Number: 433
<title> Greek, philosophy, stoicism
<desc> Description: Is there contemporary interest in the Greek philosophy of stoicism?
<narr> Narrative: Actual references to the philosophy or philosophers, productions of Greek stoic plays, and new "stoic" artistic productions are all relevant.
</top>
<top>
<num> Number: 434
<title> Estonia, economy
<desc> Description: What is the state of the economy of Estonia?
<narr> Narrative: Documents that give concrete economic information such as economic statistics, entering economic unions and treaties, or monetary performance are relevant, as are discussions of economic issues such as transportation or pollution.
</top>
<top>
<num> Number: 435
<title> curbing population growth
<desc> Description: What measures have been taken worldwide and what countries have been effective in curbing population growth?

<narr> Narrative: A relevant document must describe an actual case in which population measures have been taken and their results are known. The reduction measures must have been actively pursued; that is, passive events such as disease or famine involuntarily reducing the population are not relevant.
</top>
<top>
<num> Number: 436
<title> railway accidents
<desc> Description: What are the causes of railway accidents throughout the world?
<narr> Narrative: A relevant document provides data on railway accidents of any sort (i.e., locomotive, trolley, streetcar) where either the railroad system or the vehicle or pedestrian involved caused the accident. Documents that discuss railroading in general, new rail lines, new technology for safety, and safety and accident prevention are not relevant, unless an actual accident is described.
</top>
<top>
<num> Number: 437
<title> deregulation, gas, electric
<desc> Description: What has been the experience of residential utility customers following deregulation of gas and electric?
<narr> Narrative: Documents that discuss privatization of government- owned utilities alone are not relevant. Also, not relevant are documents that discuss the deregulation of utilities for commercial customers.
</top>
<top>
<num> Number: 438
<title> tourism, increase
<desc> Description: What countries are experiencing an increase in tourism?
<narr> Narrative: A relevant document will name a country that has experienced an increase in tourism. The increase must represent the nation as a whole and tourism in general, not be restricted to only certain regions of the country or to some specific type of tourism (e.g., adventure travel). Documents discussing only projected increases are not relevant.
</top>
<top>
<num> Number: 439
<title> inventions, scientific discoveries
<desc> Description: What new inventions or scientific discoveries have been made?
<narr> Narrative: The word "new" in the description is defined as occurring in the 1990s. Documents that indicate a "recent" invention or scientific discovery are considered relevant. Discoveries made in astronomy or any scientific discoveries that are not patentable are not relevant.
</top>
<top>
<num> Number: 440
<title> child labor
<desc> Description: What steps are being taken by governments or corporations to eliminate abuse of child labor?
<narr> Narrative: A relevant document identifies an action taken by either a private commercial corporation or governmental organization to reduce or eliminate the use of child labor in manufacturing operations.
</top>
<top>
<num> Number: 441
<title> Lyme disease
<desc> Description: How do you prevent and treat Lyme disease?

<narr> Narrative: Documents that discuss current prevention and treatment techniques for Lyme disease are relevant. Reports of research on new treatments of the disease are also relevant.
</top>
<top>
<num> Number: 442
<title> heroic acts
<desc> Description: Find accounts of selfless heroic acts by individuals or small groups for the benefit of others or a cause.
<narr> Narrative: Relevant documents will contain a description of specific acts. General statements concerning heroic acts are not relevant.
</top>
<top>
<num> Number: 443
<title> U.S., investment, Africa
<desc> Description: What is the extent of U.S. (government and private) investment in sub-Saharan Africa?
<narr> Narrative: All references to U.S. Governmental and private assistance to sub-Saharan Africa are relevant. Documents discussing contributions by reason of U.S. membership in international aid organizations are also relevant.
</top>
<top>
<num> Number: 444
<title> supercritical fluids
<desc> Description: What are the potential uses for supercritical fluids as an environmental protection measure?
<narr> Narrative: To be relevant, a document must indicate that the fluid involved is achieved by a process of pressurization producing the supercritical fluid.
</top>
<top>
<num> Number: 445
<title> women clergy
<desc> Description: What other countries besides the United States are considering or have approved women as clergy persons?
<narr> Narrative: To be relevant, a document must indicate either a country where a woman has been installed as clergy or a country that is considering such an installation. The clergy position must be as church pastor rather than some other church capacity (e.g., nun or choir member).
</top>
<top>
<num> Number: 446
<title> tourists, violence
<desc> Description: Where are tourists likely to be subjected to acts of violence causing bodily harm or death?
<narr> Narrative: A relevant document must contain accounts of known harm to tourists. Evidence of single, isolated incidents are not relevant.
</top>
<top>
<num> Number: 447
<title> Stirling engine
<desc> Description: What new developments and applications are there for the Stirling engine?
<narr> Narrative: Any discussion of new developments and applications of the Stirling engine (also known as the Stirling cycle) are relevant.
</top>
<top>
<num> Number: 448

<title> ship losses
<desc> Description: Identify instances in which weather was a main or contributing factor in the loss of a ship at sea.
<narr> Narrative: Any ship loss due to weather is relevant, either in international or coastal waters.
</top>
<top>
<num> Number: 449
<title> antibiotics ineffectiveness
<desc> Description: What has caused the current ineffectiveness of antibiotics against infections and what is the prognosis for new drugs?
<narr> Narrative: To be relevant, a document must discuss the reasons or causes for the ineffectiveness of current antibiotics. Relevant documents may also include efforts by pharmaceutical companies and federal government agencies to find new cures, updating current testing phases, new drugs being tested, and the prognosis for the availability of new and effective antibiotics.
</top>
<top>
<num> Number: 450
<title> King Hussein, peace
<desc> Description: How significant a figure over the years was the late Jordanian King Hussein in furthering peace in the Middle East?
<narr> Narrative: A relevant document must include mention of Israel; King Hussein himself as opposed to other Jordanian officials; discussion of the King's on-going, previous or upcoming efforts; and efforts pertinent to the peace process, not merely Jordan's relationship with other middle-east countries or the U.S.
</top>

B.2 Appendix: Topics in Robust

<top>
<num> 301 <title> International Organized Crime
<desc> Identify organizations that participate in international criminal activity, the activity, and, if possible, collaborating organizations and the countries involved.
<narr> A relevant document must as a minimum identify the organization and the type of illegal activity (e.g., Columbian cartel exporting cocaine). Vague references to international drug trade without identification of the organization(s) involved would not be relevant.
</top>
<top>
<num> 302 <title> Poliomyelitis and Post-Polio
<desc> Is the disease of Poliomyelitis (polio) under control in the world?
<narr> Relevant documents should contain data or outbreaks of the polio disease (large or small scale), medical protection against the disease, reports on what has been labeled as "post-polio" problems. Of interest would be location of the cases, how severe, as well as what is being done in the "post-polio" area.
</top>
<top>
<num> 303 <title> Hubble Telescope Achievements
<desc> Identify positive accomplishments of the Hubble telescope since it was launched in 1991.
<narr> Documents are relevant that show the Hubble telescope has produced new data, better quality data than previously available, data that has increased human knowledge of the universe, or data that has led to disproving previously existing theories or hypotheses. Documents limited to the shortcomings of the telescope would be irrelevant. Details of repairs or modifications to the telescope without reference to positive achievements would not be relevant.

</top>
<top>
<num> 304 <title> Endangered Species (Mammals)
<desc> Compile a list of mammals that are considered to be endangered, identify their habitat and, if possible, specify what threatens them.
<narr> Any document identifying a mammal as endangered is relevant. Statements of authorities disputing the endangered status would also be relevant. A document containing information on habitat and populations of a mammal identified elsewhere as endangered would also be relevant even if the document at hand did not identify the species as endangered. Generalized statements about endangered species without reference to specific mammals would not be relevant.
</top>
<top>
<num> 305 <title> Most Dangerous Vehicles
<desc> Which are the most crashworthy, and least crashworthy, passenger vehicles?
<narr> A relevant document will contain information on the crashworthiness of a given vehicle or vehicles that can be used to draw a comparison with other vehicles. The document will have to describe/compare vehicles, not drivers. For instance, it should be expected that vehicles preferred by 16-25 year-olds would be involved in more crashes, because that age group is involved in more crashes. I would view number of fatalities per 100 crashes to be more revealing of a vehicle's crashworthiness than the number of crashes per 100,000 miles, for example.
</top>
<top>
<num> 306 <title> African Civilian Deaths
<desc> How many civilian non-combatants have been killed in the various civil wars in Africa?
<narr> A relevant document will contain specific casualty information for a given area, country, or region. It will cite numbers of civilian deaths caused directly or indirectly by armed conflict.
</top>
<top>
<num> 307 <title> New Hydroelectric Projects
<desc> Identify hydroelectric projects proposed or under construction by country and location. Detailed description of nature, extent, purpose, problems, and consequences is desirable.
<narr> Relevant documents would contain as a minimum a clear statement that a hydroelectric project is planned or construction is under way and the location of the project. Renovation of existing facilities would be judged not relevant unless plans call for a significant increase in acre-feet or reservoir or a marked change in the environmental impact of the project. Arguments for and against proposed projects are relevant as long as they are supported by specifics, including as a minimum the name or location of the project. A statement that an individual or organization is for or against such projects in general would not be relevant. Proposals or projects underway to dismantle existing facilities or drain existing reservoirs are not relevant, nor are articles reporting a decision to drop a proposed plan.
</top>
<top>
<num> 308 <title> Implant Dentistry
<desc> What are the advantages and/or disadvantages of tooth implants?
<narr> A tooth replacement procedure, begun in the 1960s by Doctor Branemark, is becoming more widely used today. It involves the replacement of a lost tooth/teeth by an implantation process which secures the fabricated tooth to a titanium post with an adhesive resulting in a stable and sturdy denture almost like the original. A relevant document will include any clinical experiment, report, study, paper, or medical discussion which describes the advantages or disadvantages of tooth implant(s), conditions under which such a procedure is favorable, denture comfort and function compared to false teeth, bridge, or plate and comparative cost differential.
</top>
<top>
<num> 309 <title> Rap and Crime

<desc> Evidence that rap music has a negative effect on young people.
<narr> The lyrics to Rap music are replete with words which degrade women, glorify drug abuse, the killing of policemen and even romanticize suicide. A crusade was begun a few years ago in which the music industry was asked "to clean up its act" because of the detrimental effect it was having upon teenagers. There were some instances in which teenagers, who demonstrated no sign of internal strife, committed suicide. The only clues to their self-destruction were notes found near the bodies with lyrics from Rap songs relating to suicide. A relevant item should include any psychological study, evidence, poll, documentation, or psychological opinion that Rap music has a deleterious effect upon young people and has led to a discernible change in personality, life style, crime, alteration of personality or any other negative effect upon the psyche, mores, and culture of teenagers.
</top>
<top>
<num> 310 <title> Radio Waves and Brain Cancer
<desc> Evidence that radio waves from radio towers or car phones affect brain cancer occurrence.
<narr> Persons living near radio towers and more recently persons using car phones have been diagnosed with brain cancer. The argument rages regarding the direct association of one with the other. The incidence of cancer among the groups cited is considered, by some, to be higher than that found in the normal population. A relevant document includes any experiment with animals, statistical study, articles, news items which report on the incidence of brain cancer being higher/lower/same as those persons who live near a radio tower and those using car phones as compared to those in the general population.
</top>
<top>
<num> 311 <title> Industrial Espionage
<desc> Document will discuss the theft of trade secrets along with the sources of information: trade journals, business meetings, data from Patent Offices, trade shows, or analysis of a competitor's products.
<narr> A relevant document will contain specific information on the theft of trade secrets on new or existing technology. Relevant information includes, but is not limited to, espionage in the highly competitive fashion industry, toy companies, drug firms, and computer companies. An apparently legal way of stealing company secrets is by using the Freedom of Information Act, passed by the U.S. Congress in 1966 to help people and the press get information for the public good. The law has often been used instead by companies for private gain. Documents discussing company counterespionage are also considered to be relevant.
</top>
<top>
<num> 312 <title> Hydroponics
<desc> Document will discuss the science of growing plants in water or some substance other than soil.
<narr> A relevant document will contain specific information on the necessary nutrients, experiments, types of substrates, and/or any other pertinent facts related to the science of hydroponics. Related information includes, but is not limited to, the history of hydroponics, advantages over standard soil agricultural practices, or the approach of suspending roots in a humid enclosure and spraying them periodically with a nutrient solution to promote plant growth.
</top>
<top>
<num> 313 <title> Magnetic Levitation-Maglev
<desc> Commercial uses of Magnetic Levitation.
<narr> A relevant document must contain Magnetic Levitation or Maglev. It should be concerned with possible commercial applications of this phenomenon to include primarily mass transit, but also other commercial applications such as Maglev flywheels for cars. Discussions of superconductivity when linked to Maglev and government support plans when linked to Maglev are also relevant.
</top>
<top>
<num> 314 <title> Marine Vegetation

<desc> Commercial harvesting of marine vegetation such as algae, seaweed and kelp for food and drug purposes.
<narr> Recent research has shown that marine vegetation is a valuable source of both food (human and animal) and a potentially useful drug. This search will focus primarily on these two uses. Also to be considered relevant would be instances of other possible commercial uses such as fertilizer, etc.
</top>
<top>
<num> 315 <title> Unexplained Highway Accidents
<desc> How many fatal highway accidents are there each year that are not resolved as to cause.
<narr> A relevant document will contain data relating to highway accidents where the cause of the accident cannot be determined. Typical of such accidents would be those where one vehicle "suddenly swerves into oncoming traffic."
</top>
<top>
<num> 316 <title> Polygamy Polyandry Polygyny
<desc> A look at the roots and prevalence of polygamy in the world today.
<narr> Polygamy is a form of marriage which permits a person to have more than one husband or wife. Polyandry refers to one woman sharing two or more husbands at the same time. Polygyny refers to one man sharing two or more wives at the same time. Primary focus of the search will be the prevalence of these practices in the world today and societal attitudes towards these practices. Also relevant would be discussions of the roots and practical sources of these customs. A modern development in this area is serial polygamy, a phrase coined to label the practice of men who take a series of wives in sequence as a solution to practical welfare, considerations of child care, housing, etc. Documents discussing serial polygamy will not be considered relevant.
</top>
<top>
<num> 317 <title> Unsolicited Faxes
<desc> Have regulations been passed by the FCC banning junk facsimile (fax)? If so, are they effective?
<narr> Relevant documents will provide information on the cost of junk fax transmitted to individuals or businesses. Documents which contain laws or regulations passed banning junk fax or articles indicating the effectiveness or ineffectiveness of these laws are relevant. Additionally, any document showing where unsolicited/junk fax is an invasion of privacy is also relevant.
</top>
<top>
<num> 318 <title> Best Retirement Country
<desc> Aside from the United States, which country offers the best living conditions and quality of life for a U.S. retiree?
<narr> A relevant document will contain information describing the living conditions and/or costs in one or more foreign countries. It will provide information that a potential retiree could use in deciding where to establish a retirement home.
</top>
<top>
<num> 319 <title> New Fuel Sources
<desc> What research is ongoing for new fuel sources.
<narr> Relevant documents will contain information on the actual research being accomplished for new fuel sources. Documents may also reveal company (companies) involved in this research. Listing(s) of possible new fuel sources is also of relevance. Relevant documents can also show how much current fuel sources are available, however, articles which simply refer to current sources are not relevant. Documents about nuclear power are relevant only if they discuss new developments in the nuclear field, such as the processing of spent fuel to establish a new nuclear fuel system.
</top>
<top>
<num> 320 <title> Undersea Fiber Optic Cable

<desc> Fiber optic link around the globe (Flag) will be the world's longest undersea fiber optic cable. Who's involved and how extensive is the technology on this system. What problems exist?
<narr> Relevant documents will reference companies involved in building the system or the technology needed for such an endeavor. Of relevance also would be information on the link up points of FLAG or landing sites or interconnection with other telecommunication cables. Relevant documents may reference any regulatory problems with the system once constructed. A non-relevant document would contain information on other fiber optic systems currently in place.
</top>
<top>
<num> 321 <title> Women in Parliaments
<desc> Pertinent documents will reflect the fact that women continue to be poorly represented in parliaments across the world, and the gap in political power between the sexes is very wide, particularly in the Third World.
<narr> Pertinent documents relating to this issue will discuss the lack of representation by women, the countries that mandate the inclusion of a certain percentage of women in their legislatures, decreases if any in female representation in legislatures, and those countries in which there is no representation of women.
</top>
<top>
<num> 322 <title> International Art Crime
<desc> Isolate instances of fraud or embezzlement in the international art trade.
<narr> A relevant document is any report that identifies an instance of fraud or embezzlement in the international buying or selling of art objects. Objects include paintings, jewelry, sculptures and any other valuable works of art. Specific instances must be identified for a document to be relevant; generalities are not relevant.
</top>
<top>
<num> 323 <title> Literary/Journalistic Plagiarism
<desc> Find instances of plagiarism in the literary and journalistic worlds.
<narr> A relevant document will report any occasion or suspected instance of plagiarism in the areas of either literature or journalism. Relevant documents will also include such areas as doctorate and master's theses and will encompass writings as well as the ideas and concepts developed by some authors and taken or borrowed by others without attribution.
</top>
<top>
<num> 324 <title> Argentine/British Relations
<desc> Define Argentine and British international relations
<narr> It has been 15 years since the war between Argentina and the United Kingdom in 1982 over sovereignty in the Falkland Islands. A relevant report will describe their relations after that period. Any kind of international contact between the two countries is relevant, to include commercial, economic, cultural, diplomatic, or military exchanges. Negative reports on the absence of such exchanges are also desirable. Reports containing information on direct exchanges between Argentina and the Falkland Islands are also relevant.
</top>
<top>
<num> 325 <title> Cult Lifestyles
<desc> Describe a cult by name and identify the cult members' activities in their everyday life.
<narr> A relevant document would include the name of the cult and offer information about the members' lifestyles. It may include how they dress or what they do to attain the ultimate goal of the organization. A relevant document may tell what they eat or how they contribute to the cult. Just the mention of the existence of a cult by name with no other clarifying information would not be relevant.
</top>
<top>
<num> 326 <title> Ferry Sinkings

<desc> Any report of a ferry sinking where 100 or more people lost their lives.
<narr> To be relevant, a document must identify a ferry that has sunk causing the death of 100 or more humans. It must identify the ferry by name or place where the sinking occurred. Details of the cause of the sinking would be helpful but are not necessary to be relevant. A reference to a ferry sinking without the number of deaths would not be relevant.
</top>
<top>
<num> 327 <title> Modern Slavery
<desc> Identify a country or a city where there is evidence of human slavery being practiced in the eighties or nineties.
<narr> A relevant document would present evidence of current slavery practices being carried out. It would identify a specific country or city and give some information on who the slaves are, or who was buying or selling them, and for what purposes they were being used. References to slavery being carried out several years ago would not be relevant.
</top>
<top>
<num> 328 <title> Pope Beatifications
<desc> Identify an individual that has been beatified by the Pope.
<narr> To be relevant, the document must include the person's name as well as the person's deeds or actions that led to the recognition by the Pope. It must also include when such actions or deeds occurred by giving the year or period of time, i.e., "late 17th century".
</top>
<top>
<num> 329 <title> Mexican Air Pollution
<desc> Mexico City has the worst air pollution in the world. Pertinent Documents would contain the specific steps Mexican authorities have taken to combat this deplorable situation.
<narr> Relevant documents would discuss the steps the Mexican Government has taken to alleviate the air pollution in Mexico City. Steps such as reducing the number of automobiles in the city, encouraging the use of mass public transportation, and creating new mass transportation systems are relevant, among others. Mention of any new methods in the design stage would also be appropriate.
</top>
<top>
<num> 330 <title> Iran-Iraq Cooperation
<desc> This query is looking for examples of cooperation or friendly ties between Iran and Iraq, or ways in which the two countries could be considered allies.
<narr> A relevant document would mention such things as mutually beneficial economic, military, religious, or social relations; cooperation on border control or treatment of minorities; collaboration in getting around trade sanctions placed upon either country, etc. (Any mention of the possible return of the airplanes Iraq flew to Iran during the Gulf War would also be relevant).
</top>
<top>
<num> 331 <title> World Bank Criticism
<desc> What criticisms have been made of World Bank policies, activities or personnel?
<narr> This query is looking for any instances where the World Bank has been accused of things like not being responsive to the unique problems of individual countries, of being too strict in its policies, of pursuing agendas that are biased because of their benefits to western countries, of being no longer useful or practical, of its personnel being difficult to work with, etc.
</top>
<top>
<num> 332 <title> Income Tax Evasion

<desc> This query is looking for investigations that have targeted evaders of U.S. income tax.
<narr> A relevant document would mention investigations either in the U.S. or abroad of people suspected of evading U.S. income tax laws. Of particular interest are investigations involving revenue from illegal activities, as a strategy to bring known or suspected criminals to justice.
</top>
<top>
<num> 333 <title> Antibiotics Bacteria Disease
<desc> Determine the reasons why bacteria seems to be winning the war against antibiotics and rendering antibiotics now less effective in treating diseases than they were in the past.
<narr> A relevant document will address the questions of how and why, and to what degree, bacteria are able to fend off the curative effects of antibiotics. Overuse of antibiotics, as well as the increasing use of antibiotics in promoting the growth of crops and animals whose food products are meant for human consumption have played roles in creating a situation where every known bacteria-generated disease now has versions that resist at least one of the more than 100 antibiotics now in use.
</top>
<top>
<num> 334 <title> Export Controls Cryptography
<desc> Determine the usefulness and effectiveness of continuing to maintain export controls on encryption software.
<narr> Relevant documents will argue for or against continuing to make encryption software subject to export controls. Since 1993 quality encryption software to ensure the secrecy of communications has been available, but the U.S. Government has considered such software to be subject to the same export controls as munitions, and has sought to restrict the export of encryption software unless it contains a device which could allow the U.S. to read the underlying messages. Business interests say that this will make it impossible for U.S. producers to compete in the international market.
</top>
<top>
<num> 335 <title> Adoptive Biological Parents
<desc> Identify the problems, and solutions to those problems, which arise in the relationships among biological parents, adoptive parents, and the child or children involved.
<narr> A relevant document will report on court procedures and decisions which affect the rights of biological parents, adoptive parents, as well as the adopted child. Problems arise when the biological parents of a child given up for adoption decide that they or he/she would like to reclaim the child against the wishes of the adoptive parents. Problems also arise when the adopted child tries to identify and contact the biological parents, but find the way blocked by sealed court orders.
</top>
<top>
<num> 336 <title> Black Bear Attacks
<desc> A relevant document would discuss the frequency of vicious black bear attacks worldwide and the possible causes for this savage behavior.
<narr> It has been reported that food or cosmetics sometimes attract hungry black bears, causing them to viciously attack humans. Relevant documents would include the aforementioned causes as well as speculation preferably from the scientific community as to other possible causes of vicious attacks by black bears. A relevant document would also detail steps taken or new methods devised by wildlife officials to control and/or modify the savageness of the black bear.
</top>
<top>
<num> 337 <title> Viral Hepatitis
<desc> What research has been done on viral hepatitis and what progress has been made in its treatment?
<narr> A relevant document might include any of the following information:
– A medical procedure used in the treatment of hepatitis.

– Ongoing research in vaccines, including the identity of the strain of hepatitis.
– How hepatitis affects the body's immune system.
– Third World countries' reports on control of the disease.
</top>
<top>
<num> 338 <title> Risk of Aspirin
<desc> What adverse effects have people experienced while taking aspirin repeatedly?
<narr> A relevant document should identify any adverse effects experienced from the repeated use of aspirin. Possible effects might include intestinal bleeding, inflammation of the stomach, or various forms of ulcers. The purpose of the individual's repeated aspirin use should also be stated.
</top>
<top>
<num> 339 <title> Alzheimer's Drug Treatment
<desc> What drugs are being used in the treatment of Alzheimer's Disease and how successful are they?
<narr> A relevant document should name a drug used in the treatment of Alzheimer's Disease and also its manufacturer, and should give some indication of the drug's success or failure.
</top>
<top>
<num> 340 <title> Land Mine Ban
<desc> Identify any actions being taken to propel the nations of the world toward a treaty banning the production, transfer and use of land mines.
<narr> Brought about by various conflicts, there are an estimated 100 million land mines buried in 60 countries with thousands of acres of land made unusable. It is estimated that every 22 minutes a man, woman or child is killed or maimed by a land mine. Identify any actions being taken to propel the nations of the world toward a treaty banning the production, transfer and use of land mines. Also, of interest, would be identification of actions being undertaken to sanitize the thousands of acres now made unusable due to use of these mines.
</top>
<top>
<num> 341 <title> Airport Security
<desc> A relevant document would discuss how effective government orders to better scrutinize passengers and luggage on international flights and to step up screening of all carry-on baggage has been.
<narr> A relevant document would contain reports on what new steps airports worldwide have taken to better scrutinize passengers and their luggage on international flights and to step up screening of all carry-on baggage. With the increase in international terrorism and in the wake of the TWA Flight 800 disaster, articles on airport security relating in particular to additional steps taken by airports to increase flight safety would be relevant. The mere mention of enhanced security does not constitute relevance. Additional steps refer to something beyond just passenger and carry-on screening using the normal methods. Examples of new steps would be additional personnel, sophisticated monitoring and screening devices, and extraordinary measures to check luggage in the baggage compartment.
</top>
<top>
<num> 342 <title> Diplomatic Expulsion
<desc> The end of the Cold War seems to have intensified economic competition and has started to generate serious friction between nations as attempts are made by diplomatic personnel to acquire sensitive trade and technology information or to obtain information on highly classified industrial projects. Identify instances where attempts have been made by personnel with diplomatic status to obtain information of this nature.
<narr> Identify instances where attempts have been made by personnel with diplomatic status to obtain information of this nature. Of interest would be the country(s) involved, the information hopefully acquired, or if the exposure resulted in expulsion of diplomatic personnel.
</top>

<top>
<num> 343 <title> Police Deaths
<desc> Identify instances where a civilian policeman has been killed either during performance of his duty or because of other association with this occupation, e.g., killed for the gun, to keep from testifying, etc.
<narr> Whatever the reason, be it a breakdown in family values or loss of respect for an authoritative figure, there seems to have been a substantial increase in the number of policemen being killed. Identify instances where a civilian policeman has been killed either during performance of his duty or because of other association with this occupation, e.g., killed for the gun, to keep from testifying, etc. Military police or foreign military-type personnel considered policemen should not be considered relevant.
</top>
<top>
<num> 344 <title> Abuses of E-Mail
<desc> The availability of E-mail to many people through their job or school affiliation has allowed for many efficiencies in communications but also has provided the opportunity for abuses. What steps have been taken world-wide by those bearing the cost of E-mail to prevent excesses?
<narr> To be relevant, a document will concern dissatisfaction by an entity paying for the cost of electronic mail. Particularly sought are items which relate to system users (such as employees) who abuse the system by engaging in communications of the type not related to the payer's desired use of the system.
</top>
<top>
<num> 345 <title> Overseas Tobacco Sales
<desc> Health studies primarily in the U.S. have caused reductions in tobacco sales here, but the economic impact has caused U.S. tobacco companies to look overseas for customers. What impact have the health and economic factors had overseas?
<narr> To be relevant, an item will discuss either an increase or decrease in the sales of U.S. tobacco products overseas, and attribute this to health findings or more aggressive marketing by U.S. companies or the foreign companies they own or cooperate with.
</top>
<top>
<num> 346 <title> Educational Standards
<desc> There has long been a call for standards in U.S. education, these calls frequently citing the superiority of foreign school systems. Are there many countries outside the U.S. which have standards for pre-teen students? If so, which are those countries and what standards have been set?
<narr> To be relevant, an item will specify the foreign country which has adopted standards for children up to the age of 12 and more specifically give some detail of what the standards are (e.g., computer training, 2 years of foreign language).
</top>
<top>
<num> 347 <title> Wildlife Extinction
<desc> The spotted owl episode in America highlighted U.S. efforts to prevent the extinction of wildlife species. What is not well known is the effort of other countries to prevent the demise of species native to their countries. What other countries have begun efforts to prevent such declines?
<narr> A relevant item will specify the country, the involved species, and steps taken to save the species.
</top>
<top>
<num> 348 <title> Agoraphobia
<desc> Is the fear of open or public places (Agoraphobia) a widespread disorder or relatively unknown?
<narr> Relevant documents contain data on this physical/mental disorder, including information on the person affected, profession of the individual, impact on the life work of the individual, as well as any data on how these individuals cope with this disorder.

</top>
<top>
<num> 349 <title> Metabolism
<desc> Document will discuss the chemical reactions necessary to keep living cells healthy and/or producing energy.
<narr> A relevant document will contain specific information on the catabolic and anabolic reactions of the metabolic process. Relevant information includes, but is not limited to, the reactions occurring in metabolism, biochemical processes (Glycolysis or Krebs cycle for production of energy), and disorders associated with the metabolic rate.
</top>
<top>
<num> 350 <title> Health and Computer Terminals
<desc> Is it hazardous to the health of individuals to work with computer terminals on a daily basis?
<narr> Relevant documents would contain any information that expands on any physical disorder/problems that may be associated with the daily working with computer terminals. Such things as carpel tunnel, cataracts, and fatigue have been said to be associated, but how widespread are these or other problems and what is being done to alleviate any health problems.
</top>
<top>
<num> 351 <title> Falkland petroleum exploration
<desc> What information is available on petroleum exploration in the South Atlantic near the Falkland Islands?
<narr> Any document discussing petroleum exploration in the South Atlantic near the Falkland Islands is considered relevant. Documents discussing petroleum exploration in continental South America are not relevant.
</top>
<top>
<num> 352 <title> British Chunnel impact
<desc> What impact has the Chunnel had on the British economy and/or the life style of the British?
<narr> Documents discussing the following issues are relevant:

- projected and actual impact on the life styles of the British - Long term changes to economic policy and relations - major changes to other transportation systems linked with the Continent

Documents discussing the following issues are not relevant:

- expense and construction schedule - routine marketing ploys by other channel crossers (i.e., schedule changes, price drops, etc.)
</top>
<top>
<num> 353 <title> Antarctica exploration
<desc> Identify systematic explorations and scientific investigations of Antarctica, current or planned.
<narr> Documents discussing the following issues are relevant:

- systematic explorations and scientific investigations of Antarctica (e.g., seismology, ionospheric physics, possible economic development) - other research currently conducted or planned for the future - banning of mineral mining

Documents discussing tourism are non-relevant. Documents discussing "disrupting scientific experiments" are non-relevant unless a specific experiment is identified.
</top>
<top>
<num> 354 <title> journalist risks
<desc> Identify instances where a journalist has been put at risk (e.g., killed, arrested or taken hostage) in the performance of his work.
<narr> Any document identifying an instance where a journalist or correspondent has been killed, arrested or taken hostage in the performance of his work is relevant.
</top>
<top>
<num> 355 <title> ocean remote sensing

<desc> Identify documents discussing the development and application of spaceborne ocean remote sensing.
<narr> Documents discussing the development and application of spaceborne ocean remote sensing in oceanography, seabed prospecting and mining, or any marine-science activity are relevant. Documents that discuss the application of satellite remote sensing in geography, agriculture, forestry, mining and mineral prospecting or any land-bound science are not relevant, nor are references to international marketing or promotional advertizing of any remote-sensing technology. Synthetic aperture radar (SAR) employed in ocean remote sensing is relevant.
</top>
<top>
<num> 356 <title> postmenopausal estrogen Britain
<desc> Identify documents discussing the use of estrogen by postmenopausal women in Britain.
<narr> The use of hormone replacement therapy outside of the United Kingdom is not relevant. United Kingdom and British development and marketing of estrogen suppressing drugs are relevant.
</top>
<top>
<num> 357 <title> territorial waters dispute
<desc> Identify documents discussing international boundary disputes relevant to the 200-mile special economic zones or 12-mile territorial waters subsequent to the passing of the "International Convention on the Law of the Sea".
<narr> To be relevant, documents must discuss an international boundary dispute relevant to the 200-mile special economic zones or 12-mile territorial waters subsequent to the passing of the "International Convention on the Law of the Sea". Documents that merely express agreement or disagreement with any of the precepts of the "Law of the Sea" are not relevant, nor are any documents that passively allude to the existence of the convention or any of its non-controversial applications.
</top>
<top>
<num> 358 <title> blood-alcohol fatalities
<desc> What role does blood-alcohol level play in automobile accident fatalities?
<narr> Relevant documents must contain information on automobile accidents in which there was a fatality and the blood-alcohol level of the driver of the vehicle must be identified.
</top>
<top>
<num> 359 <title> mutual fund predictors
<desc> Are there reliable and consistent predictors of mutual fund performance?
<narr> A document must contain at least one factor such as: rankings, risks, yields, or costs, and fund performance to be relevant. Documents that discuss mutual fund rankings are considered relevant.
</top>
<top>
<num> 360 <title> drug legalization benefits
<desc> What are the benefits, if any, of drug legalization?
<narr> Relevant documents may contain information on perceived benefits of drug legalization, such as crime reduction, improved treatment using monies which otherwise would have gone for crime fighting, reduced drug addiction, and increased governmental income. Documents that discuss drug legalization and whether legalization is or is not perceived to be beneficial are relevant.
</top>
<top>
<num> 361 <title> clothing sweatshops
<desc> Identify documents that discuss clothing sweatshops.
<narr> A relevant document must identify the country, the working conditions, salary, and type of clothing or shoes being produced. Relevant documents may also include the name of the business or company or the type of manufacturing, such as: "designer label".

</top>
<top>
<num> 362 <title> human smuggling
<desc> Identify incidents of human smuggling.
<narr> A relevant document shows an incident of humans (at least ten) being smuggled. The smugglers would have to realize a monetary gain for their actions, while the people being smuggled may or may not be willing participants.
</top>
<top>
<num> 363 <title> transportation tunnel disasters
<desc> What disasters have occurred in tunnels used for transportation?
<narr> A relevant document identifies a disaster in a tunnel used for trains, motor vehicles, or people. Wind tunnels and tunnels used for wiring, sewage, water, oil, etc. are not relevant. The cause of the problem may be fire, earthquake, flood, or explosion and can be accidental or planned. Documents that discuss tunnel disasters occurring during construction of a tunnel are relevant if lives were threatened.
</top>
<top>
<num> 364 <title> rabies
<desc> Identify documents discussing cases where rabies have been confirmed and what, if anything, is being done about it.
<narr> A relevant document identifies confirmed cases of rabies and may contain actions taken to correct the problem.
</top>
<top>
<num> 365 <title> El Nino
<desc> What effects have been attributed to El Nino?
<narr> A document is relevant if it describes a particular phenomenon (either specific event or generalization) like flood, drought, warming, etc. and names El Nino as the cause or as being a contributing factor.
</top>
<top>
<num> 366 <title> commercial cyanide uses
<desc> What are the industrial or commercial uses of cyanide or its derivatives?
<narr> A document is relevant if it names or describes a process that uses cyanide commercially or mentions that cyanide-rich waste comes from a particular industry.
</top>
<top>
<num> 367 <title> piracy
<desc> What modern instances have there been of old fashioned piracy, the boarding or taking control of boats?
<narr> Documents discussing piracy on any body of water are relevant. Documents discussing the legal taking of ships or their contents by a national authority are non-relevant. Clashes between fishing vessels over fishing are not relevant, unless one vessel is boarded.
</top>
<top>
<num> 368 <title> in vitro fertilization
<desc> Identify documents that discuss in vitro fertilization.
<narr> A relevant document will describe any aspect of the process of in vitro fertilization: uniting two human germ cells (sperm and egg) outside of the human body and in an artificial environment. Documents that describe related techniques, such as the freezing of eggs, sperm, or embryos for future implantation are also relevant.
</top>
<top>
<num> 369 <title> anorexia nervosa bulimia

<desc> What are the causes and treatments of anorexia nervosa and bulimia?
<narr> A relevant document will describe the causes of the eating disorders, anorexia nervosa or bulimia. Documents that describe symptoms including the taking of laxatives, self-induced vomiting and excessive exercise are relevant. Discussions of the treatment of the disorders are also relevant.
</top>
<top>
<num> 370 <title> food/drug laws
<desc> What are the laws dealing with the quality and processing of food, beverages, or drugs?
<narr> A relevant document will contain specific information on the laws dealing with such matters as quality control in processing, the use of additives and preservatives, the avoidance of impurities and poisonous substances, spoilage prevention, nutritional enrichment, and/or the grading of meat and vegetables. Relevant information includes, but is not limited to, federal regulations targeting three major areas of label abuse: deceptive definitions, misleading health claims, and untrue serving sizes and proposed standard definitions for such terms as high fiber and low fat.
</top>
<top>
<num> 371 <title> health insurance holistic
<desc> What is the extent of health insurance coverage of holistic or other non-traditional medicine/medical treatments (for example, acupuncture)?
<narr> Discussions of whether or not particular health insurance plans cover alternative treatments are relevant. Discussions of coverage for preventative programs such as health education or wellness sessions are also relevant.
</top>
<top>
<num> 372 <title> Native American casino
<desc> Identify documents that discuss the growth of Native American casino gambling.
<narr> Relevant documents include discussions regarding Native American casino gambling: its social implications, effects on local and Native American economies, and legal aspects related to Native American tribal autonomy.
</top>
<top>
<num> 373 <title> encryption equipment export
<desc> Identify documents that discuss the concerns of the United States regarding the export of encryption equipment.
<narr> Documents that merely mention the name of a company or group that produces encryption equipment but does not mention the exportation and/or commercial exploitation of the encryption equipment are not relevant. Documents which refer to governmental access into the encryption systems for the purposes of counter-intelligence or anti-crime activities are relevant.
</top>
<top>
<num> 374 <title> Nobel prize winners
<desc> Identify and provide background information on Nobel prize winners.
<narr> At a minimum, relevant documents must contain the following information: year of Nobel prize award, field of study, and recipients name. When the document announces what is obviously a current award, no year is required.
</top>
<top>
<num> 375 <title> hydrogen energy
<desc> What is the status of research on hydrogen as a feasible energy source?
<narr> A relevant document will describe progress in research on controlled hydrogen fusion or the use of hydrogen as fuel to power engines.
</top>
<top>
<num> 376 <title> World Court

<desc> What types of cases were heard by the World Court (International Court of Justice)?
<narr> Documents that report on particular cases heard by the World Court, including war tribunals created by the Court, are relevant.
</top>
<top>
<num> 377 <title> cigar smoking
<desc> Identify documents that discuss the renewed popularity of cigar smoking.
<narr> A relevant document will discuss the extent of the resurgence of cigar smoking or the social and economic issues attendant to it. Documents that discuss "Cigar Nights", "Cigar Rooms" and cigar production are relevant.
</top>
<top>
<num> 378 <title> euro opposition
<desc> Identify documents that discuss opposition to the introduction of the euro, the European currency.
<narr> A relevant document should include the countries or individuals who oppose the use of the euro and the reason(s) for their opposition to its use.
</top>
<top>
<num> 379 <title> mainstreaming
<desc> Identify documents that discuss mainstreaming children with physical or mental impairments.
<narr> A relevant document will include the pros and cons of mainstreaming children with physical or mental impairments, the benefits to the impaired child, as well as the attitude, beliefs and concerns of teachers and school administrators with regard to taking time away from the "normal children".
</top>
<top>
<num> 380 <title> obesity medical treatment
<desc> Identify documents that discuss medical treatment of obesity.
<narr> A relevant document should identify prescribed legal medications or treatments used to combat obesity and the positive or negative affects resulting from the applications.
</top>
<top>
<num> 381 <title> alternative medicine
<desc> What forms of alternative medicine are being used in the treatment of illnesses or diseases and how successful are they?
<narr> A relevant document should identify a form of alternative medicine which is being utilized in the treatment of a disease or illness, identify the illness or disease being treated, and provide an indication of the success of the procedure.
</top>
<top>
<num> 382 <title> hydrogen fuel automobiles
<desc> Identify documents that discuss the use of hydrogen as a fuel for piston driven automobiles (safe storage a concern) or the use of hydrogen in fuel cells to generate electricity to drive the car.
<narr> A relevant document may discuss either hydrogen using fuel cells to electrically power automobiles or it may discuss the safe storage of hydrogen gas in a fuel tank through the use of metal hydrides for use in piston engines.
</top>
<top>
<num> 383 <title> mental illness drugs
<desc> Identify drugs used in the treatment of mental illness.
<narr> A relevant document will include the name of a specific or generic type of drug. Generalities are not relevant.
</top>
<top>

<num> 384 <title> space station moon
<desc> Identify documents that discuss the building of a space station with the intent of colonizing the moon.
<narr> A relevant document will discuss the purpose of a space station, initiatives towards colonizing the moon, impediments which thus far have thwarted such a project, plans currently underway or in the planning stages for such a venture; cost, countries prepared to make a commitment of men, resources, facilities and money to accomplish such a feat.
</top>
<top>
<num> 385 <title> hybrid fuel cars
<desc> Identify documents that discuss the current status of hybrid automobile engines, (i.e., cars fueled by something other than gasoline only).
<narr> A relevant document may include research on non-gasoline powered engines or prototypes that may be fueled by natural gas, methanol, alcohol; cost to the consumer; health benefits derived; and shortcomings in horsepower and passenger comfort.
</top>
<top>
<num> 386 <title> teaching disabled children
<desc> What methods are currently utilized or anticipated in the teaching of disabled children?
<narr> A relevant document should identify a method or procedure currently used in the teaching of disabled children or one that is anticipated being available in the near future.
</top>
<top>
<num> 387 <title> radioactive waste
<desc> Identify documents that discuss effective and safe ways to permanently handle long-lived radioactive wastes.
<narr> Documents that discuss incineration, cementation, bitumenization, vitrification, and in underground nuclear explosion are relevant.
</top>
<top>
<num> 388 <title> organic soil enhancement
<desc> Identify documents that discuss the use of organic fertilizers (composted sludge, ash, vegetable waste, microorganisms, etc.) as soil enhancers.
<narr> The focus of the topic is on soil enhancement. Documents that discuss other uses of organic material are not relevant, nor are documents that concentrate solely on chemical fertilizers.
</top>
<top>
<num> 389 <title> illegal technology transfer
<desc> What specific entities have been accused of illegal technology transfer such as: selling their products, formulas, etc. directly or indirectly to foreign entities for other than peaceful purposes?
<narr> To be relevant, a selected document must specifically identify a person, company, or governmental entity which provided articles useful for hostile purposes to entities outside of their own country and such provision violated the laws of the provider's country.
</top>
<top>
<num> 390 <title> orphan drugs
<desc> Find documents that discuss issues associated with so-called "orphan drugs", that is, drugs that treat diseases affecting relatively few people.
<narr> A relevant document will discuss how the Orphan Drug Act is working on behalf of those who suffer from orphan diseases and conditions, or how this matter is handled in other countries.
</top>
<top>

<num> 391 <title> R&D drug prices
<desc> Identify documents that discuss the impact of the cost of research and development (R&D) on the price of drugs.
<narr> Documents that describe how any aspect of the development of a drug affects its price are relevant. Documents that discuss other factors that affect drug prices, such as advertising, without also discussing R&D costs, are not relevant.
</top>
<top>
<num> 392 <title> robotics
<desc> What are the applications of robotics in the world today?
<narr> Documents concerning R&D or testing of robotic equipment are not relevant, nor are documents discussing future applications of robotics. If the task completed seems to be merely another computer function rather than a task usually done by people, then it is not relevant.
</top>
<top>
<num> 393 <title> mercy killing
<desc> Identify documents that discuss mercy killings.
<narr> All individual cases of mercy killing are relevant, except that "letters to the editor" mentioning cases are not relevant. The removal of life support systems is relevant. A general mention or description of a case without specifics, such as victim's name are not relevant. Cases determined to be a murder-suicide are not relevant.
</top>
<top>
<num> 394 <title> home schooling
<desc> Identify documents that discuss the education of children at home (home schooling).
<narr> A relevant document would contain any data on the education of children in a home environment. Included could be data on where, by whom, community acceptance, and successes. Additional data on programs aiding the education of disadvantaged children in an away-from-school environment such as mobile vehicle classrooms would be considered as home schooling and are therefore relevant.
</top>
<top>
<num> 395 <title> tourism
<desc> Provide examples of successful attempts to attract tourism as a means to improve a local economy.
<narr> To be relevant, a selected document will specify the entity (city, state, country, governmental unit) which has achieved an economic increase due to the entity's efforts at boosting tourism. Documents which only concern plans for increasing tourism are not relevant, only documents which detail an actual increase are relevant.
</top>
<top>
<num> 396 <title> sick building syndrome
<desc> Identify documents that discuss sick building syndrome or building-related illnesses.
<narr> A relevant document would contain any data that refers to the sick building or building-related illnesses, including illnesses caused by asbestos, air conditioning, pollution controls. Work-related illnesses not caused by the building, such as carpal tunnel syndrome, are not relevant.
</top>
<top>
<num> 397 <title> automobile recalls
<desc> Identify documents that discuss the reasons for automobile recalls.
<narr> A relevant document will specify major or minor reasons for automobile recalls by car manufacturers. Documents that discuss truck recalls are not relevant.
</top>
<top>

<num> 398 <title> dismantling Europe's arsenal
<desc> Identify documents that discuss the European Conventional Arms Cut as it relates to the dismantling of Europe's arsenal.
<narr> Relevant documents may address any of the following issues: reducing conventional (non- nuclear) forces in Europe (CFE), efforts toward reduction that have begun or have been accomplished, the numbers of weapons or manpower reduced through this treaty, comparisons of numbers of reductions made by the Warsaw Pact and NATO, and the effect this arms cut is having on European nations.
</top>
<top>
<num> 399 <title> oceanographic vessels
<desc> Identify documents that discuss the activities or equipment of oceanographic vessels.
<narr> Relevant documents will contain information on the activity of oceanographic vessels regardless of nationality. Such items as undersea mapping, monitoring global warming, marine environmental monitoring and the use of deep sea robots to search for sunken ships are relevant to oceanography activity. Also relevant are such actions as using the vessel for research and development of new equipment or exploring oceanic resources. Any reference to equipment in use aboard an oceanographic vessel is relevant. Documents that identify a vessel as a survey ship which is involved in oceanographic activity is also relevant.
</top>
<top>
<num> 400 <title> Amazon rain forest
<desc> What measures are being taken by local South American authorities to preserve the Amazon tropical rain forest?
<narr> Relevant documents may identify: the official organizations, institutions, and individuals of the countries included in the Amazon rain forest; the measures being taken by them to preserve the rain forest; and indications of degrees of success in these endeavors.
</top>
<top>
<num> 401 <title> foreign minorities, Germany
<desc> What language and cultural differences impede the integration of foreign minorities in Germany?
<narr> A relevant document will focus on the causes of the lack of integration in a significant way; that is, the mere mention of immigration difficulties is not relevant. Documents that discuss immigration problems unrelated to Germany are also not relevant.
</top>
<top>
<num> 402 <title> behavioral genetics
<desc> What is happening in the field of behavioral genetics, the study of the relative influence of genetic and environmental factors on an individual's behavior or personality?
<narr> Documents describing genetic or environmental factors relating to understanding and preventing substance abuse and addictions are relevant. Documents pertaining to attention deficit disorders tied in with genetics are also relevant, as are genetic disorders affecting hearing or muscles. The genome project is relevant when tied in with behavior disorders (i.e., mood disorders, Alzheimer's disease).
</top>
<top>
<num> 403 <title> osteoporosis
<desc> Find information on the effects of the dietary intakes of potassium, magnesium and fruits and vegetables as determinants of bone mineral density in elderly men and women thus preventing osteoporosis (bone decay).
<narr> A relevant document may include one or more of the dietary intakes in the prevention of osteoporosis. Any discussion of the disturbance of nutrition and mineral metabolism that results in a decrease in bone mass is also relevant.
</top>

<top>
<num> 404 <title> Ireland, peace talks
<desc> How often were the peace talks in Ireland delayed or disrupted as a result of acts of violence?
<narr> Any interruptions to the peace process not directly attributable to acts of violence are not relevant.
</top>
<top>
<num> 405 <title> cosmic events
<desc> What unexpected or unexplained cosmic events or celestial phenomena, such as radiation and supernova outbursts or new comets, have been detected?
<narr> New theories or new interpretations concerning known celestial objects made as a result of new technology are not relevant.
</top>
<top>
<num> 406 <title> Parkinson's disease
<desc> What is being done to treat the symptoms of Parkinson's disease and keep the patient functional as long as possible?
<narr> A relevant document identifies a drug or treatment program utilized in patient care and provides an indication of success or failure.
</top>
<top>
<num> 407 <title> poaching, wildlife preserves
<desc> What is the impact of poaching on the world's various wildlife preserves?
<narr> A relevant document must discuss poaching in wildlife preserves, not in the wild itself. Also deemed relevant is evidence of preventive measures being taken by local authorities.
</top>
<top>
<num> 408 <title> tropical storms
<desc> What tropical storms (hurricanes and typhoons) have caused significant property damage and loss of life?
<narr> The date of the storm, the area affected, and the extent of damage/casualties are all of interest. Documents that describe the damage caused by a tropical storm as "slight", "limited", or "small" are not relevant.
</top>
<top>
<num> 409 <title> legal, Pan Am, 103
<desc> What legal actions have resulted from the destruction of Pan Am Flight 103 over Lockerbie, Scotland, on December 21, 1988?
<narr> Documents describing any charges, claims, or fines presented to or imposed by any court or tribunal are relevant, but documents that discuss charges made in diplomatic jousting are not relevant.
</top>
<top>
<num> 410 <title> Schengen agreement
<desc> Who is involved in the Schengen agreement to eliminate border controls in Western Europe and what do they hope to accomplish?
<narr> Relevant documents will contain any information about the actions of signatories of the Schengen agreement such as: measures to eliminate border controls (removal of traffic obstacles, lifting of traffic restrictions); implementation of the information system data bank that contains unified visa issuance procedures; or strengthening of border controls at the external borders of the treaty area in exchange for free movement at the internal borders. Discussions of border crossovers for business purposes are not relevant.
</top>
<top>
<num> 411 <title> salvaging, shipwreck, treasure

<desc> Find information on shipwreck salvaging: the recovery or attempted recovery of treasure from sunken ships.
<narr> A relevant document will provide information on the actual locating and recovery of treasure; on the technology which makes possible the discovery, location and investigation of wreckages which contain or are suspected of containing treasure; or on the disposition of the recovered treasure.
</top>
<top>
<num> 412 <title> airport security
<desc> What security measures are in effect or are proposed to go into effect in airports?
<narr> A relevant document could identify a specific airport and describe the security measures already in effect or proposed for use at that airport. Relevant items could also describe a failure of security that was cited as a contributing cause of a tragedy which came to pass or which was later averted. Comparisons between and among airports based on the effectiveness of the security of each are also relevant.
</top>
<top>
<num> 413 <title> steel production
<desc> What are new methods of producing steel?
<narr> Relevant documents will discuss the processes adapted by entrepreneurs who have organized so-called "mini-imills" and are producing steel by methods which differ from the old blast furnace method of production. Documents that identify the new companies, the problems they have have encountered, and/or their successes or failures in the national and international markets are also relevant.
</top>
<top>
<num> 414 <title> Cuba, sugar, exports
<desc> How much sugar does Cuba export and which countries import it?
<narr> A relevant document will provide information regarding Cuba's sugar trade. Sugar production statistics are not relevant unless exports are mentioned explicitly.
</top>
<top>
<num> 415 <title> drugs, Golden Triangle
<desc> What is known about drug trafficking in the "Golden Triangle", the area where Burma, Thailand and Laos meet?
<narr> A relevant document will discuss drug trafficking in the Golden Triangle, including organizations that produce or distribute the drugs; international efforts to combat the traffic; or the quantities of drugs produced in the area.
</top>
<top>
<num> 416 <title> Three Gorges Project
<desc> What is the status of The Three Gorges Project?
<narr> A relevant document will provide the projected date of completion of the project, its estimated total cost, or the estimated electrical output of the the finished project. Discussions of the social, political, or ecological impact of the project are not relevant.
</top>
<top>
<num> 417 <title> creativity
<desc> Find ways of measuring creativity.
<narr> Relevant items include definitions of creativity, descriptions of characteristics associated with creativity, and factors linked to creativity.
</top>
<top>
<num> 418 <title> quilts, income
<desc> In what ways have quilts been used to generate income?

<narr> Documents mentioning quilting books, quilting classes, quilted objects, and museum exhibits of quilts are all relevant. Documents that discuss AIDs quilts are irrelevant, unless there is specific mention that the quilts are being used for fundraising.
</top>
<top>
<num> 419 <title> recycle, automobile tires
<desc> What new uses have been developed for old automobile tires as a means of tire recycling?
<narr> A relevant document must show advantageous uses of recycled tires, such as: destructive distillation of scrap rubber for valuable chemicals, reef building for fish habitats, filler or binder in asphalt roadway mixes, and burning in a controlled environment for heat generation.
</top>
<top>
<num> 420 <title> carbon monoxide poisoning
<desc> How widespread is carbon monoxide poisoning on a global scale?
<narr> Relevant documents will contain data on what carbon monoxide poisoning is, symptoms, causes, and/or prevention. Advertisements for carbon monoxide protection products or services are not relevant. Discussions of auto emissions and air pollution are not relevant even though they can contain carbon monoxide.
</top>
<top>
<num> 421 <title> industrial waste disposal
<desc> How is the disposal of industrial waste being accomplished by industrial management throughout the world?
<narr> Documents that discuss the disposal, storage, or management of industrial waste—both standard and hazardous—are relevant. However, documents that discuss disposal or storage of nuclear or radioactive waste, or the illegal shipment or dumping of waste to avoid legal disposal methods are not relevant.
</top>
<top>
<num> 422 <title> art, stolen, forged
<desc> What incidents have there been of stolen or forged art?
<narr> Instances of stolen or forged art in any media are relevant. Stolen mass-produced things, even though they might be decorative, are not relevant (unless they are mass- produced art reproductions). Pirated software, music, movies, etc. are not relevant.
</top>
<top>
<num> 423 <title> Milosevic, Mirjana Markovic
<desc> Find references to Milosevic's wife, Mirjana Markovic.
<narr> Any mention of the Serbian president's wife is relevant, even if she is not named. She may be referred to by her nickname, Mira. A general mention of his family, without specifying his wife, is not relevant.
</top>
<top>
<num> 424 <title> suicides
<desc> Give examples of alleged suicides that aroused suspicion of the death actually being murder.
<narr> The intent of this query is to find criminal murders that are being disguised as suicide, but assisted suicides done out of compassion would be relevant if someone refers to them as murder.
</top>
<top>
<num> 425 <title> counterfeiting money
<desc> What counterfeiting of money is being done in modern times?
<narr> Relevant documents must cite actual instances of counterfeiting. Anti-counterfeiting measures by themselves are not relevant.
</top>

<top>
<num> 426 <title> law enforcement, dogs
<desc> Provide information on the use of dogs worldwide for law enforcement purposes.
<narr> Relevant items include specific information on the use of dogs during an operation. Training of dogs and their handlers are also relevant.
</top>
<top>
<num> 427 <title> UV damage, eyes
<desc> Find documents that discuss the damage ultraviolet (UV) light from the sun can do to eyes.
<narr> A relevant document will discuss diseases that result from exposure of the eyes to UV light, treatments for the damage, and/or education programs that help prevent damage. Documents discussing treatment methods for cataracts and ocular melanoma are relevant even when a specific cause is not mentioned. However, documents that discuss radiation damage from nuclear sources or lasers are not relevant.
</top>
<top>
<num> 428 <title> declining birth rates
<desc> Do any countries other than the U.S. and China have a declining birth rate?
<narr> To be relevant, a document will name a country other than the U.S. or China in which the birth rate fell from the rate of the previous year. The decline need not have occurred in more than the one preceding year.
</top>
<top>
<num> 429 <title> Legionnaires' disease
<desc> Identify outbreaks of Legionnaires' disease.
<narr> To be relevant, a document must discuss a specific outbreak of Legionnaires' disease. Documents that address prevention of or cures for the disease without citing a specific case are not relevant.
</top>
<top>
<num> 430 <title> killer bee attacks
<desc> Identify instances of attacks on humans by Africanized (killer) bees.
<narr> Relevant documents must cite a specific instance of a human attacked by killer bees. Documents that note migration patterns or report attacks on other animals are not relevant unless they also cite an attack on a human.
</top>
<top>
<num> 431 <title> robotic technology
<desc> What are the latest developments in robotic technology?
<narr> A relevant document will contain information on current applications of robotic technology. Discussions of robotics research or simulations of robots are not relevant.
</top>
<top>
<num> 432 <title> profiling, motorists, police
<desc> Do police departments use "profiling" to stop motorists?
<narr> A relevant document will report or discuss police department criteria for identifying motorists considered likely to be carrying contraband. Documents discussing the detention of individuals by foreign security forces are not relevant.
</top>
<top>
<num> 433 <title> Greek, philosophy, stoicism
<desc> Is there contemporary interest in the Greek philosophy of stoicism?
<narr> Actual references to the philosophy or philosophers, productions of Greek stoic plays, and new "stoic" artistic productions are all relevant.

</top>
<top>
<num> 434 <title> Estonia, economy
<desc> What is the state of the economy of Estonia?
<narr> Documents that give concrete economic information such as economic statistics, entering economic unions and treaties, or monetary performance are relevant, as are discussions of economic issues such as transportation or pollution.
</top>
<top>
<num> 435 <title> curbing population growth
<desc> What measures have been taken worldwide and what countries have been effective in curbing population growth?
<narr> A relevant document must describe an actual case in which population measures have been taken and their results are known. The reduction measures must have been actively pursued; that is, passive events such as disease or famine involuntarily reducing the population are not relevant.
</top>
<top>
<num> 436 <title> railway accidents
<desc> What are the causes of railway accidents throughout the world?
<narr> A relevant document provides data on railway accidents of any sort (i.e., locomotive, trolley, streetcar) where either the railroad system or the vehicle or pedestrian involved caused the accident. Documents that discuss railroading in general, new rail lines, new technology for safety, and safety and accident prevention are not relevant, unless an actual accident is described.
</top>
<top>
<num> 437 <title> deregulation, gas, electric
<desc> What has been the experience of residential utility customers following deregulation of gas and electric?
<narr> Documents that discuss privatization of government- owned utilities alone are not relevant. Also, not relevant are documents that discuss the deregulation of utilities for commercial customers.
</top>
<top>
<num> 438 <title> tourism, increase
<desc> What countries are experiencing an increase in tourism?
<narr> A relevant document will name a country that has experienced an increase in tourism. The increase must represent the nation as a whole and tourism in general, not be restricted to only certain regions of the country or to some specific type of tourism (e.g., adventure travel). Documents discussing only projected increases are not relevant.
</top>
<top>
<num> 439 <title> inventions, scientific discoveries
<desc> What new inventions or scientific discoveries have been made?
<narr> The word "new" in the description is defined as occurring in the 1990s. Documents that indicate a "recent" invention or scientific discovery are considered relevant. Discoveries made in astronomy or any scientific discoveries that are not patentable are not relevant.
</top>
<top>
<num> 440 <title> child labor
<desc> What steps are being taken by governments or corporations to eliminate abuse of child labor?
<narr> A relevant document identifies an action taken by either a private commercial corporation or governmental organization to reduce or eliminate the use of child labor in manufacturing operations.

</top>
<top>
<num> 441 <title> Lyme disease
<desc> How do you prevent and treat Lyme disease?
<narr> Documents that discuss current prevention and treatment techniques for Lyme disease are relevant. Reports of research on new treatments of the disease are also relevant.
</top>
<top>
<num> 442 <title> heroic acts
<desc> Find accounts of selfless heroic acts by individuals or small groups for the benefit of others or a cause.
<narr> Relevant documents will contain a description of specific acts. General statements concerning heroic acts are not relevant.
</top>
<top>
<num> 443 <title> U.S., investment, Africa
<desc> What is the extent of U.S. (government and private) investment in sub-Saharan Africa?
<narr> All references to U.S. Governmental and private assistance to sub-Saharan Africa are relevant. Documents discussing contributions by reason of U.S. membership in international aid organizations are also relevant.
</top>
<top>
<num> 444 <title> supercritical fluids
<desc> What are the potential uses for supercritical fluids as an environmental protection measure?
<narr> To be relevant, a document must indicate that the fluid involved is achieved by a process of pressurization producing the supercritical fluid.
</top>
<top>
<num> 445 <title> women clergy
<desc> What other countries besides the United States are considering or have approved women as clergy persons?
<narr> To be relevant, a document must indicate either a country where a woman has been installed as clergy or a country that is considering such an installation. The clergy position must be as church pastor rather than some other church capacity (e.g., nun or choir member).
</top>
<top>
<num> 446 <title> tourists, violence
<desc> Where are tourists likely to be subjected to acts of violence causing bodily harm or death?
<narr> A relevant document must contain accounts of known harm to tourists. Evidence of single, isolated incidents are not relevant.
</top>
<top>
<num> 447 <title> Stirling engine
<desc> What new developments and applications are there for the Stirling engine?
<narr> Any discussion of new developments and applications of the Stirling engine (also known as the Stirling cycle) are relevant.
</top>
<top>
<num> 448 <title> ship losses
<desc> Identify instances in which weather was a main or contributing factor in the loss of a ship at sea.
<narr> Any ship loss due to weather is relevant, either in international or coastal waters.
</top>
<top>

<num> 449 <title> antibiotics ineffectiveness
<desc> What has caused the current ineffectiveness of antibiotics against infections and what is the prognosis for new drugs?
<narr> To be relevant, a document must discuss the reasons or causes for the ineffectiveness of current antibiotics. Relevant documents may also include efforts by pharmaceutical companies and federal government agencies to find new cures, updating current testing phases, new drugs being tested, and the prognosis for the availability of new and effective antibiotics.
</top>
<top>
<num> 450 <title> King Hussein, peace
<desc> How significant a figure over the years was the late Jordanian King Hussein in furthering peace in the Middle East?
<narr> A relevant document must include mention of Israel; King Hussein himself as opposed to other Jordanian officials; discussion of the King's on-going, previous or upcoming efforts; and efforts pertinent to the peace process, not merely Jordan's relationship with other middle-east countries or the U.S.
</top>
<top> <num> 601 <title> Turkey Iraq water
<desc> What is the effect of Turkish river control projects on Iraqi water resources?
<narr> A relevant document will deal specifically with water issues between Turkey and Iraq. Other political and economic concerns between the two countries (e.g. the Kurds or water to Syria, Israel and Lebanon) are not relevant.
</top>
<top> <num> 602 <title> Czech, Slovak sovereignty
<desc> Retrieve information regarding the process by which the Czech and Slovak republics of Czechoslovakia established separate sovereign countries.
<narr> A relevant document will provide specific dates and details regarding the separation movement. Documents relating to normal activities of the separate nations, both internal and external are not relevant. </top>
<top> <num> 603 <title> Tobacco cigarette lawsuit
<desc> Retrieve documents regarding U.S. lawsuits against the tobacco industry for causing health problems and/or death from cigarettes.
<narr> Relevant documents must contain who, where and why details, and must be about a specific suit or suits. Specific documents are relevant regardless of the outcome of the suit. Documents concerning lawsuits about fraud, lack of warning, and nicotine 'spiking' are all relevant. </top>
<top> <num> 604 <title> Lyme disease arthritis
<desc> What evidence is there to link tick-borne Lyme disease with arthritis?
<narr> Documents providing evidence to support or refute the connection between Lyme disease and arthritis are relevant. Documents discussing Lyme disease and inflammatory joint disorder are also relevant. </top>
<top> <num> 605 <title> Great Britain health care
<desc> What are the pros and cons of Great Britain's universal health care system?
<narr> Documents that discuss recommendations for change or list criticisms of the current system are relevant. Documents about an individual's experience with the health care system in Great Britain are irrelevant unless the document also contains a specific recommendation or criticism. </top>
<top> <num> 606 <title> leg traps ban
<desc> Find documents that discuss banning leg traps used to capture animals.
<narr> Both 'leg traps' and 'steel-jawed traps' are relevant. Documents describing related prohibitions, such as banning the sale of furs when the animals are caught using leg traps, are also relevant. </top>
<top> <num> 607 <title> human genetic code
<desc> What progress is being made in the effort to map and sequence the human genetic code?
<narr> Documents must discuss specific progress in mapping the human genome. Documents that simply describe applications of the research, such as using DNA in criminal cases, using the genetic code to treat disease, or creating genetically engineered organisms are irrelevant. </top>

<top> <num> 608 <title> taxing social security
<desc> Find articles that discuss the pros and cons of taxing U.S. social security benefits.
<narr> Only documents that discuss income tax on social security benefits in the U.S. are relevant. Documents that discuss the social security tax itself or other countries' taxation are irrelevant. </top>
<top> <num> 609 <title> per capita alcohol consumption
<desc> Find documents that discuss per capita consumption of alcohol by political entity—country, state, province, etc.
<narr> Documents that provide per capita figures for specific beverages such as beer, wine, or spirits are relevant. However, to be relevant documents must provide specific figures on alcohol consumption: documents that simply discuss trends without providing specific figures are not relevant. Documents with figures for world per capita consumption are also relevant. </top>
<top> <num> 610 <title> minimum wage adverse impact
<desc> Find claims made by U.S. small businesses regarding the adverse impact on their businesses of raising the minimum wage.
<narr> A relevant document will include a specific reason for opposition to raising the minimum wage by U.S. small businesses. </top>
<top> <num> 611 <title> Kurds Germany violence
<desc> What violent activities have Kurds, or members of the Workers Party of Kurdistan (PKK), carried out in Germany?
<narr> Relevant documents contain information on the types of violent, destructive, or terroristic activities perpetrated by Kurds or members of the Workers Party of Kurdistan (PKK) in Germany. Information on whom or what is targeted by these kinds of acts is also relevant. Relevant activities include: killing or wounding of individuals; destruction or burning of property; and rioting. Relevant activities must be both attributed to Kurds or the PKK and occur in Germany. Violence by Kurds or the PKK against other Kurds or PKK members is relevant if it occurs in Germany, though self immolation is irrelevant. Peaceful demonstrations in Germany are irrelevant. </top>
<top> <num> 612 <title> Tibet protesters
<desc> What has been the outcome for the pro-independence protesters in Tibet who were arrested by Chinese authorities?
<narr> Documents containing information on the sentencing, releasing and/or treatment of arrested protesters are relevant. Only arrests for anti-Chinese/pro-independence (for Tibet) are relevant; criminal arrests are irrelevant. </top>
<top> <num> 613 <title> Berlin wall disposal
<desc> How were pieces of the Berlin wall disposed of after their removal?
<narr> Relevant documents contain specific instances of the fate of pieces of the Berlin wall. Example uses such as selling or taking pieces as souvenirs or using pieces in monuments are relevant. However, the simple fact of the wall's removal is not relevant, nor are schemes and plans that were not implemented. </top>
<top> <num> 614 <title> Flavr Savr tomato
<desc> Find information about the first genetically modified food product to go on the market, Flavr Savr (also Flavor Saver) Tomato developed by Calgene.
<narr> Documents about genetically engineered food in general are not relevant; relevant documents must include specifics regarding the Flavr Savr tomato. </top>
<top> <num> 615 <title> timber exports Asia
<desc> What is the extent of U.S. raw timber exports to Asia, and what effect do these exports have on the U.S. lumber industry?
<narr> Documents containing information about economic or environmental concerns related to the export of timber to Asia are relevant. Documents must specifically address exports to Asia, rather than the timber industry in general, to be relevant. </top>
<top> <num> 616 <title> Volkswagen Mexico
<desc> What is the history and extent of Volkswagen production in Mexico?
<narr> Relevant documents must contain information specific to Volkswagen production in Mexico. </top>

<top> <num> 617 <title> Russia Cuba economy
<desc> What effect has the reduction of Russian support had on the Cuban economy?
<narr> Relevant documents must contain information specific to how the reduction of Russian support has affected the Cuban economy. Documents referring to the state of the Cuban economy but not mentioning Russia are irrelevant, as are documents describing joint Russo-Cuban projects. </top>
<top> <num> 618 <title> Ayatollah Khomeini death
<desc> Find documents that describe the death of Iranian President Ayatollah Khomeini and the ramifications of his death.
<narr> Relevant documents must contain substantive information regarding the death of Khomeini; incidental mentions of his death are not relevant. Documents about the movement Khomeini led when alive are not relevant. </top>
<top> <num> 619 <title> Winnie Mandela scandal
<desc> What part did Winnie Mandela herself play in the kidnapping, beating and murder scandal in South Africa in December 1988 through January 1989?
<narr> Documents must contain some mention of what Mandela's personal involvement may have been to be relevant. Documents about her conviction on charges stemming from the scandal are relevant, though her protestations of innocence are irrelevant. </top>
<top> <num> 620 <title> France nuclear testing
<desc> How has France responded to protests against its nuclear testing in the South Pacific?
<narr> A document containing information regarding any type of protest by any one is relevant as long as France's response is also included in the document. </top>
<top> <num> 621 <title> women ordained Church of England
<desc> What are the arguments for and against Great Britain's approval of women being ordained as Church of England priests?
<narr> Documents primarily about voting on the matter by church officials and the British parliament are relevant. </top>
<top> <num> 622 <title> price fixing
<desc> Identify companies or corporations that have been accused or indicted of price fixing including the product or type of product involved.
<narr> Relevant documents must contain both the company and the product involved. Documents about price fixing by the stock market are not relevant. </top>
<top> <num> 623 <title> toxic chemical weapon
<desc> Gather any information that mentions ricin, sarin, soman, or anthrax as a toxic chemical used as a weapon.
<narr> To be relevant, a document must pertain to the use of toxic chemicals as weapons. </top>
<top> <num> 624 <title> SDI Star Wars
<desc> What are the pros and cons of developing the Strategic Defense Initiative (SDI) also known as "Star Wars"?
<narr> A relevant document must contain information regarding the technical aspects of the Strategic Defense Initiative (SDI). Documents containing simple statements that SDI is part of the U.S. defense strategy, or that discuss funding or arms reduction treaties related to SDI but do not discuss the technical aspects are irrelevant. Documents containing statements by the administration that SDI capabilities had been oversold are relevant. </top>
<top> <num> 625 <title> arrests bombing WTC
<desc> Identify documents that provide information on the arrest and/or conviction of the bombers of the World Trade Center (WTC) in February 1993.
<narr> A relevant document must contain information about an arrest or conviction related to the bombing. Simple claims of involvement are not relevant. </top>
<top> <num> 626 <title> human stampede
<desc> Find reports of human stampedes that have resulted in 20 or more deaths.
<narr> Documents that mention people being killed by animal stampedes are not relevant. Relevant documents could mention human stampedes in any public place such as a sporting event, bar, restaurant or entertainment event. </top>

<top> <num> 627 <title> Russian food crisis
<desc> What steps are being taken by the U.S. to help Russia solve the food crisis in Russia?
<narr> A relevant document must contain information specific to U.S. involvement (actual or planned) to alleviate the Russian food crisis. Documents describing Russia's internal food shortage debates, statistics depicting the severity of the shortage, or other countries' involvement are not relevant. </top>
<top> <num> 628 <title> U.S. invasion of Panama
<desc> What justification was used by the U.S. government to invade Panama, and why did some oppose the invasion?
<narr> Relevant documents must include specific reference to the U.S. government's justification for the invasion. Documents simply noting support for or criticism of the invasion by other governments are not relevant. </top>
<top> <num> 629 <title> abortion clinic attack
<desc> What is the incidence of violent attacks on abortion clinics and the doctors and staff of the clinics by anti-abortionists?
<narr> Documents describing bombings and arson attacks on an abortion clinic, or other violent, destructive actions, are relevant. Violent attacks on doctors and staff, including shootings, are also relevant. Reports of attempted attacks are relevant. Reports of non-violent activities such as trespassing, picketing, etc. are irrelevant. </top>
<top> <num> 630 <title> Gulf War Syndrome
<desc> Retrieve documents containing information about the symptoms of individuals suffering from 'Gulf War Syndrome' as a result of serving in the Gulf War.
<narr> Documents regarding law suits that claim causes of illness from service in the Gulf War are relevant, as are reports of cases resulting from contact with an ill Gulf War veteran. 'Dessert Storm Syndrome' is a synonym for the condition and is considered relevant. </top>
<top> <num> 631 <title> Mandela South Africa President
<desc> Find documents relating to the election of Nelson Mandela as president of the Republic of South Africa.
<narr> Documents must include specific information about Mandela becoming president of South Africa to be relevant. </top>
<top> <num> 632 <title> southeast Asia tin mining
<desc> What are the major tin mining countries in southeast Asia?
<narr> Relevant document contain information about tin mining in southeast Asia. Countries considered to be in southeast Asia include China, Malaysia, Indonesia, Viet Nam, Burma, Thailand and Laos. </top>
<top> <num> 633 <title> Welsh devolution
<desc> What is the history of the Welsh devolution movement?
<narr> Relevant documents contain information about some aspect of the Welsh devolution movement. </top>
<top> <num> 634 <title> L-tryptophan deaths
<desc> How many deaths are attributed to having taken tainted L-tryptophan dietary supplements?
<narr> A relevant document will describe a specific case of death attributed to tainted L=tryptophan or will give figures regarding total number of deaths. Reports of illness caused by the supplement that did not result in death are not relevant. </top>
<top> <num> 635 <title> doctor assisted suicides
<desc> What are the arguments for and against doctor assisted suicide in the U.S.?
<narr> A relevant document must contain a viewpoint regarding doctor assisted suicide to be relevant. Reports describing actual doctor assisted suicides are not relevant unless the document also presents a pro or con argument. </top>
<top> <num> 636 <title> jury duty exemptions
<desc> Find documents that discuss reasons why people may be exempted from serving on a jury.
<narr> A relevant document must include a specific reason that provides an exemption or list an occupation that causes exemption from jury duty. </top>
<top> <num> 637 <title> human growth hormone (HGH)
<desc> What are the pros and cons of adults using human growth hormone (HGH)?

<narr> A relevant document must discuss a specific application of human growth hormone in adults. Documents about using growth hormone to help short children are irrelevant. </top>
<top> <num> 638 <title> wrongful convictions
<desc> Find documents that discuss freed prisoners who have been wrongfully convicted based on faulty forensic evidence, poor police work, or false testimony.
<narr> Documents about political prisoners who were freed because of incompetent prosecutions are relevant. However, documents that discuss prisoners who are pardoned or released on bond when their convictions are overturned are not relevant, nor are documents about prisoners freed to make a political statement or prisoners freed for an exchange. </top>
<top> <num> 639 <title> consumer on-line shopping
<desc> What factors contributed to the growth of consumer on-line shopping?
<narr> A relevant document will describe a factor that has contributed to the increase use of on-line shopping by consumers. Documents containing statistical data supporting a growth factor are also relevant. </top>
<top> <num> 640 <title> maternity leave policies
<desc> What are the maternity leave policies of various governments?
<narr> Relevant documents will contain specific details about a legally-mandated maternity leave policy. Current or passed legislation is relevant, but proposed or in-process legislation is not. Vetoes of bills for maternity leave are irrelevant. Legislation that is passed, but not yet enacted (i.e. to take effect at a specific time in future) is relevant. Issues of cost to businesses and/or government are irrelevant. Statements to the effect of what benefits are not provided are relevant. </top>
<top> <num> 641 <title> Valdez wildlife marine life
<desc> What was the impact of the Exxon Valdez oil spill on the marine life and wildlife of the area?
<narr> Relevant documents must have some specific detail, such as types of animals, the kinds of problems found, and outcome (such as number of animals saved or dead). </top>
<top> <num> 642 <title> Tiananmen Square protesters
<desc> What happened to protesters arrested in connection with the Tiananmen Square demonstrations in Beijing in the spring of 1989?
<narr> Documents must be attributed to the Tiananmen Square demonstrations to be relevant: documents about protests elsewhere in response to Tiananmen Square are not relevant. Specific names, etc. are not necessary for a document to be relevant. </top>
<top> <num> 643 <title> salmon dams Pacific northwest
<desc> What harm have power dams in the Pacific northwest caused to salmon fisheries?
<narr> To be relevant, a document must include at least one negative impact of dams on the salmon population in the U.S. Pacific northwest. </top>
<top> <num> 644 <title> exotic animals import
<desc> Identify documents that discuss exotic species of animals that are imported into the U.S. or U.K.
<narr> Documents about importing birds are relevant. Documents about importing insects or pests are not relevant. Discussion of exotic animal diseases is not relevant. </top>
<top> <num> 645 <title> software piracy
<desc> Find documents that discuss the financial impact of software piracy upon the software-producing industry.
<narr> Relevant documents must specifically include software piracy: documents that lump the piracy of video, music, software, etc. together are not relevant. </top>
<top> <num> 646 <title> food stamps increase
<desc> Find documents that discuss an increase in the number of people receiving food stamp benefits.
<narr> Documents that discuss changes in the law allowing more eligible food stamp recipients are relevant. Documents that imply that a certain group of people, such as immigrants, will be eligible for food stamps are also relevant. </top>
<top> <num> 647 <title> windmill electricity
<desc> Has the use of windmill technology to generate electricity been economically productive?

<narr> A relevant document will contain a comparison of the costs of generating electricity through the use of windmills to other more conventional means of generation. </top>
<top> <num> 648 <title> family leave law
<desc> Identify documents that discuss details of a family leave law, such as how long, compensation, if any, for what reason allowed, etc.
<narr> A relevant document must contain some detail about a family leave law to be relevant. The mere mention of the existence of such a law is not relevant. </top>
<top> <num> 649 <title> computer viruses
<desc> How do computers get infected by computer viruses?
<narr> A relevant document will discuss a means by which a computer can become a host to a computer virus. </top>
<top> <num> 650 <title> tax evasion indicted
<desc> Identify individuals or corporations that have been indicted on charges of tax evasion of more than two million dollars in the U.S. or U.K.
<narr> A relevant document will contain details about large-scale tax evasion. Documents about people who lost in excess of two million dollars as a result of doing business with an organization indicted for tax fraud are relevant. </top>
<top> <num> 651
<title> U.S. ethnic population
<desc> How is the ethnic make-up of the U.S. population changing?
<narr> Documents must indicate a shift in the ethnic make-up of the U.S. population. </top>
<top> <num> 652
<title> OIC Balkans 1990s
<desc> What was the OIC's involvement in the Balkans in 1990-94?
<narr> Relevant documents describe the role the OIC played in the Balkan region. Also relevant are documents reflecting the Balkans' players (nations, groups) positions pro or con regarding OIC involvement. </top>
<top> <num> 653
<title> ETA Basque terrorism
<desc> Find documents that describe the activities of ETA, the Basque terrorist organization, in Spain.
<narr> Non-ETA politicians' comments/speculation regarding the organization are not relevant, nor are descriptions of ETA activities outside of Spain. </top>
<top> <num> 654
<title> same-sex schools
<desc> What are the advantages and disadvantages of same-sex schools?
<narr> Any discussion of the relative merits of same-sex schooling is relevant. </top>
<top> <num> 655
<title> ADD diagnosis treatment
<desc> How is Attention Deficit Disorder (ADD) diagnosed and treated in young children?
<narr> Relevant documents specifically address diagnosis and/or treatment of ADD. Simple definitions and descriptions of ADD are not relevant. Documents that discuss studies that could lead to treatment are relevant. </top>
<top> <num> 656
<title> lead poisoning children
<desc> How are young children being protected against lead poisoning from paint and water pipes?
<narr> Documents describing the extent of the problem, including suits against manufacturers and product recalls, are relevant. Descriptions of future plans for lead poisoning abatement projects are also relevant. Worker problems with lead are not relevant. Other poison hazards for children are not relevant. </top>
<top> <num> 657
<title> school prayer banned
<desc> Has prayer in U.S. schools been banned completely?

<narr> Relevant documents discuss circumstances under which some schools continue to permit prayer and court rulings on the same. Prayer in non-public schools is relevant, but prayer in schools outside the U.S. is not relevant. </top>
<top> <num> 658
<title> teenage pregnancy
<desc> Find documents that discuss teenage pregnancy in the United States: the birth rate for teenage mothers, causes and results of teenage pregnancies, and steps taken to reduce the number of teenage pregnancies.
<narr> Relevant documents will discuss teenage pregnancy in the United States. Also relevant is information on teenage abortions. </top>
<top>
<num> 659
<title> cruise health safety
<desc> What standards do cruise ships use for health and safety maintenance?
<narr> Relevant documents refer to health and safety practices and standards for recreational cruise ships. Not relevant are standards for small pleasure craft or commercial freight ships, tankers, etc. Documents referring to a specific ship's problems are not relevant. </top>
<top>
<num> 660
<title> whale watching California
<desc> Find information about whale watching off the coast of California.
<narr> Relevant documents make mention of one or more Californian site or guide services for watching whales in their habitat. </top>
<top>
<num> 661
<title> melanoma treatment causes
<desc> What are the causes and treatments for melanoma?
<narr> Relevant documents describe causes, diagnosis, and/or treatment of melanoma. Articles on research being conducted that could lead to diagnosis or treatment are also relevant. Articles on other diseases, including other skin cancers, are not relevant. </top>
<top>
<num> 662
<title> telemarketer protection
<desc> How can consumers protect against telemarketers?
<narr> Relevant are documents describing legislation, regulations, or strategies that help protect consumers against either telemarketing scams or unwanted interruptions by telemarketers. Not relevant are documents describing telemarketing as a business. General complaints about telemarketing are not relevant. "900" number calls are not relevant since they are made by the consumer. </top>
<top> <num> 663
<title> Agent Orange exposure
<desc> What were the health effects of Vietnam veterans' exposure to Agent Orange?
<narr> Only U.S. veterans' outcomes are relevant. Information on studies without results included are not relevant. </top>
<top> <num> 664
<title> American Indian Museum
<desc> What are the plans for a National Museum of the American Indian?
<narr> Relevant documents will describe the intended location, the types, and extent of the collections to be housed, and any relationships with similar museums elsewhere. </top>
<top> <num> 665
<title> poverty Africa sub-Sahara
<desc> How extensive is poverty in sub-Saharan Africa?

<narr> A relevant document must address the extent of poverty in sub-Saharan Africa. Items solely about hunger/famine conditions or debt status are not relevant. Mortality rates around the world compared to sub-Sahara Africa are not relevant. </top>
<top> <num> 666
<title> Thatcher resignation impact
<desc> Find documents that discuss the impact Prime Minister Margaret Thatchers' resignation may have on U.S. and U.K. relations.
<narr> A relevant document must specifically discuss the impact of Thatcher's resignation on relations between the U.S. and U.K. </top>
<top> <num> 667
<title> unmarried-partner households
<desc> Find documents that discuss the increasing trend toward creation of unmarried-partner households in the U.S.
<narr> Reference to any laws pertaining to such households as well as same sex households are relevant. </top>
<top> <num> 668
<title> poverty, disease
<desc> What is the relationship between poverty and disease?
<narr> Documents that do not link poverty to diseases directly but mention a link between poverty and health care are relevant. Documents that simply mention poverty and disease but do not draw a connection are not relevant. </top>
<top> <num> 669
<title> Islamic Revolution
<desc> What were the causes for the Islamic Revolution relative to relations with the U.S.?
<narr> Relevant documents must discuss the reasons that relations between the Islamic world and the United States have deteriorated. </top>
<top> <num> 670
<title> U.S. elections apathy
<desc> Why is there such apathy in U.S. elections?
<narr> Relevant documents must provide reasons for poor turnout at U.S. elections. </top>
<top> <num> 671
<title> Salvation Army benefits
<desc> Find documents that cite the specific benefits the Salvation Army provides those in need.
<narr> Relevant documents must cite specific work by the Salvation Army to help the needy. </top>
<top> <num> 672
<title> NRA membership profile
<desc> Find documents that detail the membership profile of the National Rifle Association (NRA).
<narr> Relevant documents provide details such as the age, race, or personality of NRA members. Documents that merely state the NRA position on current issues are not relevant. </top>
<top> <num> 673
<title> Soviet withdrawal Afghanistan
<desc> What factors led to the withdrawal of Soviet troops from Afghanistan?
<narr>

Documents must provide a reason for the withdrawal; documents that simply report the fact of the withdrawal are not relevant. </top>
<top> <num> 674
<title> Greenpeace prosecuted
<desc> Has Greenpeace been prosecuted or its members arrested for any of its actions?
<narr> Relevant documents concern arrest and/or prosecution of members of Greenpeace or the organization itself for actions it has taken. Suits brought by Greenpeace against others are not relevant. </top>
<top> <num> 675

<title> Olympics training swimming
<desc> Find information regarding training for Olympic swim meets.
<narr> Relevant documents will include discussion about training conditions, sites and requirements. Stories about individual competitors or about the Olympics in general are irrelevant. </top>
<top> <num> 676
<title> poppy cultivation
<desc> Find information on poppy cultivation and export worldwide.
<narr> Relevant documents contain information specific to the cultivation of poppy for exportation, including historical reports on the development of the poppy "industry". Documents regarding legal issues or programs to stop cultivation are not relevant. Incidental mention of/reference to cultivation, export, smuggling of poppies or their products in documents with other emphasis are not relevant. Editorial comment and eradication data are not relevant. Applications of U.S. companies to import poppy (for legal purposes) are not relevant. </top>
<top> <num> 677
<title> Leaning Tower of Pisa
<desc> What efforts are being made to stabilize the Leaning Tower of Pisa, and how successful have the efforts been?
<narr> Relevant documents provide discussions of the current condition of the tower, describe reinforcement measures taken, and/or provide measurements reflecting change in the tower. </top>
<top> <num> 678
<title> joint custody impact
<desc> Find information on joint/shared custody's impact on children.
<narr> Both economic and emotional impacts on children of divorce are relevant. Impact on parents is not relevant unless directly connected to the children involved. </top>
<top> <num> 679
<title> opening adoption records
<desc> Find documents that discuss the U.S. debate about the opening of sealed adoption records to adoptees.
<narr> Only documents referring to court-sanctioned adoptions are relevant. Mere descriptions of an individual's experience are not relevant. </top>
<top> <num> 680
<title> immigrants Spanish school
<desc> Find documents that discuss how the use of Spanish in U.S. schools has improved the lives of Mexican immigrants.
<narr> Documents must discuss how the use of Spanish in school improves the lives of Spanish-speaking immigrants. Documents that simply mention the use of Spanish are not relevant. </top>
<top> <num> 681
<title> wind power location
<desc> Where are wind power installations located?
<narr> Documents must provide the location of specific wind power installations. Descriptions of wind power installations, applications to construct installations, and arguments for/against wind power are all not relevant. </top>
<top> <num> 682
<title> adult immigrants English
<desc> What is being done to teach English to recently admitted adult immigrants?
<narr> Descriptions of any program to teach English to adult immigrants are relevant, as are proposals to establish such programs. Testing programs and programs targeting only children are not relevant. </top>
<top> <num> 683
<title> Czechoslovakia breakup
<desc> Find information on the breakup of Czechoslovakia into the Czech Republic and Slovakia and its social and political impact on the two countries' people.

<narr> Editorials are not relevant. Mention of economic impact and difficulties are not relevant. Pre-breakup machinations between the two parties/parts are relevant. </top>
<top> <num> 684
<title> part-time benefits
<desc> What businesses or government entities give medical or other benefits to part-time workers?
<narr> Documents that mention benefits given specifically to part-time workers are relevant. Not relevant are benefit plans not currently in effect, nor benefits given to full-time workers. </top>
<top> <num> 685
<title> Oscar winner selection
<desc> How are Oscar winners selected?
<narr> Relevant documents describe the process by which Oscar winners are determined, including the methods by which potential awardees are selected and nominated, the qualifications needed to become an eligible voter, and the number of eligible voters. </top>
<top> <num> 686
<title> Argentina pegging dollar
<desc> What are the negative impacts of Argentina's policy of pegging their peso to the U.S. dollar?
<narr> Documents must identify problems that have been created due to the pegging of the peso to the dollar. Documents that mention pegging but do not identify problems in the Argentine economy are not relevant. </top>
<top> <num> 687
<title> Northern Ireland industry
<desc> What businesses and industries form the basis of the economy of Northern Ireland?
<narr> Relevant documents name at least one business or industry that is a major employer in Northern Ireland. Documents about unemployment or industry closing are not relevant. </top>
<top> <num> 688
<title> non-U.S. media bias
<desc> What bias exists in the media of countries other than the U.S.?
<narr> Allegations or evidence of bias in the media of countries other than the U.S. are relevant. External criticism of U.S. media, i.e., bias allegations against U.S. media by external persons, is not relevant. </top>
<top> <num> 689
<title> family-planning aid
<desc> To which countries does the U.S. provide aid to support family planning, and for which countries has the U.S. refused or limited support?
<narr> Relevant documents indicate where U.S. aid supports family planning or where such aid has been denied. Discussions of why aid for family planning has been refused are also relevant. Documents that mention U.S. aid to countries, but not specifically for family planning are not relevant. Descriptions of funds for family planning in the U.S. itself are not relevant. </top>
<top> <num> 690
<title> college education advantage
<desc> Find documents which describe an advantage in hiring potential or increased income for graduates of U.S. colleges.
<narr> Relevant documents cite some advantage of a college education for job opportunities. Documents citing better opportunities for non-college vocational-training is not relevant. </top>
<top> <num> 691
<title> clear-cutting forests
<desc> What are the objections to the practice of "clear-cutting"?
<narr> Relevant documents discuss the reasons for resistance to the practice of "clear-cutting." Discusion of cutting old forests is not associated with clear-cutting and is not relevant. </top>
<top> <num> 692
<title> prostate cancer detection treatment
<desc> Find information on prostate cancer detection and treatment.

<narr> Different kinds of treatment are relevant, but only when tied to prostate cancer. The affliction or treatment of a particular (well-known) individual is not relevant, nor is treatment or detection of benign prostate enlargement (BHP). </top>
<top> <num> 693
<title> newspapers electronic media
<desc> What has been the effect of the electronic media on the newspaper industry?
<narr> Relevant documents must explicitly attribute effects to the electronic media: information about declining readership is irrelevant unless it attributes the cause to the electronic media. </top>
<top> <num> 694
<title> compost pile
<desc> How do you make a compost pile?
<narr> Relevant documents must reflect methods or procedures used to create a compost heap (e.g. dimensions, size, depth, contents or care). </top>
<top> <num> 695
<title> white collar crime sentence
<desc> What is the usual sentence for those convicted of white collar crimes?
<narr> To be relevant, a document must indicate the actual sentence imposed for a white collar crime. </top>
<top> <num> 696
<title> safety plastic surgery
<desc> Find documents that discuss the safety of or the hazards of cosmetic plastic surgery.
<narr> Relevant document must refer to a safety issue regarding an elective cosmetic procedure performed for enhancement of an individual's body image. </top>
<top> <num> 697
<title> air traffic controller
<desc> What are working conditions and pay for U.S. air traffic controllers?
<narr> Relevant documents tell something about working conditions or pay for American controllers. Documents about foreign controllers or an individual controller are not relevant. </top>
<top> <num> 698
<title> literacy rates Africa
<desc> What are literacy rates in African countries?
<narr> A relevant document will contain information about the literacy rate in an African country. General education levels that do not specifically include literacy rates are not relevant. </top>
<top> <num> 699
<title> term limits
<desc> What are the pros and cons of term limits?
<narr> Relevant documents reflect an opinion on the value of term limits with accompanying reason(s). Documents that cite the status of term limit legislation or opinions on the issue sans reasons for the opinion are not relevant. </top>
<top> <num> 700
<title> gasoline tax U.S.
<desc> What are the arguments for and against an increase in gasoline taxes in the U.S.?
<narr> Relevant documents present reasons for or against raising gasoline taxes in the U.S. Documents discussing rises or decreases in the price of gasoline are not relevant. </top>

B.3 Appendix: Topics in AP88-89

<top>
<head> Tipster Topic Description
<num> 051
<dom> Domain: International Economics

<title> Airbus Subsidies
<desc>
Document will discuss government assistance to Airbus Industrie, or mention a trade dispute between Airbus and a U.S. aircraft producer over the issue of subsidies.
<smry> Summary:
Document will discuss government assistance to Airbus Industrie, or mention a trade dispute between Airbus and a U.S. aircraft producer over the issue of subsidies.
<narr>
A relevant document will cite or discuss assistance to Airbus Industrie by the French, German, British or Spanish government(s), or will discuss a trade dispute between Airbus or the European governments and a U.S. aircraft producer, most likely Boeing Co. or McDonnell Douglas Corp., or the U.S. government, over federal subsidies to Airbus.
<con> Concept(s):
1. Airbus Industrie
2. European aircraft consortium, Messerschmitt-Boelkow-Blohm GmbH, British Aerospace PLC, Aerospatiale, Construcciones Aeronauticas S.A.
3. federal subsidies, government assistance, aid, loan, financing
4. trade dispute, trade controversy, trade tension
5. General Agreement on Tariffs and Trade (GATT) aircraft code
6. Trade Policy Review Group (TPRG)
7. complaint, objection
8. retaliation, anti-dumping duty petition, countervailing duty petition, sanctions
<fac> Factor(s):
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 052
<dom> Domain: International Economics
<title> South African Sanctions
<desc>
Document discusses sanctions against South Africa.
<smry> Summary:
Document discusses international sanctions against South Africa.
<narr>
A relevant document will discuss any aspect of South African sanctions, such as: sanctions declared/proposed by a country against the South African government in response to its apartheid policy, or in response to pressure by an individual, organization or another country; international sanctions against Pretoria imposed by the United Nations; the effects of sanctions against S. Africa; opposition to sanctions; or, compliance with sanctions by a company. The document will identify the sanctions instituted or being considered, e.g., corporate disinvestment, trade ban, academic boycott, arms embargo.
<con> Concept(s):
1. sanctions, international sanctions, economic sanctions
2. corporate exodus, corporate disinvestment, stock divestiture, ban on new investment, trade ban, import ban on South African diamonds, U.N. arms embargo, curtailment of defense contracts, cutoff of nonmilitary goods, academic boycott, reduction of cultural ties
3. apartheid, white domination, racism
4. antiapartheid, black majority rule
5. Pretoria
<fac> Factor(s):

<nat> Nationality: South Africa
</fac>
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 053
<dom> Domain: International Economics
<title> Leveraged Buyouts
<desc>
Document mentions a leveraged buyout valued at or above 200 million dollars.
<smry> Summary:
Document mentions a leveraged buyout valued at or above 200 million dollars.
<narr>
A relevant document will cite a leveraged buyout (LBO) valued at or above 200 million dollars. The LBO may be at any stage, e.g., considered, proposed, pending, a fact. The company (being) taken private must be identified. The offer may be expressed in dollars a share.
<con> Concept(s):
1. leveraged buyout, LBO
2. take private, go private
3. management-led leveraged buyout
<fac> Factor(s): <price> Price: >= 200 million dollars </fac>
<def> Definition(s):
Leveraged Buyout (LBO) - Takeover of a company using borrowed funds, with the target company's assets serving as security for the loans taken out by the acquiring firm, which repays the loans out of the cash flow of the acquired company or from the sale of the assets of the acquired firm.
</top>
<top>
<head> Tipster Topic Description
<num> 054
<dom> Domain: International Economics
<title> Satellite Launch Contracts
<desc>
Document will cite the signing of a contract or preliminary agreement, or the making of a tentative reservation, to launch a commercial satellite.
<smry> Summary:
Document will cite the signing of a contract or preliminary agreement, or the making of a tentative reservation, to launch a commercial satellite.
<narr>
A relevant document will mention the signing of a contract or preliminary agreement , or the making of a tentative reservation, to launch a commercial satellite.
<con> Concept(s):
1. contract, agreement
2. launch vehicle, rocket, payload, satellite
3. launch services, commercial space industry, commercial launch industry
4. Arianespace, Martin Marietta, General Dynamics, McDonnell Douglas
5. Titan, Delta II, Atlas, Ariane, Proton
<fac> Factor(s):
<def> Definition(s):
</top>

<top>
<head> Tipster Topic Description
<num> 055
<dom> Domain: International Economics
<title> Insider Trading
<desc>
Document discusses an insider-trading case.
<smry> Summary:
Document discusses an insider-trading case.
<narr>
A relevant document will discuss an insider-trading case, identifying the accused/the defendant(s), as well as the government doing the investigation. It will also mention at least one of the following specifics of the case: the alleged illegal activity, whether charged with providing or using insider information, possible conspirator's role in the scheme, the monetary amount of illegal profit and of damage; if pronounced guilty, then the sentence terms, e.g., monetary penalty (cash payment), time in prison, cooperation with the government, probation, community service, being barred from the industry for a certain time period.
<con> Concept(s):
1. insider-trading case
2. insider-trading scheme, illegal insider-trading activity, illegal securities transactions, white-collar crime, securities-law violations
3. insider-trading investigation, insider-trading probe
4. Securities and Exchange Commission, SEC
5. investment banking, investment banker, arbitrage, arbitrager, securities analyst
6. inside information, advance knowledge of corporate takeovers
7. leaking information, misappropriating information, leaking word, leaking news, passing a tip
8. pre-bid market activity, takeover speculation, matched-book transaction
<fac> Factor(s):
<def> Definition(s):
Insider Trading - Dealing in shares with the advantage of inside information. The owners of shares in a public company are all supposed to be equal - so if one of them knows that the company is about to go broke before the others and sells his shares while their price is still good, that is unfair. In many countries, including the U.S. and the U.K., it is also illegal. However, it is very difficult to prove that an investor is buying or selling shares on the basis of inside information.
</top>
<top>
<head> Tipster Topic Description
<num> 056
<dom> Domain: International Finance
<title> Prime (Lending) Rate Moves, Predictions
<desc>
Document will include a prediction about the prime lending rate, or will report an actual prime rate move.
<smry> Summary:
Document will include a prediction about the prime lending rate, or will report an actual prime rate move.
<narr>
A relevant document will include a prediction about the prime lending rate (national-level or major banks'), or will report a prime rate move by major banks, in response to or in anticipation of a federal/national-level action, such as a cut in the discount rate.
<con> Concept(s):
1. prime lending rate, prime rate, base rate, reference rate, minimum lending rate, bank rate, weekly bank rate
2. Federal Reserve Board, Bank of Canada, major commercial banks

3. foresee, expect, predict, speculate
4. move, increase, decrease, rise, decline, boost, cut

<fac> Factor(s):
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 057
<dom> Domain: U.S. Economics
<title> MCI
<desc>

Document will discuss how MCI has been doing since the Bell System breakup.

<smry> Summary:

Document will discuss how MCI (Multiport Communications Interface) has been doing since the Bell System breakup.

<narr>

A relevant document will discuss the financial health of MCI Communications Corp. since the breakup of the Bell System (AT&T and the seven regional Baby Bells) in January 1984. The status indicated may not necessarily be a direct or indirect result of the breakup of the system and ensuing regulation and deregulation of Ma Bell or of the restrictions placed upon the seven Bells; it may result from any number of factors, such as advances in telecommunications technology, MCI initiative, etc. MCI's financial health may be reported directly: a broad statement about its earnings or cash flow, or a report containing financial data such as a quarterly report; or it may be reflected by one or more of the following: credit ratings, share of customers, volume growth, cuts in capital spending, $$ figure net loss, pre-tax charge, analysts' or MCI's own forecast about how well they will be doing, or MCI's response to price cuts that AT & T makes at its own initiative or under orders from the Federal Communications Commission (FCC), such as price reductions, layoffs of employees out of a perceived need to cut costs, etc. Daily OTC trading stock market and monthly short interest reports are NOT relevant; the inventory must be longer term, at least quarterly.

<con> Concept(s):

1. MCI Communications Corp.
2. Bell System breakup
3. Federal Communications Commission, FCC
4. regulation, deregulation
5. profits, revenue, net income, net loss, write-downs
6. NOT daily OTC trading, NOT monthly short interest

<fac> Factor(s):

Time: after January 1984

</fac>
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 058
<dom> Domain: International Economics
<title> Rail Strikes
<desc>

Document will predict or anticipate a rail strike or report an ongoing rail strike.

<smry> Summary:

Document will predict or anticipate a rail strike or report an ongoing rail strike.

<narr>

A relevant document will either report an impending rail strike, describing the conditions which may lead to a strike, or will provide an update on an ongoing strike. To be relevant, the document will identify the location of the strike or potential strike. For an impending strike, the document will report the status of negotiations, contract talks, etc. to enable an assessment of the probability of a strike. For an ongoing strike, the document will report the length of the strike to the current date and the status of negotiations or mediation.
<con> Concept(s):

1. rail strike, picket, stoppage, lockout, walkout, wildcat
2. rail union, negotiator, railroad, federal conciliator, brotherhood
3. union proposal, talks, settlement, featherbedding, cost cutting
4. working without a contract, expired contract, cooling off period

<fac> Factor(s):
<time> Time: Current
</fac>
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 059
<dom> Domain: Environment
<title> Weather Related Fatalities
<desc>

Document will report a type of weather event which has directly caused at least one fatality in some location.
<smry> Summary:

Document will report a type of weather event which has directly caused at least one fatality in some location.
<narr>

A relevant document will include the number of people killed and injured by the weather event, as well as reporting the type of weather event and the location of the event.
<con> Concept(s):

1. lightning, avalanche, tornado, typhoon, hurricane, heat, heat wave, flood, snow, rain, downpour, blizzard, storm, freezing temperatures
2. dead, killed, fatal, death, fatality, victim
3. NOT man-made disasters, NOT war-induced famine
4. NOT earthquakes, NOT volcanic eruptions

<fac> Factor(s):
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 060
<dom> Domain: International Economics
<title> Merit-Pay vs. Seniority
<desc>

Document will describe either one or both sides of the controversy over the use of standards of performance to determine salary levels and incentive pay as contrasted with determining pay solely on the basis of seniority or longevity on the job.
<smry> Summary:

Document will describe either one or both sides of the controversy over a change in the use of standards of performance to determine salary levels and incentive pay as contrasted with determining pay solely on the basis of seniority or longevity on the job.
<narr>

A relevant document will name an employer (company, institution, or governmental entity) which has announced a change in its policy regarding merit vs. seniority in paying employees, or name an employee group (union or lobbying group) which is protesting a change in merit vs. seniority status for a particular employee group. The change announced should be a move away from one policy toward another, not an announcement of across-the-board cuts as a cost-cutting measure.
<con> Concept(s):

1. merit-pay, merit pay, performance, award program, bonus
2. seniority, longevity, cost-of-living
3. salary, pay, compensation, raise
4. appraisal, evaluation, assessment, testing, career ladder, peer review, certifying

<fac> Factor(s):
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 061
<dom> Domain: International Politics
<title> Israeli Role in Iran-Contra Affair
<desc>
Document will discuss the role of Israel in the Iran-Contra Affair.
<smry> Summary:
Document will discuss the role of Israel in the Iran-Contra Affair.
<narr>
A relevant document will discuss Israel's role in the Iran-Contra Affair. Any mention or hint of Israeli involvement in the scandal, e.g., selling arms to Iran, suggesting diverting profits from sales of arms to Iran to Nicaraguan insurgents, participating in diversion of funds to the Contras, kicking back money to Iranian officials, kidnapping relatives of top Iranian officials by the Mossad (Israeli intelligence service) to be exchanged for U.S. hostages in Lebanon, is relevant.
<con> Concept(s):

1. Iran-Contra affair, initiative, controversy, scandal
2. role of Israel, Israeli involvement
3. Israeli intermediaries, middlemen, arms dealers, government officials
4. David Kimche, Michael Ledeen, Jaacov Nimrodi

<fac> Factor(s):
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 062
<dom> Domain: Military
<title> Military Coups D'etat
<desc>
Document will report a military coup d'etat, either attempted or successful, in any country.
<smry> Summary:
Document will report a military coup d'etat, either attempted or successful, in any country.
<narr>
A relevant document will identify the country involved, the group responsible for the coup or coup attempt, the target of the coup, and the motivation of the coup plotters. It should NOT be about civilian government shake ups.
<con> Concept(s):

1. coup d'etat, coup attempt, skirmish, revolt, mutiny

2. oust, take charge, rebel, storm, surrender
3. regime
4. amnesty

<fac> Factor(s):
<Time> Time: Current
</fac>
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 063
<dom> Domain: Science and Technology
<title> Machine Translation
<desc>

Document will identify a machine translation system.

<smry> Summary:

Document will identify a machine translation system.

<narr>

A relevant document will identify a machine translation system which is being developed or marketed in any country. It will identify the developer or vendor, name the system, and identify one or more features of the system.

<con> Concept(s):

1. machine translation system
2. language, dictionary, font
3. batch, interactive, process, user interface

<fac> Factor(s):
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 064
<dom> Domain: International Relations
<title> Hostage-Taking
<desc>

Document will report an event or result of politically motivated hostage-taking.

<smry> Summary:

Document will report an event or result of politically motivated hostage-taking.

<narr>

A relevant document will report an attempt (successful or unsuccessful) to seize a hostage for some political purpose, report on hostages currently held, or report a current hostage release event. The document will identify the nationality of the victim, some identification of the perpetrators (nationality or religion), and location where the hostage-taking occurred.

<con> Concept(s):

1. kidnap, seize, hijack
2. negotiate, trade, exchange, swap
3. free, release
4. retaliate
5. captivity, terrorism
6. captors, hostages, negotiator, victims

<fac> Factor(s):
<def> Definition(s):

</top>
<top>
<head> Tipster Topic Description
<num> 065
<dom> Domain: Science and Technology
<title> Information Retrieval Systems
<desc>
Document will identify a type of information retrieval system.
<smry> Summary:
A relevant document will identify a new information retrieval system, identify the company or person marketing the system, and identify some of the characteristics of the system.
<narr>
A relevant document will identify an information retrieval system, identify the company or person marketing the system, and identify some of the characteristics of the system.
<con> Concept(s):
1. information retrieval system
2. storage, database, data, query
<fac> Factor(s):
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 066
<dom> Domain: Science and Technology
<title> Natural Language Processing
<desc>
Document will identify a type of natural language processing technology which is being developed or marketed in the U.S.
<smry> Summary:
Document will identify a type of natural language processing technology which is being developed or marketed in the U.S.
<narr>
A relevant document will identify a company or institution developing or marketing a natural language processing technology, identify the technology, and identify one or more features of the company's product.
<con> Concept(s):
1. natural language processing
2. translation, language, dictionary, font
3. software applications
<fac> Factor(s):
<nat> Nationality: U.S.
</fac>
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 067
<dom> Domain: International Relations
<title> Politically Motivated Civil Disturbances
<desc>

Document will report a current civil disturbance in any country, involving citizens of that country protesting a political position of their own country's government.
<smry> Summary:

Document will report a current civil disturbance in any country, involving citizens of that country protesting a political position of their own country's government.
<narr>

A relevant document will report the location of the disturbance, the identity of the group causing the disturbance, the nature of the disturbance, the identity of the group suppressing the disturbance and the political goals of the protesters. It should NOT be about economically-motivated civil disturbances and NOT be about a civil disturbance directed against a second country.
<con> Concept(s):

1. protest, unrest, demonstration, march, riot, clash, uprising, rally, boycott, sit-in
2. students, agitators, dissidents
3. police, riot police, troops, army, National Guard, government forces
4. NOT economically-motivated

<fac> Factor(s):
<time> Time: Current
</fac>
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 068
<dom> Domain: Science and Technology
<title> Health Hazards from Fine-Diameter Fibers
<desc>

Document will report actual studies, or even unsubstantiated concerns about the safety to manufacturing employees and installation workers of fine-diameter fibers used in insulation and other products.
<smry> Summary:

Document will report actual studies, or even unsubstantiated concerns about the safety to manufacturing employees and installation workers of fine-diameter fibers used in insulation and other products.
<narr>

A relevant document will report the type of fiber in question, and whether the fiber has been determined to be harmful or is merely suspected of being harmful.
<con> Concept(s):

1. fine-diameter fibers, glass, ceramic, mineral-wool, asbestos, cellulose
2. diminished lung capacity, cancer
3. workplace safety, OSHA

<fac> Factor(s):
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 069
<dom> Domain: International Relations
<title> Attempts to Revive the SALT II Treaty
<desc>

Document will report an attempt by the U.S. House of Representatives or a European country to revive the SALT II Treaty ceilings on weapons in order to limit President Reagan's military build up.
<smry> Summary:

Document will report an attempt by the U.S. House of Representatives or a European country to revive the SALT II Treaty ceilings on weapons in order to limit President Reagan's military build up.
<narr>
A relevant document will describe an attempt to push the U.S. Administration to adhere to the limits on weapons proposed in SALT II. To be relevant the document must identify what group is pushing the policy and what methods they are using to influence the U.S. Administration.
<con> Concept(s):
1. SALT II, treaty, ceiling, missiles
2. House, Senate, President, Europe
3. war, enemy, fight
<fac> Factor(s):
<def> Definition(s):
SALT II - Treaty signed by Carter and Brezhnev in 1979 limiting the number of several categories of strategic weapons which the Soviet Union and the U.S. would build and deploy. The treaty was never ratified by the U.S. Senate, and expired in 1985.
</top>
<top>
<head> Tipster Topic Description
<num> 070
<dom> Domain: Law and Government
<title> Surrogate Motherhood
<desc>
Document will report judicial proceedings and opinions on contracts for surrogate motherhood.
<smry> Summary:
Document will report judicial proceedings and opinions on contracts for surrogate motherhood.
<narr>
A relevant document will report legal opinions, judgments, and decisions regarding surrogate motherhood and the custody of any children which result from surrogate motherhood. To be relevant, a document must identify the case, state the issues which are or were being decided and report at least one ethical or legal question which arises from the case.
<con> Concept(s):
1. surrogate, mothers, motherhood
2. judge, lawyer, court, lawsuit, custody, hearing, opinion, finding
<fac> Factor(s):
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 071
<dom> Domain: Military
<title> Border Incursions
<desc>
Document will report incursions by land, air, or water into the border area of one country by military forces of a second country or a guerrilla group based in a second country.
<smry> Summary:
Document will report brief incursions by land, air, or water into the border area of one country by military forces of a second country or a guerrilla group based in a second country.
<narr>
A relevant document will name the invading country or group, name the invaded country, and identify the target or goal of the attacking force. The target or goal must be a military objective. It should NOT be about a war or

invasion in which troops remain on foreign soil for a sustained engagement. It should NOT be about a territorial dispute or economic dispute (such as fishing rights) which is being negotiated or argued by civilian groups.
<con> Concept(s):

1. rebels, refugees, soldiers, guerrillas
2. raid, assault, skirmish, dispute, fighting, clash, retaliation
3. border, frontier, ford, thalweg, air space, territorial limit
4. NOT sustained engagement
5. NOT fishing rights

<fac> Factor(s):
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 072
<dom> Domain: U.S. Economics
<title> Demographic Shifts in the U.S.
<desc>

Document will report a movement of people from one area of the U.S. to another for a variety of reasons. The movement of population described should be large enough to have some economic impact.
<smry> Summary:

Document will report a movement of people from one area of the U.S. to another for a variety of reasons. The movement of population described should be large enough to have some economic impact.
<narr>

A relevant document will name the location (state, county, city, town, or geographical area) which is undergoing a demographic shift, report whether the shift is a gain or loss of population, and give some indication of the magnitude of the shift such as a number, percentage, ratio, or ranking. The movement of population described should be large enough to have some impact on the economy of either the losing area, the receiving area, or both.
<con> Concept(s):

1. population
2. shift, gain, loss
3. census

<fac> Factor(s):
<nat> Nationality: U.S.
</fac>
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 073
<dom> Domain: International Economics
<title> Demographic Shifts across National Boundaries
<desc>

Document will report a movement of people from one country to another for a variety of reasons. The movement of population described should be large enough to have some economic impact.
<smry> Summary:

Document will report a movement of people from one country to another for a variety of reasons. The movement of population described should be large enough to have some economic impact.
<narr>

A relevant document will name the location (country or geographical area of a country) which is undergoing a demographic shift, report whether the shift is a gain or loss of population, and give some indication of the magnitude

of the shift such as a number, percentage, ratio, or ranking. The movement of population described should be large enough to have some impact on the economy of either the losing country, the receiving country, or both.
<con> Concept(s):
1. population
2. shift, gain, loss
3. immigration, emigration, deportation, diaspora, exile
<fac> Factor(s):
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 074
<dom> Domain: Political
<title> Conflicting Policy
<desc>
Document will cite an instance in which the U.S. government propounds two conflicting or opposing policies.
<smry> Summary:
Document will cite an instance in which the U.S. government propounds two conflicting or opposing policies.
<narr>
A relevant document will cite an instance in which the U.S. federal government propounds two policies which are in conflict with or opposition to each other, or seem at least hypocritical. These policies may be cited in the same document, or in two separate documents within one month's time. The target is conflicting policies, NOT a reversal of policy. Examples of two-faced policy of interest include: expending funds in the campaign against tobacco use while subsidizing the U.S. tobacco industry and the marketing of U.S. tobacco products abroad, allowing imports of produce treated with pesticide which is banned in the U.S., disapproving the sale of deficient infant formula in the U.S. while allowing it to be exported to the third world.
<con> Concept(s):
1. condemn, prohibit, ban
2. subsidize
<fac> Factor(s):
<nat> Nationality: U.S.
</fac>
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 075
<dom> Domain: Science and Technology
<title> Automation
<desc>
Document will identify an instance in which automation has clearly paid off, or conversely, has failed.
<smry> Summary:
Document will identify an instance in which automation has clearly paid off, or conversely, has failed.
<narr>
A relevant document will cite a case in which automation has been at least a factor in cutting costs, by either increased efficiency or a smaller payroll or both, or a case in which automation has failed for some reason in reducing costs, or had disappointing results. The case may involve an industry or a particular firm.
<con> Concept(s):
1. automation
2. increased efficiency, smaller payroll, work force reduction

3. pay off, cut costs
<fac> Factor(s):
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 076
<dom> Domain: Law and Government
<title> U.S. Constitution - Original Intent
<desc>

Document will include a discussion of, or debate over, the issue of the original intent of the U.S. Constitution, or the meaning of a particular amendment to the Constitution, or will present an individual's interpretation of an amendment.
<smry> Summary:

Document will include a discussion of, or debate over, the issue of the original intent of the U.S. Constitution, or the meaning of a particular amendment to the Constitution, or will present an individual's interpretation of an amendment.
<narr>

A relevant document will include a discussion of, or debate over, the issue of the original intent of the U.S. Constitution, or the meaning of a particular amendment to the Constitution, or will present an individual's interpretation of an amendment. (A mere reference to the problem is not relevant.)
<con> Concept(s):

1. U.S. Constitution, separation of powers, body of the Constitution, Bill of Rights, First Amendment, Fifth Amendment
2. original intent

<fac> Factor(s):
<nat> Nationality: U.S.
</fac>
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 077
<dom> Domain: Environment
<title> Poaching
<desc>

Document will report a poaching method used against a certain type of wildlife.
<smry> Summary:

Document will report specific poaching activities including identification of the type of wildlife being poached, the poaching technique or method of killing the wildlife which is used by the poacher, and the reason for the poaching.
<narr>

A relevant document will identify the type of wildlife being poached, the poaching technique or method of killing the wildlife which is used by the poacher, and the reason for the poaching (e.g. for a trophy, meat, or money). A report of poaching or someone caught for poaching without mention of the technique used and the reason is not relevant.
<con> Concept(s):

1. poaching, illegal hunting, fishing, trapping, equipment
2. territorial waters, economic zone, game preserve, refuge, park
3. arrest, impound, fine
4. issue, agreement, license, treaty, legal limit

<fac> Factor(s):
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 078
<dom> Domain: Environment
<title> Greenpeace
<desc>
Document will report activity by Greenpeace to carry out their environmental protection goals.
<smry> Summary:
Document will report activity by Greenpeace to carry out their environmental protection goals.
<narr>
A relevant document will report a protest action by Greenpeace, citing the company or country targeted by the organization, the specific target (such as a ship, train, etc.), the action carried out by Greenpeace, and the objective which Greenpeace sought to attain.
<con> Concept(s):
1. Greenpeace, environment, group, activist
2. protest, disrupt, block, harass, scuttle, trespass, confront
3. anti-nuclear, uranium, radioactive, missile
<fac> Factor(s):
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 079
<dom> Domain: International Relations
<title> FRG Political Party Positions
<desc>
Document will identify at least one German political party and report some position which the party has taken in regard to a political, social, economic, military, environmental, or international objective.
<smry> Summary:
Document will identify at least one German political party or faction and report some position which the party has taken in regard to a political, social, economic, military, environmental, or international objective.
<narr>
A relevant document will report at least one goal or objective of one of the German political parties. The position will be stated clearly enough that the reader will be able to understand where the party stands on a specific issue.
<con> Concept(s):
1. Social Democrats, SPD, Green Party, Christian Democratic Union, CDU, Reds, Green, Blacks, Christian Social Union
2. oppose, support, demand
3. proponents, advocates, allies
4. swing, shift, aim, goal
<fac> Factor(s):
<nat> Nationality: Germany
</fac>
<def> Definition(s):
</top>
<top>

<head> Tipster Topic Description
<num> 080
<dom> Domain: Political
<title> 1988 Presidential Candidates Platforms
<desc>
Document will identify something about the platform of a 1988 presidential candidate.
<smry> Summary:
Document will identify something about the platform of a 1988 presidential candidate.
<narr>
A relevant document will identify something about the platform of a 1988 presidential candidate: his campaign, plans, promises, philosophy, position on an issue, issue(s) important to him, or how he would address an issue if elected. The platform of an individual candidate is of interest; the party platform is not relevant.
<con> Concept(s):
1. Bush, Dukakis, Gephardt, Dole, Haig, Simon, Robertson, Jackson, Kemp, Babbitt, DuPont, Gore, Hart, Biden
2. platform, plans, idea, promise, position, stance, philosophy, posturing, campaign, issue
3. presidential candidate, aspirant, contender, hopeful
4. NOT Democratic party platform, NOT Republican party platform
<fac> Factor(s):
<nat> Nationality: U.S.
</fac>
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 081
<dom> Domain: Finance
<title> Financial crunch for televangelists in the wake of the PTL scandal
<desc>
Document will report a loss of revenue of a televangelist in the aftermath of the PTL scandal, or a financial crisis triggered by the scandal.
<smry> Summary:
Document will report a loss of revenue of a televangelist in the aftermath of the PTL scandal, or a financial crisis triggered by the scandal.
<narr>
A relevant document will identify a religious broadcaster, and report a specific dollar amount or a percentage loss, indicating a decline in revenue suffered by that broadcaster as a result of consumer reaction to the Jim Bakker scandal.
<con> Concept(s):
1. donations, confidence, checks, money, dollars
2. plead, beg, cry
3. down, fall, slip
4. religious broadcasters, televangelists
5. Bakker, Falwell, Robertson, Swaggart, Schuller, Roberts
<fac> Factor(s):
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 082

<dom> Domain: Science and Technology
<title> Genetic Engineering
<desc>

Document discusses a genetic engineering application, a product that has been, is being, or will be developed by genetic manipulation, or attitudes toward genetic engineering.
<smry> Summary:

Document discusses a genetic engineering application, a product that has been, is being, or will be developed by genetic manipulation, or attitudes toward genetic engineering.
<narr>

A relevant document will discuss a product, e.g., drug, microorganism, vaccine, animal, plant, agricultural product, developed by genetic engineering techniques; identify an application, such as to clean up the environment or human gene therapy for a specific problem; or, present human attitudes toward genetic engineering.
<con> Concept(s):

1. genetic engineering, molecular manipulation
2. biotechnology
3. genetically engineered product: plant, animal, drug, microorganism, vaccine, agricultural product
4. cure a disease, clean up the environment, increase agricultural productivity

<fac> Factor(s):
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 083
<dom> Domain: Environment
<title> Measures to Protect the Atmosphere
<desc>

Document will discuss plans, projects, agreements, treaties, or laws concluded or under consideration aimed at protecting the atmosphere.
<smry> Summary:

Document will discuss plans, projects, agreements, treaties, or laws concluded or under consideration aimed at protecting the atmosphere.
<narr>

A relevant document will discuss specific steps taken or under consideration at the multilateral, bilateral, or national level to protect the atmosphere. The plan, project, agreement, treaty, or law discussed in the document will identify steps taken or to be taken under the measure and its purpose (limiting chlorofluorocarbon emissions, reducing acid rain, containing carbon dioxide buildup, or blocking deforestation) and not focus on scientific debate or research into such atmospheric phenomena as ozone depletion and global warming.
<con> Concept(s):

1. Greenhouse effect, global warming, carbon dioxide buildup
2. Acid rain, smokestack emissions
3. Deforestation
4. Chlorofluorocarbons

<fac> Factor(s):
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 084
<dom> Domain: U.S. Economics
<title> Alternative/renewable Energy Plant & Equipment Installation

<desc>
Document will identify planned or under-construction alternative/renewable energy plant or equipment installations.
<smry> Summary:
Document will identify planned or under-construction alternative/renewable energy plant or equipment installations.
<narr>
A relevant document will identify a specific alternative/renewable energy plant or equipment installation planned or under-construction, to include the technology employed, projected capacity or output, location, and specific financial information, e.g. total cost, source of funding, estimated unit production costs, etc. Document will not focus on political discussions of developing sources of alternate energy, nor will it focus on macroeconomic discussions of national energy needs.
<con> Concept(s):
1. Alternative energy, renewable energy
2. Cogeneration, hydro power, solar energy, windmills, geothermal, biomass
<fac> Factor(s):
<nat> Nationality: U.S.
<time> Time: future
</fac>
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 085
<dom> Domain: Law and Government
<title> Official Corruption
<desc>
Document will discuss allegations, or measures being taken against, corrupt public officials of any governmental jurisdiction worldwide.
<smry> Summary:
Document will discuss allegations, or measures being taken against, corrupt public officials of any governmental jurisdiction worldwide.
<narr>
A relevant document will discuss charges or actions being taken against corrupt public officials (be they elected, appointed, or career civil servant) anywhere in the world. The allegations or charges must be specific, e.g. bribes taken >from a named group or individual with a given objective, rather than generalized allegations of endemic political corruption, or moves against corporate or private malfeasance (unless linked to an official corruption case).
<con> Concept(s):
1. Official corruption, public corruption
2. Bribery, prostitution, cover-up, undercover investigation, influence-peddling, illegal payment, collusion
<fac> Factor(s):
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 086
<dom> Domain: Finance
<title> Bank Failures
<desc>

Document will identify recent bank failures within the United States, specifying name, location, assets, and federal or state authorities taking action.
<smry> Summary:
Document will identify recent closings by the Federal Deposit Insurance Corporation of a failed commercial bank chartered within the U.S.
<narr>
A relevant document, in addition to giving specific identifying information on a failed bank (or banks), must be clear that the bank was chartered within the United State and that it is a bank, as opposed to another financial institution such as a savings and loan or credit union, and must specify the actions taken or being taken by the public authorities.
<con> Concept(s):
1. Bank failure
2. Federal Deposit Insurance Corporation, FDIC, Comptroller of the Currency, state banking authorities, state bank regulators
<fac> Factor(s):
<nat> Nationality: U.S.
<time> Time: current
</fac>
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 087
<dom> Domain: Law & Government
<title> Criminal Actions Against Officers of Failed Financial Institutions
<desc>
Document will report on current criminal actions against officers of a failed U.S. financial institution.
<smry> Summary:
Document will report on current criminal actions against officers of a failed U.S. banking institution (including all categories of banking).
<narr>
To be relevant, a document will discuss a recently concluded or pending criminal case being pursued by U.S. federal or state authorities against officers of a failed financial institution. Civil actions, such as creditor claims filed for recovery of losses, are not relevant. The failed institution may have been insured by either federal or state funds and may belong to any category of financial institution, e.g. bank, savings and loan, credit union, etc. Although not essential, it would be preferable if the document provided such details on the case as specific crime, losses to insurance funds, penalties imposed, etc.
<con> Concept(s):
1. FDIC, bank, thrift, savings & loan, bank
2. suit, trial, restitution, default, conspiracy, audit, fraud, felony
<fac> Factor(s):
<nat> Nationality: U.S.
<time> Time: current
</fac>
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 088
<dom> Domain: International Economics

<title> Crude Oil Price Trends
<desc>
Document will provide quantitative data usable in charting international crude oil price trends.
<smry> Summary:
Document will provide quantitative data usable in charting international crude oil price trends.
<narr>
To be relevant, document must contain international crude oil production and price data in at least two out of the three following categories: reported daily output by any member of the Organization of Petroleum Exporting Countries (OPEC) or the larger non-OPEC producers (e.g. U.S., USSR, Mexico, Norway, and Great Britain); the official production quota set by OPEC for any of its members; and, prices in at least one of the so-called "benchmark" crude oil markets (specifically, West Texas Intermediate, North Sea Brent, Rotterdam crude oil spot market, and crude oil futures on the New York Mercantile Exchange). Documents referencing OPEC and its members likely will report or speculate about political and diplomatic moves related to the "official" OPEC price, but such information by itself is not directly relevant. The best documents would be those which summarize and statistically analyze crude oil production and price data over a period of time, rather than those which simply report daily market numbers.
<con> Concept(s):
1. Organization of Petroleum Exporting Countries, OPEC
2. Algeria, Ecuador, Gabon, Indonesia, Iran, Iraq, Kuwait, Libya, Nigeria, Qatar, Saudi Arabia, United Arab Emirates (UAE), Venezuela.
3. U.S., USSR, Mexico, Norway, Great Britain
3. West Texas Intermediate, North Sea Brent, Rotterdam spot, New York Mercantile Exchange.
<fac> Factor(s):
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 089
<dom> Domain: International Economics
<title> "Downstream" Investments by OPEC Member States
<desc>
Document must identify an existing or pending investment by an OPEC member state in any "downstream" operation.
<smry> Summary:
Document must identify an existing or pending investment by an OPEC member state in any "downstream" operation.
<narr>
To be relevant, a document must identify an existing or pending investment by an OPEC member state (or its national oil company) in any "downstream" installation or enterprise. The investment may be through a joint venture, acquisition, construction, or stock purchase of any operation in the refining, petrochemical, oil industry equipment manufacture, drilling and exploration, shipping, marketing and retail sales, or other ancillary "downstream" activity.
<con> Concept(s):
1. OPEC, Organization of Petroleum Exporting Countries
2. Algeria, Ecuador, Gabon, Indonesia, Iran, Iraq, Kuwait, Libya, Nigeria, Qatar, Saudi Arabia, United Arab Emirates (UAE), Venezuela.
3. Kuwait Petroleum Co., Kuwait Investment Co.
4. Petroleos de Venezuela
<fac> Factor(s):
<def> Definition(s):
1. "Downstream:" any petroleum industry activity which occurs after initial production of crude oil, or which supports such production.

</top>
<top>
<head> Tipster Topic Description
<num> 090
<dom> Domain: International Economics
<title> Data on Proven Reserves of Oil & Natural Gas Producers
<desc>

Document will provide totals or specific data on changes to the proven reserve figures for any oil or natural gas producer.

<smry> Summary:

Document will provide totals or specific data on changes to the proven reserve figures for any oil or natural gas producer.

<narr>

To be relevant, a document must give specific data on totals, additions, subtractions, recalculations, or adjustments to the official proven reserve figures used by any national or corporate entity which produces oil or natural gas. The document should specify whether a change is due to such factors as new investment in drilling activity or oil field installations, depletions or losses, recalculations of existing figures, acquisition of other producers, etc.

<con> Concept(s):

1. Proven reserve
2. Oil, natural gas

<fac> Factor(s):

<def> Definition(s):

1. Proven Reserve: although not always used with precision or industry-wide agreement on precise parameters, and thus subject to change or redefinition, a "proven reserve" can be distinguished from a probable or potential reserve as one which is ready for production near-term. Oil or natural gas which can be economically extracted using installed equipment employing current technology is considered to be a proven reserve.

</top>
<top>
<head> Tipster Topic Description
<num> 091
<dom> Domain: Military
<title> U.S. Army Acquisition of Advanced Weapons Systems
<desc>

Document will identify acquisition by the U.S. Army of specified advanced weapons systems.

<smry> Summary:

Document will identify acquisition by the U.S. Army of specified advanced weapons systems.

<narr>

To be relevant, document will provide such specific information as contracts concluded, payments made to contractors, or actual deliveries to the U.S. Army of the following advanced weapons systems: Abrams Tank (M-1), Apache Helicopter (AH-64), Patriot Missile, Blackhawk Helicopter, or Bradley Fighting Vehicle. Congressional discussions of such systems, DoD budget proposals, RDT&E activities, overseas sales, component/spare parts/ancilliary equipment/services acquisitions, and system modifications are not relevant. In other words, document must report actual or agreed-to deliveries to the U.S. Army of core elements of their most modern weapons systems.

<con> Concept(s)

1. U.S. Army
2. Abrams Tank, M-1
3. Apache Helicopter, AH-64
4. Patriot Missile
5. Blackhawk Helicopter
6. Bradley Fighting Vehicle

<fac> Factor(s):
<time> Time: current and future
<nat> Nationality: U.S.
</fac>
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 092
<dom> Domain: Military
<title> International Military Equipment Sales
<desc>

Document will identify a proposed or recently concluded sale of military equipment in the international arms market.

<smry> Summary:

Document will identify a proposed or recently concluded official sale of military equipment between countries in the international arms market.

<narr>

To be relevant, document will identify a proposed or recently concluded sale of military equipment, to include weapons platforms, munitions, spares, ancilliary equipment, technology, and services. The equipment and quantities must be identified, and recent should be defined as within the previous 12 months, but the price and conditions of sale are optional information. Any international sale, including transfers between NATO members or equipment being provided under military aid programs, is relevant. However, political discussions of international military equipment sales, such as debate of the Iran-Contra Affair, discussion of the UN embargo against South Africa, or Congressional expressions of hostility against equipment sales to Arab states, unless linked to a specific transaction, are not relevant.

<con> Concept(s):

1. FMS, foreign military sales

<fac> Factor(s):
<time> Time: current </fac>
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 093
<dom> Domain: Politics
<title> What Backing Does the National Rifle Association Have?
<desc>

Document must describe or identify supporters of the National Rifle Association (NRA), or its assets.

<smry> Summary:

Document must describe or identify members or supporters of the National Rifle Association (NRA), or discuss its finances.

<narr>

To be relevant, a document must describe or name individuals or organizations who are members of the NRA, or who contribute money to it. A document is also relevant if it quantifies the NRA's financial assets or identifies any other NRA holdings.

<con> Concept(s):

1. National Rifle Association, NRA
2. contributor, member, supporter
3. holdings, assets, finances

<fac> Factor(s):
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 094
<dom> Domain: Science and Technology
<title> Computer-aided Crime
<desc>
Document must identify a crime perpetrated with the aid of a computer.
<smry> Summary:
Document must identify a criminal activity involving computers.
<narr>
To be relevant, a document must describe an illegal activity which was carried out with the aid of a computer, either used as a planning tool, such as in target research; or used in the conduct of the crime, such as by illegally gaining access to someone else's computer files. A document is NOT relevant if it merely mentions the illegal spread of a computer virus or worm. However, a document WOULD be relevant if the computer virus/worm were used in conjunction with another crime, such as extortion.
<con> Concept(s):
1. crime, fraud
2. illegal access, corporate spying, extortion
<fac> Factor(s):
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 095
<dom> Domain: Science and Technology
<title> Computer-aided Crime Detection
<desc>
Document must describe a computer application to crime solving.
<smry> Summary:
Document must describe an actual or theoretical computer application used in crime solving.
<narr>
To be relevant, a document must describe either an actual or a theoretical computer application to detective work, by the police or by another law enforcement organization. A relevant document could include techniques such as profiling criminals and their methods of operation, identifying finger prints, spotting anomalies, etc.
<con> Concept(s):
1. police, detective, sleuth, enforcement agency
2. clue, records, fingerprints, methods
<fac> Factor(s):
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 096
<dom> Domain: Science and Technology
<title> Computer-Aided Medical Diagnosis
<desc>
Document must describe computer programs, or computerized equipment, which aid in medical diagnosis.

<smry> Summary:

Document must identify computer programs, or computerized equipment, which are in operational use in medical diagnosis.

<narr>

To be relevant, a document must describe actual computer software or hardware which is in operational use for the purpose of diagnosing, or assisting in diagnosing, patients' medical conditions.

<con> Concept(s):

1. diagnosis, scanning, testing
2. tomagraph, digital subtraction angiography, magnetic resonance scanners

<fac> Factor(s):

<def> Definition(s):

</top>

<top>

<head> Tipster Topic Description

<num> 097

<dom> Domain: Science and Technology

<title> Fiber Optics Applications

<desc>

Document must identify instances of fiber optics technology actually in use.

<smry> Summary:

Document must identify instances of fiber optics technology actually in use or contracted for.

<narr>

To be relevant, a document must describe actual operational situations in which fiber optics are being employed, or will be employed. A document describing future fiber optics use will be relevant only if contracts have been signed concerning the future application.

<con> Concept(s):

1. fiber optic, light
2. telephone, LAN, television

<fac> Factor(s):

<def> Definition(s):

1. Fiber optics refers to technology in which information is passed via laser light transmitted through glass or plastic fibers.

</top>

<top>

<head> Tipster Topic Description

<num> 098

<dom> Domain: Science and Technology

<title> Fiber Optics Equipment Manufacturers

<desc>

Document must identify individuals or organizations which produce fiber optics equipment.

<smry> Summary:

Document must identify individuals or organizations which produce fiber optics equipment or systems.

<narr>

To be relevant, a document must identify by name either individuals or companies which manufacture equipment or materials used in fiber optics technology.

<con> Concept(s):

1. Fiber optics, lasers
2. cables, connectors, fibers

<fac> Factor(s):

<def> Definition(s):

1. Fiber optics refers to the technology by which information is passed via laser light transmitted through glass or plastic fibers.
</top>
<top>
<head> Tipster Topic Description
<num> 099
<dom> Domain: International Politics
<title> Iran-Contra Affair
<desc>
Document will identify a development in the Iran-Contra Affair.
<smry> Summary:
Document will identify a development in the Iran-Contra Affair.
<narr>
A relevant document will identify a development in the unfolding saga termed the Iran-Contra Affair. A development consists of an agent and an action. Of interest are at least piecemeal disclosures of who did what in order to be able to come up with a description and chronology of events so that a determination can be made of who was behind the scheme.
<con> Concept(s):
1. Iran-Contra affair, controversy, scandal
2. President Reagan, George Bush, Oliver North, John Poindexter, Richard Secord, Robert McFarlane, Albert Hakim, Manucher Ghorbanifar, Hashemi Rafsanjani, Alexander Haig, George Shultz, Caspar Weinberger, Don Regan, William Casey, Michael Ledeen, David Kimche, Fawn Hall, Edward de Garay, Robert Dutton, Adnan Khashoggi, Benjamin Weir, William Buckley, White House
3. Iran, Nicaragua, Lebanon
4. Tower Commission, John Tower, Tower report, Brent Scowcroft, Edmund Muskie
5. independent counsel, Lawrence Walsh
6. management style, hands-off management
7. arms sales to Iran
8. diversion of profits to Nicaraguan rebels, insurgents, Contras, arms-for-hostages, kickbacks, kickback scheme
9. Swiss bank accounts, records
10. sharing U.S. intelligence secrets with Tehran
11. Israeli intermediaries, front companies
12. contributions to the Contras from Brunei, Saudi Arabia, Taiwan
<fac> Factor(s):
<def> Definition(s):
</top>
<top>
<head> Tipster Topic Description
<num> 100
<dom> Domain: International Relations
<title> Controlling the Transfer of High Technology
<desc>
Document will identify efforts by the non-communist, industrialized states to regulate the transfer of high-tech goods or "dual-use" technologies to undesirable nations.
<smry> Summary:
Document will identify efforts by the non-communist, industrialized states to regulate the transfer of high-tech goods or "dual-use" technologies to undesirable nations.
<narr>
A relevant document will identify activities of the U.S. Departments of Commerce, Defense, and State to set policy or regulate exports of strategic goods and technologies, or it will report on similar activities by allied

nations, or it will discuss technology transfer policy issues and determinations under consideration by the Paris-based COCOM. Unless explicitly related to policy issues or establishment of regulations aimed at controlling transfers of high technology goods and technologies, such events as investigations of alleged violations of COCOM rules or arms shipments in violation of embargoes are not relevant.
<con> Concept(s):
1. technology transfer, export license, export controls, strategic goods, international trade security policy, high technology
2. COCOM, Coordinating Committee on Multilateral Export Controls, Japan, U.S., NATO
3. Toshiba Machine, Kongsberg Vaapenfabrik, C. Itoh
<fac> Factor(s):
<def> Definition(s):
1. "dual-use" technology: a technology usable in both non-military and military products
</top>

B.4 Appendix: Topics in WT2G

<top>
<num> Number: 401 <title> foreign minorities, Germany
<desc> Description: What language and cultural differences impede the integration of foreign minorities in Germany?
<narr> Narrative: A relevant document will focus on the causes of the lack of integration in a significant way; that is, the mere mention of immigration difficulties is not relevant. Documents that discuss immigration problems unrelated to Germany are also not relevant.
</top>
<top>
<num> Number: 402 <title> behavioral genetics
<desc> Description: What is happening in the field of behavioral genetics, the study of the relative influence of genetic and environmental factors on an individual's behavior or personality?
<narr> Narrative: Documents describing genetic or environmental factors relating to understanding and preventing substance abuse and addictions are relevant. Documents pertaining to attention deficit disorders tied in with genetics are also relevant, as are genetic disorders affecting hearing or muscles. The genome project is relevant when tied in with behavior disorders (i.e., mood disorders, Alzheimer's disease).
</top>
<top>
<num> Number: 403 <title> osteoporosis
<desc> Description: Find information on the effects of the dietary intakes of potassium, magnesium and fruits and vegetables as determinants of bone mineral density in elderly men and women thus preventing osteoporosis (bone decay).
<narr> Narrative: A relevant document may include one or more of the dietary intakes in the prevention of osteoporosis. Any discussion of the disturbance of nutrition and mineral metabolism that results in a decrease in bone mass is also relevant.
</top>
<top>
<num> Number: 404 <title> Ireland, peace talks
<desc> Description: How often were the peace talks in Ireland delayed or disrupted as a result of acts of violence?
<narr> Narrative: Any interruptions to the peace process not directly attributable to acts of violence are not relevant.
</top>
<top>
<num> Number: 405 <title> cosmic events

<desc> Description: What unexpected or unexplained cosmic events or celestial phenomena, such as radiation and supernova outbursts or new comets, have been detected?
<narr> Narrative: New theories or new interpretations concerning known celestial objects made as a result of new technology are not relevant.
</top>
<top>
<num> Number: 406 <title> Parkinson's disease
<desc> Description: What is being done to treat the symptoms of Parkinson's disease and keep the patient functional as long as possible?
<narr> Narrative: A relevant document identifies a drug or treatment program utilized in patient care and provides an indication of success or failure.
</top>
<top>
<num> Number: 407 <title> poaching, wildlife preserves
<desc> Description: What is the impact of poaching on the world's various wildlife preserves?
<narr> Narrative: A relevant document must discuss poaching in wildlife preserves, not in the wild itself. Also deemed relevant is evidence of preventive measures being taken by local authorities.
</top>
<top>
<num> Number: 408 <title> tropical storms
<desc> Description: What tropical storms (hurricanes and typhoons) have caused significant property damage and loss of life?
<narr> Narrative: The date of the storm, the area affected, and the extent of damage/casualties are all of interest. Documents that describe the damage caused by a tropical storm as "slight", "limited", or "small" are not relevant.
</top>
<top>
<num> Number: 409 <title> legal, Pan Am, 103
<desc> Description: What legal actions have resulted from the destruction of Pan Am Flight 103 over Lockerbie, Scotland, on December 21, 1988?
<narr> Narrative: Documents describing any charges, claims, or fines presented to or imposed by any court or tribunal are relevant, but documents that discuss charges made in diplomatic jousting are not relevant.
</top>
<top>
<num> Number: 410 <title> Schengen agreement
<desc> Description: Who is involved in the Schengen agreement to eliminate border controls in Western Europe and what do they hope to accomplish?
<narr> Narrative: Relevant documents will contain any information about the actions of signatories of the Schengen agreement such as: measures to eliminate border controls (removal of traffic obstacles, lifting of traffic restrictions); implementation of the information system data bank that contains unified visa issuance procedures; or strengthening of border controls at the external borders of the treaty area in exchange for free movement at the internal borders. Discussions of border crossovers for business purposes are not relevant.
</top>
<top>
<num> Number: 411 <title> salvaging, shipwreck, treasure
<desc> Description: Find information on shipwreck salvaging: the recovery or attempted recovery of treasure from sunken ships.
<narr> Narrative: A relevant document will provide information on the actual locating and recovery of treasure; on the technology which makes possible the discovery, location and investigation of wreckages which contain or are suspected of containing treasure; or on the disposition of the recovered treasure.
</top>

<top>
<num> Number: 412 <title> airport security
<desc> Description: What security measures are in effect or are proposed to go into effect in airports?
<narr> Narrative: A relevant document could identify a specific airport and describe the security measures already in effect or proposed for use at that airport. Relevant items could also describe a failure of security that was cited as a contributing cause of a tragedy which came to pass or which was later averted. Comparisons between and among airports based on the effectiveness of the security of each are also relevant.
</top>
<top>
<num> Number: 413 <title> steel production
<desc> Description: What are new methods of producing steel?
<narr> Narrative: Relevant documents will discuss the processes adapted by entrepreneurs who have organized so-called "minimills" and are producing steel by methods which differ from the old blast furnace method of production. Documents that identify the new companies, the problems they have have encountered, and/or their successes or failures in the national and international markets are also relevant.
</top>
<top>
<num> Number: 414 <title> Cuba, sugar, exports
<desc> Description: How much sugar does Cuba export and which countries import it?
<narr> Narrative: A relevant document will provide information regarding Cuba's sugar trade. Sugar production statistics are not relevant unless exports are mentioned explicitly.
</top>
<top>
<num> Number: 415 <title> drugs, Golden Triangle
<desc> Description: What is known about drug trafficking in the "Golden Triangle", the area where Burma, Thailand and Laos meet?
<narr> Narrative: A relevant document will discuss drug trafficking in the Golden Triangle, including organizations that produce or distribute the drugs; international efforts to combat the traffic; or the quantities of drugs produced in the area.
</top>
<top>
<num> Number: 416 <title> Three Gorges Project
<desc> Description: What is the status of The Three Gorges Project?
<narr> Narrative: A relevant document will provide the projected date of completion of the project, its estimated total cost, or the estimated electrical output of the the finished project. Discussions of the social, political, or ecological impact of the project are not relevant.
</top>
<top>
<num> Number: 417 <title> creativity
<desc> Description: Find ways of measuring creativity.
<narr> Narrative: Relevant items include definitions of creativity, descriptions of characteristics associated with creativity, and factors linked to creativity.
</top>
<top>
<num> Number: 418 <title> quilts, income
<desc> Description: In what ways have quilts been used to generate income?
<narr> Narrative: Documents mentioning quilting books, quilting classes, quilted objects, and museum exhibits of quilts are all relevant. Documents that discuss AIDs quilts are irrelevant, unless there is specific mention that the quilts are being used for fundraising.
</top>

<top>
<num> Number: 419 <title> recycle, automobile tires
<desc> Description: What new uses have been developed for old automobile tires as a means of tire recycling?
<narr> Narrative: A relevant document must show advantageous uses of recycled tires, such as: destructive distillation of scrap rubber for valuable chemicals, reef building for fish habitats, filler or binder in asphalt roadway mixes, and burning in a controlled environment for heat generation.
</top>
<top>
<num> Number: 420 <title> carbon monoxide poisoning
<desc> Description: How widespread is carbon monoxide poisoning on a global scale?
<narr> Narrative: Relevant documents will contain data on what carbon monoxide poisoning is, symptoms, causes, and/or prevention. Advertisements for carbon monoxide protection products or services are not relevant. Discussions of auto emissions and air pollution are not relevant even though they can contain carbon monoxide.
</top>
<top>
<num> Number: 421 <title> industrial waste disposal
<desc> Description: How is the disposal of industrial waste being accomplished by industrial management throughout the world?
<narr> Narrative: Documents that discuss the disposal, storage, or management of industrial waste—both standard and hazardous—are relevant. However, documents that discuss disposal or storage of nuclear or radioactive waste, or the illegal shipment or dumping of waste to avoid legal disposal methods are not relevant.
</top>
<top>
<num> Number: 422 <title> art, stolen, forged
<desc> Description: What incidents have there been of stolen or forged art?
<narr> Narrative: Instances of stolen or forged art in any media are relevant. Stolen mass-produced things, even though they might be decorative, are not relevant (unless they are mass- produced art reproductions). Pirated software, music, movies, etc. are not relevant.
</top>
<top>
<num> Number: 423 <title> Milosevic, Mirjana Markovic
<desc> Description: Find references to Milosevic's wife, Mirjana Markovic.
<narr> Narrative: Any mention of the Serbian president's wife is relevant, even if she is not named. She may be referred to by her nickname, Mira. A general mention of his family, without specifying his wife, is not relevant.
</top>
<top>
<num> Number: 424 <title> suicides
<desc> Description: Give examples of alleged suicides that aroused suspicion of the death actually being murder.
<narr> Narrative: The intent of this query is to find criminal murders that are being disguised as suicide, but assisted suicides done out of compassion would be relevant if someone refers to them as murder.
</top>
<top>
<num> Number: 425 <title> counterfeiting money
<desc> Description: What counterfeiting of money is being done in modern times?
<narr> Narrative: Relevant documents must cite actual instances of counterfeiting. Anti-counterfeiting measures by themselves are not relevant.
</top>
<top>
<num> Number: 426 <title> law enforcement, dogs
<desc> Description: Provide information on the use of dogs worldwide for law enforcement purposes.

<narr> Narrative: Relevant items include specific information on the use of dogs during an operation. Training of dogs and their handlers are also relevant.
</top>
<top>
<num> Number: 427 <title> UV damage, eyes
<desc> Description: Find documents that discuss the damage ultraviolet (UV) light from the sun can do to eyes.
<narr> Narrative: A relevant document will discuss diseases that result from exposure of the eyes to UV light, treatments for the damage, and/or education programs that help prevent damage. Documents discussing treatment methods for cataracts and ocular melanoma are relevant even when a specific cause is not mentioned. However, documents that discuss radiation damage from nuclear sources or lasers are not relevant.
</top>
<top>
<num> Number: 428 <title> declining birth rates
<desc> Description: Do any countries other than the U.S. and China have a declining birth rate?
<narr> Narrative: To be relevant, a document will name a country other than the U.S. or China in which the birth rate fell from the rate of the previous year. The decline need not have occurred in more than the one preceding year.
</top>
<top>
<num> Number: 429 <title> Legionnaires' disease
<desc> Description: Identify outbreaks of Legionnaires' disease.
<narr> Narrative: To be relevant, a document must discuss a specific outbreak of Legionnaires' disease. Documents that address prevention of or cures for the disease without citing a specific case are not relevant.
</top>
<top>
<num> Number: 430 <title> killer bee attacks
<desc> Description: Identify instances of attacks on humans by Africanized (killer) bees.
<narr> Narrative: Relevant documents must cite a specific instance of a human attacked by killer bees. Documents that note migration patterns or report attacks on other animals are not relevant unless they also cite an attack on a human.
</top>
<top>
<num> Number: 431 <title> robotic technology
<desc> Description: What are the latest developments in robotic technology?
<narr> Narrative: A relevant document will contain information on current applications of robotic technology. Discussions of robotics research or simulations of robots are not relevant.
</top>
<top>
<num> Number: 432 <title> profiling, motorists, police
<desc> Description: Do police departments use "profiling" to stop motorists?
<narr> Narrative: A relevant document will report or discuss police department criteria for identifying motorists considered likely to be carrying contraband. Documents discussing the detention of individuals by foreign security forces are not relevant.
</top>
<top>
<num> Number: 433 <title> Greek, philosophy, stoicism
<desc> Description: Is there contemporary interest in the Greek philosophy of stoicism?
<narr> Narrative: Actual references to the philosophy or philosophers, productions of Greek stoic plays, and new "stoic" artistic productions are all relevant.
</top>
<top>

<num> Number: 434 <title> Estonia, economy
<desc> Description: What is the state of the economy of Estonia?
<narr> Narrative: Documents that give concrete economic information such as economic statistics, entering economic unions and treaties, or monetary performance are relevant, as are discussions of economic issues such as transportation or pollution.
</top>
<top>
<num> Number: 435 <title> curbing population growth
<desc> Description: What measures have been taken worldwide and what countries have been effective in curbing population growth?
<narr> Narrative: A relevant document must describe an actual case in which population measures have been taken and their results are known. The reduction measures must have been actively pursued; that is, passive events such as disease or famine involuntarily reducing the population are not relevant.
</top>
<top>
<num> Number: 436 <title> railway accidents
<desc> Description: What are the causes of railway accidents throughout the world?
<narr> Narrative: A relevant document provides data on railway accidents of any sort (i.e., locomotive, trolley, streetcar) where either the railroad system or the vehicle or pedestrian involved caused the accident. Documents that discuss railroading in general, new rail lines, new technology for safety, and safety and accident prevention are not relevant, unless an actual accident is described.
</top>
<top>
<num> Number: 437 <title> deregulation, gas, electric
<desc> Description: What has been the experience of residential utility customers following deregulation of gas and electric?
<narr> Narrative: Documents that discuss privatization of government- owned utilities alone are not relevant. Also, not relevant are documents that discuss the deregulation of utilities for commercial customers.
</top>
<top>
<num> Number: 438 <title> tourism, increase
<desc> Description: What countries are experiencing an increase in tourism?
<narr> Narrative: A relevant document will name a country that has experienced an increase in tourism. The increase must represent the nation as a whole and tourism in general, not be restricted to only certain regions of the country or to some specific type of tourism (e.g., adventure travel). Documents discussing only projected increases are not relevant.
</top>
<top>
<num> Number: 439 <title> inventions, scientific discoveries
<desc> Description: What new inventions or scientific discoveries have been made?
<narr> Narrative: The word "new" in the description is defined as occurring in the 1990s. Documents that indicate a "recent" invention or scientific discovery are considered relevant. Discoveries made in astronomy or any scientific discoveries that are not patentable are not relevant.
</top>
<top>
<num> Number: 440 <title> child labor
<desc> Description: What steps are being taken by governments or corporations to eliminate abuse of child labor?
<narr> Narrative: A relevant document identifies an action taken by either a private commercial corporation or governmental organization to reduce or eliminate the use of child labor in manufacturing operations.
</top>

<top>
<num> Number: 441 <title> Lyme disease
<desc> Description: How do you prevent and treat Lyme disease?
<narr> Narrative: Documents that discuss current prevention and treatment techniques for Lyme disease are relevant. Reports of research on new treatments of the disease are also relevant.
</top>
<top>
<num> Number: 442 <title> heroic acts
<desc> Description: Find accounts of selfless heroic acts by individuals or small groups for the benefit of others or a cause.
<narr> Narrative: Relevant documents will contain a description of specific acts. General statements concerning heroic acts are not relevant.
</top>
<top>
<num> Number: 443 <title> U.S., investment, Africa
<desc> Description: What is the extent of U.S. (government and private) investment in sub-Saharan Africa?
<narr> Narrative: All references to U.S. Governmental and private assistance to sub-Saharan Africa are relevant. Documents discussing contributions by reason of U.S. membership in international aid organizations are also relevant.
</top>
<top>
<num> Number: 444 <title> supercritical fluids
<desc> Description: What are the potential uses for supercritical fluids as an environmental protection measure?
<narr> Narrative: To be relevant, a document must indicate that the fluid involved is achieved by a process of pressurization producing the supercritical fluid.
</top>
<top>
<num> Number: 445 <title> women clergy
<desc> Description: What other countries besides the United States are considering or have approved women as clergy persons?
<narr> Narrative: To be relevant, a document must indicate either a country where a woman has been installed as clergy or a country that is considering such an installation. The clergy position must be as church pastor rather than some other church capacity (e.g., nun or choir member).
</top>
<top>
<num> Number: 446 <title> tourists, violence
<desc> Description: Where are tourists likely to be subjected to acts of violence causing bodily harm or death?
<narr> Narrative: A relevant document must contain accounts of known harm to tourists. Evidence of single, isolated incidents are not relevant.
</top>
<top>
<num> Number: 447 <title> Stirling engine
<desc> Description: What new developments and applications are there for the Stirling engine?
<narr> Narrative: Any discussion of new developments and applications of the Stirling engine (also known as the Stirling cycle) are relevant.
</top>
<top>
<num> Number: 448 <title> ship losses
<desc> Description: Identify instances in which weather was a main or contributing factor in the loss of a ship at sea.
<narr> Narrative: Any ship loss due to weather is relevant, either in international or coastal waters.

</top>
<top>
<num> Number: 449 <title> antibiotics ineffectiveness
<desc> Description: What has caused the current ineffectiveness of antibiotics against infections and what is the prognosis for new drugs?
<narr> Narrative: To be relevant, a document must discuss the reasons or causes for the ineffectiveness of current antibiotics. Relevant documents may also include efforts by pharmaceutical companies and federal government agencies to find new cures, updating current testing phases, new drugs being tested, and the prognosis for the availability of new and effective antibiotics.
</top>
<top>
<num> Number: 450 <title> King Hussein, peace
<desc> Description: How significant a figure over the years was the late Jordanian King Hussein in furthering peace in the Middle East?
<narr> Narrative: A relevant document must include mention of Israel; King Hussein himself as opposed to other Jordanian officials; discussion of the King's on-going, previous or upcoming efforts; and efforts pertinent to the peace process, not merely Jordan's relationship with other middle-east countries or the U.S.
</top>

B.5 Appendix: Topics in WT10G

<top>
<num> Number: 451 <title> What is a Bengals cat?
<desc> Description: Provide information on the Bengal cat breed.
<narr> Narrative: Item should include any information on the Bengal cat breed, including description, origin, characteristics, breeding program, names of breeders and catteries carrying bengals. References which discuss bengal clubs only are not relevant. Discussions of bengal tigers are not relevant.
</top>
<top>
<num> Number: 452 <title> do beavers live in salt water
<desc> Description: Describe the normal habitat for beavers; note exceptions, if any.
<narr> Narrative: Relevant documents describe the habitat range as well as references to specific areas and bodies of water.
</top>
<top>
<num> Number: 453 <title> hunger
<desc> Description: Find documents that discuss organizations/groups that are aiding in the eradication of the worldwide hunger problem.
<narr> Narrative: Relevant documents contain the name of any organization or group that is attempting to relieve the hunger problem in the world. Documents that address the problem only without providing names of organizations/groups that are working on hunger are irrelevant.
</top>
<top>
<num> Number: 454 <title> parkinson's disease
<desc> Description: What are the symptoms and treatment of Parkinson's Disease, and what segments of the population have this disease?
<narr> Narrative: Documents discussing research projects and funding for research projects were considered relevant only when clinical trials were included. Documents regarding legislation which discussed funding and programs were considered irrelevant.
</top>

<top>
<num> Number: 455 <title> whan did Jackie Robinson appear at his first game
<desc> Description: Find documents that indicate when Jackie Robinson made his major league debut.
<narr> Narrative: A relevant document must contain the year that Jackie Robinson broke into major league baseball.
</top>
<top>
<num> Number: 456 <title> is the world going to end 2000
<desc> Description: Identify individuals or groups predicting the end of the world in the year 2000.
<narr> Narrative: References to the apocalypse are taken as equivalent to "end of the world" and are therefore relevant. Documents that give imprecise references to "those who believe...", for example, are not relevant.
</top>
<top>
<num> Number: 457 <title> CHEVROLET TRUCKS
<desc> Description: Find documents that address the types of Chevrolet trucks available.
<narr> Narrative: Relevant documents must contain information such as: the length, weight, cargo size, wheelbase, horsepower, cost, etc.
</top>
<top>
<num> Number: 458 <title> fasting
<desc> Description: Find documents that discuss fasting for religious reasons.
<narr> Narrative: A relevant document discusses fasting as related to periods of religious significance. Relevant documents should state the reason for fasting and the benefits to be derived.
</top>
<top>
<num> Number: 459 <title> when can a lender foreclose on property
<desc> Description: Identify documents which state the circumstances under which a lender can legally foreclose on a property.
<narr> Narrative: A relevant document will include the statutes under which a lender can foreclose on a property, the appeal process available to the property owner, the maximum amount the property can be sold for and any mitigating or legal options available to the property owner to stall the sale by the lender.
</top>
<top>
<num> Number: 460 <title> Who was Moses?
<desc> Description: Find documents that discuss the biblical figure of Moses.
<narr> Narrative: A relevant document includes any information concerning Moses and his deeds regarding the Israelites.
</top>
<top>
<num> Number: 461 <title> lava lamps
<desc> Description: Find documents that discuss the origin or operation of lava lamps.
<narr> Narrative: A relevant document must contain information on the origin or the operation of the lava lamp.
</top>
<top>
<num> Number: 462 <title> real estate and new jersey
<desc> Description: Find documents that contain residential real estate listings within New Jersey.
<narr> Narrative: Documents containing realtor data such as point of contact, address, web site or email address are considered as a real estate listing are are relevant. Listings of commercial real estate for sale or auction are not relevant.
</top>
<top>

<num> Number: 463 <title> tartin
<desc> Description: Find information on Scottish tartans: their history, current use, how they are made, and how to wear them.
<narr> Narrative: Simple listings of clan/tartan names or price lists are not relevant. Pictures or descriptions of individual plaids are not relevant unless accompanied by history of their development.
</top>
<top>
<num> Number: 464 <title> nativity scenes
<desc> Description: Identify U.S. cities or states that have banned displays of the nativity scene.
<narr> Narrative: Relevant documents will identify those U.S. cities or states which have banned the display of the nativity scene. Documents which reveal the reasoning behind the banning are also relevant as well as any effort to do away with the ban.
</top>
<top>
<num> Number: 465 <title> deer
<desc> Description: What kinds of diseases can infect humans due to contact with deer or consumption of deer meat?
<narr> Narrative: Documents explaining the transference of Lyme disease to humans from deer ticks are relevant.
</top>
<top>
<num> Number: 466 <title> information about the Peer Gynt Suite?
<desc> Description: Identify documents that provide background on the Peer Gynt Suite.
<narr> Narrative: Relevant documents will contain information on what inspired Grieg to compose the Peer Gynt Suite, folklore basis of the suite, or details of the composition. Notices of performances of the Peer Gynt Suite are not relevant. Reviews of performances are not relevant unless they analyze the work itself (as opposed to the performance).
</top>
<top>
<num> Number: 467 <title> dachshund dachshunds "wiener dog"
<desc> Description: Identify documents that contain information on buying and owning dachshund dogs.
<narr> Narrative: Documents that discuss general dog information which is directly applicable to buying and owning dachshunds (i.e., how to chose a breeder) are relevant. Documents that list names of dachshund breeders and names of clubs for dachshund owners are relevant.
</top>
<top>
<num> Number: 468 <title> incandescent light bulb
<desc> Description: Find documents that address the history of the incandescent light bulb.
<narr> Narrative: A relevant document must provide information on who worked on the development of the incandescent light bulb. Relevant documents should include locations and dates of the development efforts. Documents that discuss unsuccessful development attempts and non- commercial use of incandescent light bulbs are considered relevant.
</top>
<top>
<num> Number: 469 <title> steinbach nutcracker
<desc> Description: Find documents that discuss Steinbach nutcrackers.
<narr> Narrative: Relevant documents include the name and locations of Steinbach nutcracker manufacturers, the variety of nutcracker designs, the origin of this type of ornament, and where they can be purchased.
</top>
<top>
<num> Number: 470 <title> mistletoe

<desc> Description: Identify documents that discuss beneficial uses of mistletoe.
<narr> Narrative: A document is relevant if it mentions any beneficial uses of mistletoe for humans or wildlife. References to the use of mistletoe for decorations, festivals, and superstitious beliefs are not relevant.
</top>
<top>
<num> Number: 471 <title> mexican food culture
<desc> Description: Find documents that discuss the popularity or appeal of Mexican food outside of the United States.
<narr> Narrative: Documents that discuss the popularity of Mexican food in the United States, Central and South America are not relevant. Relevant documents discuss the extent to which Mexican food is enjoyed or used in Europe, Asia, Africa, or Australia.
</top>
<top>
<num> Number: 472 <title> antique appliance restoration
<desc> Description: Find documents that identify museums, collectors, or dealers that hold or sell restored antique electrical appliances.
<narr> Narrative: Relevant documents will identify museums which contain restored antique appliances and individuals who have collected or restored such appliances. References to dealers or businesses for antique appliance restoration are relevant.
</top>
<top>
<num> Number: 473 <title> Toronto FIlm Awards
<desc> Description: Find documents that discuss the Toronto Film Festival awards.
<narr> Narrative: Relevant documents must identify the name of the film which won the award at the Toronto Film Festival.
</top>
<top>
<num> Number: 474 <title> how e-mail bennefits businesses
<desc> Description: Find documents that discuss the profitability of businesses selling their products via the internet versus in-store sales.
<narr> Narrative: A relevant document must compare the benefits of selling products via the internet to sales made in the store.
</top>
<top>
<num> Number: 475 <title> what is the compostion of zirconium
<desc> Description: Find documents that describe the physical properties of zirconium.
<narr> Narrative: A document is relevant if it describes the element itself or its behavior under various physical conditions. A document is not relevant if it gives only the uses of zirconium or its compounds without stating why it combines with certain elements.
</top>
<top>
<num> Number: 476 <title> Jennifer Aniston
<desc> Description: Find documents that identify movies and/or television programs that Jennifer Aniston has appeared in.
<narr> Narrative: Relevant documents include movies and/or television programs that Jennifer Aniston has appeared in.
</top>
<top>
<num> Number: 477 <title> Royal Carribean Cruise Lines

<desc> Description: Find documents that indicate how many ships are operated by Royal Caribbean Cruise Lines (RCCL).
<narr> Narrative: Relevant documents should include only the ships that are presently sailing. Documents that discuss ships that are planned or under construction are not relevant.
</top>
<top>
<num> Number: 478 <title> baltimore
<desc> Description: Find documents that identify any individual who has served as the mayor of Baltimore.
<narr> Narrative: Relevant documents must identify the individual by name and must state that he is/was mayor of Baltimore.
</top>
<top>
<num> Number: 479 <title> where can I find information about kappa alpha psi?
<desc> Description: Find documents that explain if Kappa Alpha Psi is a fraternity or a sorority and which schools it is affiliated with.
<narr> Narrative: Documents that give background information or objectives of Kappa Alpha Psi are relevant.
</top>
<top>
<num> Number: 480 <title> car traffic report
<desc> Description: Find reports on automobile traffic in the Washington, DC metropolitan area.
<narr> Narrative: Any reference to automobile traffic conditions in the Washington, DC metropolitan area (including Maryland and Virginia suburbs) is relevant. Statements of existing problems are relevant. Transportation planning to improve traffic conditions is not relevant.
</top>
<top>
<num> Number: 481 <title> what did babe ruth do in the 1920's?
<desc> Description: Find documents that discuss Babe Ruth's baseball accomplishments in 1920.
<narr> Narrative: Any document containing information of any baseball activities relating to Babe Ruth in 1920 is relevant.
</top>
<top>
<num> Number: 482 <title> where can i find growth rates for the pine tree?
<desc> Description: Find documents that give growth rates of pine trees.
<narr> Narrative: Document that give heights of trees but not the rate of growth are not relevant.
</top>
<top>
<num> Number: 483 <title> rosebowl parade
<desc> Description: Find documents that identify the Rose Bowl Parade and what preparations are undertaken for it.
<narr> Narrative: Relative documents would contain any data on the Rose Bowl Parade including date, place, and preparation for the event. Documents containing information on Rose Bowl events not associated with the parade, such as the football game, are irrelevant.
</top>
<top>
<num> Number: 484 <title> auto skoda
<desc> Description: Skoda is a heavy industrial complex in Czechoslovakia. Does it manufacture vehicles?
<narr> Narrative: Relevant documents would include references to historic and contemporary automobile and truck production. Non-relevant documents would pertain to armament production.
</top>
<top>

<num> Number: 485 <title> gps clock
<desc> Description: Clock reliance is a very important consideration in the operation of a global positioning system (GPS). What entity is responsible for clock accuracy and what is the accuracy?
<narr> Narrative: A relevant document cites both the entity responsible for clock operation as well as indicating the achieved accuracy or accuracies.
</top>
<top>
<num> Number: 486 <title> where is the Eldorado Casino in Reno ?
<desc> Description: The Eldorado (El Dorado) Casino is reportedly located in Reno. Is this so and what is the address?
<narr> Narrative: A relevant document will provide the street address of an Eldorado or El Dorado Casino in Reno, Nevada.
</top>
<top>
<num> Number: 487 <title> angioplasty
<desc> Description: How often do patients who have undergone a successful angioplasty procedure require a repeat/s angioplasty?
<narr> Narrative: Any document which refers to the necessity of performing a follow-up angioplasty procedure on a patient who had had an earlier angioplasty would be relevant. The item could refer to a specific patient or could provide statistics on such occurrences.
</top>
<top>
<num> Number: 488 <title> newport beach california
<desc> Description: What forms of entertainment are available in Newport Beach, California?
<narr> Narrative: Any document which refers to entertainment in Newport Beach is relevant. This would include spectator and participation sports, shows, theaters, tourist attractions, etc.
</top>
<top>
<num> Number: 489 <title> calcium
<desc> Description: How do members of the medical profession view the effectiveness of calcium supplements?
<narr> Narrative: Any document which cites the benefits of humans using calcium supplements or advises how calcium supplements should be used are relevant. A relevant document must establish that the information comes from a qualified medical source and not from the claims of a manufacturer or vendor of calcium supplements or from the opinion of anyone not recognized by the medical profession.
</top>
<top>
<num> Number: 490 <title> motorcycle safety helmets
<desc> Description: Are there laws governing the use of motorcycle safety helmets?
<narr> Narrative: A relevant document will note any law governing the use of a motorcycle safety helmet.
</top>
<top>
<num> Number: 491 <title> Japanese Wave
<desc> Description: Identify occurrences in which a Japanese wave or tsunami caused loss of life or damage.
<narr> Narrative: Any reports that describe the occurrence of a Japanese wave or tsunami causing loss of life or damage are relevant. A relevant report must describe an actual event occurring at any location.
</top>
<top>
<num> Number: 492 <title> "us savings bonds"
<desc> Description: A document should address the information needed to make decisions on purchases or redemptions of U.S. Savings Bonds.

<narr> Narrative: A document must contain information on types of U.S. Savings Bonds, prices, interest rates, maturity periods and denominations. It need not contain all of the above information.
</top>
<top>
<num> Number: 493 <title> retire
<desc> Description: A document should list retirement communities in the U.S. and Canada.
<narr> Narrative: A document must list retirement communities, villages, centers and resorts available in the U.S. and Canada. It must provide such information as location and name.
</top>
<top>
<num> Number: 494 <title> nirvana
<desc> Description: Find information on members of the rock group Nirvana.
<narr> Narrative: Descriptions of members' behavior at various concerts and their performing style is relevant. Information on who wrote certain songs or a band member's role in producing a song is relevant. Biographical information on members is also relevant.
</top>
<top>
<num> Number: 495 <title> Where can I find information on the decade of the 1920's?
<desc> Description: Find information on the decade of the 1920's, known also as the Roaring Twenties.
<narr> Narrative: Information on life or happenings during the 1920's decade anywhere in the world is relevant. Simple dates of birth or death in the 1920's are not relevant unless they have broader significance.
</top>
<top>
<num> Number: 496 <title> TMJ
<desc> Description: Describe TMJ syndrome, its causes, symptoms and treatment.
<narr> Narrative: TMJ is the abbreviation for temporal mandible joint, but is often used to indicate TMJ syndrome. Documents that contain valid information on TMJ, even though in themselves are meant to be humorous or pejorative, are relevant. Documents that mention a name of a treatment such as "splint therapy" without making the treatment clear will be non-relevant.
</top>
<top>
<num> Number: 497 <title> orchids
<desc> Description: How big is the orchid growing industry in the U.S.? What are some of the major growers and how big of an industry is it in dollars?
<narr> Narrative: Documents that merely list U.S. orchid growers are relevant. Documents that are advertisement for individual growers are not relevant.
</top>
<top>
<num> Number: 498 <title> hair transplant
<desc> Description: How many hair transplant procedures take place in the U.S. per year? What is the cost of such a procedure? How successful are they, and what are the inherent risks?
<narr> Narrative: Documents revealing any risks or discomfort whatsoever are relevant. Documents that only discuss hair replacement procedures other than transplanting are not relevant.
</top>
<top>
<num> Number: 499 <title> pool cue
<desc> Description: This search is to find information on the origin, development, selection, or use of the pool cue.
<narr> Narrative: Documents that merely mention the pool cue are not relevant. Advertisements of pool cues are not relevant.
</top>

<top>
<num> Number: 500 <title> DNA Testing
<desc> Description: This search seeks information on the state of the art of DNA testing; what it is and what its goals are.
<narr> Narrative: Relevant documents may discuss those things which are essentially steps in the DNA testing procedure, such as: sequencing, analysis, fingerprinting, and profiling. Documents that provide descriptions of elaborate scientific DNA testing are not relevant.
</top>
<top>
<num> Number: 501 <title> deduction and induction in English?
<desc> Description: What is the difference between deduction and induction in the process of reasoning?
<narr> Narrative: A relevant document will contrast inductive and deductive reasoning. A document that discusses only one or the other is not relevant.
</top>
<top>
<num> Number: 502 <title> prime factor?
<desc> Description: What is a prime factor?
<narr> Narrative: A relevant document will define prime numbers or prime factors of mathematical expressions. Documents that use prime factors without defining them are not relevant.
</top>
<top>
<num> Number: 503 <title> Vikings in Scotland?
<desc> Description: What hard evidence proves that the Vikings visited or lived in Scotland?
<narr> Narrative: A document that merely states that the Vikings visited or lived in Scotland is not relevant. A relevant document must mention the source of the information, such as relics, sagas, runes or other records from those times.
</top>
<top>
<num> Number: 504 <title> information about what manatees eat
<desc> Description: Find documents that describe the diet of the manatee.
<narr> Narrative: Relevant documents will identify any foods providing sustenance to the manatees.
</top>
<top>
<num> Number: 505 <title> edmund hillary; sir?
<desc> Description: Who is/was Edmund Hillary?
<narr> Narrative: A relevant document will provide biographical information on Edmund Hillary.
</top>
<top>
<num> Number: 506 <title> history of skateboarding?
<desc> Description: What is the history of skateboarding?
<narr> Narrative: To be relevant, a document will report on when skateboarding first originated and how it came about. Stories about harm done by skateboarding and skateboarding dangers are not relevant. Current events about skateboarding are not relevant.
</top>
<top>
<num> Number: 507 <title> dodge recalls?
<desc> Description: Find documents that report on the recall of any Dodge automobile products.
<narr> Narrative: A recall notice must be specific to the Dodge brand to be relevant.
</top>
<top>

<num> Number: 508 <title> hair loss is a symptom of what diseases
<desc> Description: Find diseases for which hair loss is a symptom.
<narr> Narrative: A document is relevant if it positively connects the loss of head hair in humans with a specific disease. In this context, "thinning hair" and "hair loss" are synonymous. Loss of body and/or facial hair is irrelevant, as is hair loss caused by drug therapy.
</top>
<top>
<num> Number: 509 <title> steroids;what does it do to your body
<desc> Description: What effect do steroids have on the human body?
<narr> Narrative: A relevant document will describe any physiological effect of steroids on humans. Animal studies are not relevant.
</top>
<top>
<num> Number: 510 <title> do you have any information on j. robert oppenheimer?
<desc> Description: Find biographical data on J. Robert Oppenheimer.
<narr> Narrative: A relevant document will contain biographical information on J. Robert Oppenheimer.
</top>
<top>
<num> Number: 511 <title> diseases caused by smoking?
<desc> Description: What diseases does smoking cause?
<narr> Narrative: A relevant document must describe smoking tobacco products as a cause of a disease. Diseases caused by second-hand smoke and smokeless tobacco are not relevant.
</top>
<top>
<num> Number: 512 <title> how are tornadoes formed?
<desc> Description: How are tornadoes formed?
<narr> Narrative: A relevant document will provide the meteorological and atmospheric conditions necessary to create a tornado and explain how the conditions interact to form the funnel-shaped cloud.
</top>
<top>
<num> Number: 513 <title> earthquakes?
<desc> Description: What causes earthquakes, and where do they occur most often?
<narr> Narrative: A relevant document will discuss scientific causes of earthquakes or tremors and/or report geographic areas where earthquake activity occurs most frequently.
</top>
<top>
<num> Number: 514 <title> how much money for retirement?
<desc> Description: Find documents that discuss the financial aspects of retirement planning.
<narr> Narrative: A relevant document will describe how to calculate the amount of money needed during retirement and investment strategies to fund one's lifestyle after retiring from regular employment.
</top>
<top>
<num> Number: 515 <title> what about alexander graham bell
<desc> Description: What did Alexander Graham Bell invent?
<narr> Narrative: A relevant document will list an invention by Alexander Graham Bell. The document must provide the date or approximate date of the invention or patent to be relevant.
</top>
<top>
<num> Number: 516 <title> halloween?
<desc> Description: When, where, and how did Halloween evolve?

<narr> Narrative: A relevant document will discuss the origin of Halloween and the original customs of Halloween. Modern day trick-or-treating stories are not relevant.
</top>
<top>
<num> Number: 517 <title> titanic what went wrong
<desc> Description: Find documents that discuss the reasons for or problems leading to the sinking of the Titanic.
<narr> Narrative: A relevant document will discuss what caused the Titanic to sink.
</top>
<top>
<num> Number: 518 <title> how we use statistics to aid our decision making?
<desc> Description: Find documents that reference the use of statistical data in decision-making.
<narr> Narrative: A relevant document will describe a specific statistical method that is used to assist decision-making.
</top>
<top>
<num> Number: 519 <title> info on where frogs live
<desc> Description: Find documents that describe the habitat of frogs.
<narr> Narrative: A relevant document will identify the natural habitat of any type of frog. A frog's diet is not relevant.
</top>
<top>
<num> Number: 520 <title> how was the black plague stopped?
<desc> Description: What measures were taken to end the black plague?
<narr> Narrative: A relevant document will identify measures taken to end the black plague (also known as the black death or bubonic plague) of the 14th century.
</top>
<top>
<num> Number: 521 <title> surveys on the best places to live
<desc> Description: Identify the best places to live as judged by various surveys and magazines.
<narr> Narrative: A document that identifies the best places to live for some segment of the population as judged by a third party is relevant. Tourism promotions and areas that proclaim themselves to be best are not relevant.
</top>
<top>
<num> Number: 522 <title> how is water supplied to the mojave desert region?
<desc> Description: Identify methods being used to provide water to the Mojave Desert region
<narr> Narrative: A relevant document will describe ways in which water is supplied to the region.
</top>
<top>
<num> Number: 523 <title> facts about the five main clouds?
<desc> Description: How are the five main types of clouds formed?
<narr> Narrative: A document that explains the process of cloud formation for any of the five main types of clouds is relevant. A document that discusses clouds, but does not explain their formation processes is not relevant.
</top>
<top>
<num> Number: 524 <title> how to erase scar?
<desc> Description: What methods are used for removal of scar tissue?
<narr> Narrative: A relevant document must disclose the name of a procedure or describe it, or identify the instrument used to remove scar tissue or skin defects. Mere references to "surgical removal" are insufficient.
</top>
<top>

<num> Number: 525 <title> how does water get into the atmosphere?
<desc> Description: How does water enter the earth's atmosphere?
<narr> Narrative: A relevant document will explain the meteorological process whereby the earth's atmosphere becomes impregnated with water.
</top>
<top>
<num> Number: 526 <title> bmi
<desc> Description: What does BMI stand for?
<narr> Narrative: Any document that gives defines or explains BMI is relevant.
</top>
<top>
<num> Number: 527 <title> can you info on booker t. washington?
<desc> Description: What biographical data is available on Booker T. Washington?
<narr> Narrative: A relevant document will provide biographical information on Booker T. Washington.
</top>
<top>
<num> Number: 528 <title> How does a hygrometer measure the humidity in the atmosphere?
<desc> Description: How does a hygrometer measure the humidity in the atmosphere?
<narr> Narrative: A relevant document will describe a hygrometer in enough detail to determine the principle of operation.
</top>
<top>
<num> Number: 529 <title> history on cambodia?
<desc> Description: Find accounts of the history of Cambodia.
<narr> Narrative: A relevant document will provide historical information on Cambodia. Current events in Cambodia are not relevant.
</top>
<top>
<num> Number: 530 <title> do pheromone scents work?
<desc> Description: What is the scientific evidence that suggests pheromones stimulate the opposite sex?
<narr> Narrative: A relevant document will discuss how pheromones act as an attractor or repellent among humans, other animals, or plants.
</top>
<top>
<num> Number: 531 <title> who and whom
<desc> Description: What is the proper grammatical use of "who" versus "whom"?.
<narr> Narrative: A relevant document will provide explicit guidance for the proper grammatical use of "who" and "whom".
</top>
<top>
<num> Number: 532 <title> hypnosis?
<desc> Description: Find documents that describe the uses of hypnosis, and report on how effective hypnosis is.
<narr> Narrative: A relevant document will describe a medical/social use of hypnosis and its effectiveness in addressing the problem. The use of hypnosis as entertainment is not relevant.
</top>
<top>
<num> Number: 533 <title> school uniforms in public schools?
<desc> Description: Find documents either pro or con regarding children wearing uniforms in public schools.

<narr> Narrative: A relevant document will contain an argument for or against requiring students to wear uniforms in public schools. Advertisements for uniforms and simple statements that particular school districts require uniforms are not relevant.
</top>
<top>
<num> Number: 534 <title> artists who died in the 1700's
<desc> Description: Find items referencing artists who died in the 1700's.
<narr> Narrative: A relevant document must include the date of the artist's death.
</top>
<top>
<num> Number: 535 <title> canadian building codes?
<desc> Description: Find documents describing Canadian building codes.
<narr> Narrative: Regulations pertaining to any type of construction, including road building, are relevant.
</top>
<top>
<num> Number: 536 <title> can babies eat honey?
<desc> Description: Would eating honey affect the health of a baby?
<narr> Narrative: A relevant document will provide evidence as to whether it is safe to feed a human baby honey.
</top>
<top>
<num> Number: 537 <title> are sun beds safe
<desc> Description: Are sun beds detrimental to the health of an individual?
<narr> Narrative: A relevant document will identify negative health effects of sun (tanning) beds.
</top>
<top>
<num> Number: 538 <title> fha
<desc> Description: Find documents describing the Federal Housing Administration (FHA): when and why it was originally established and its current mission.
<narr> Narrative: A relevant document will discuss the history and current purpose of the Federal Housing Administration (FHA).
</top>
<top>
<num> Number: 539 <title> authors who suffered from depression
<desc> Description: Which authors suffered from depression?
<narr> Narrative: A relevant document will name authors who were depressed.
</top>
<top>
<num> Number: 540 <title> does stress cause obesity?
<desc> Description: Find descriptions of studies that correlate stress and obesity.
<narr> Narrative: A relevant document will discuss the relationship between stress and obesity.
</top>
<top>
<num> Number: 541 <title> instruments to forecast the weather?
<desc> Description: What instruments are used to forecast the weather?
<narr> Narrative: A relevant document will state that a particular instrument is used to forecast weather. Weather forecasts themselves are not relevant.
</top>
<top>
<num> Number: 542 <title> good things fire do for environment
<desc> Description: What are the positive effects of forest fires?

<narr> Narrative: A relevant document should describe specific ways the environment is helped as a result of forest fires.
</top>
<top>
<num> Number: 543 <title> radiography what are the risks
<desc> Description: What are the hazards associated with radiography?
<narr> Narrative: A relevant document should describe hazards/risks to either the patient or operator during the x-ray procedure.
</top>
<top>
<num> Number: 544 <title> estrogen why needed
<desc> Description: Find documents that describe the roles estrogen plays in the human body.
<narr> Narrative: Relevant documents will describe the positive effects of estrogen's presence or the negative effects of its absence. Discussions of the benefits of hormones in general are not relevant.
</top>
<top>
<num> Number: 545 <title> to what extent did peter the great reform russia
<desc> Description: What did Peter the Great accomplish for Russia?
<narr> Narrative: A relevant document will discuss how Peter the Great improved Russia. Documents that discuss the growth of Russian power and influence during his reign without specifying his particular contribution to the increase are not relevant.
</top>
<top>
<num> Number: 546 <title> recycle cans and why?
<desc> Description: What are the benefits of recycling cans?
<narr> Narrative: A relevant document will describe an economic or environmental benefit of recycling cans. Documents that describe the advantages of recycling in general terms are also relevant provided such advantages apply to can recycling.
</top>
<top>
<num> Number: 547 <title> camels why they were domesticated
<desc> Description: Why were camels domesticated?
<narr> Narrative: A relevant document will list a specific use of a camel that requires it to be domesticated.
</top>
<top>
<num> Number: 548 <title> how do you use solar heat to heat a pool?
<desc> Description: What are the methods of using solar heat to warm up the water in a swimming pool?
<narr> Narrative: A relevant document will explain a technique or method for warming the water in a swimming pool using heat from the sun. General discussions of solar heating are not relevant; the document must describe its application to swimming pools.
</top>
<top>
<num> Number: 549 <title> how is cancer related to cell reproduction?
<desc> Description: Find documents that contrast cancer with normal cell reproduction.
<narr> Narrative: A relevant document will discuss what causes cell reproduction to become cancerous or describe ways the body defends against run-away cell reproduction.
</top>
<top>
<num> Number: 550 <title> how are the volcanoes made?
<desc> Description: Find documents that describe how volcanoes are formed.

<narr> Narrative: A relevant document will explain how volcanoes get started, or will describe how plumes or hot spots cause eruptions. Reports that a particular volcano erupted are not relevant.
</top>

B.6 Appendix: Topics in .GOV2

<top> <num> Number: 701
<title> U.S. oil industry history
<desc> Description: Describe the history of the U.S. oil industry
<narr> Narrative: Relevant documents will include those on historical exploration and drilling as well as history of regulatory bodies. Relevant are history of the oil industry in various states, even if drilling began in 1950 or later.
</top>
<top> <num> Number: 702
<title> Pearl farming
<desc> Description: Pearl farming operations: actual farming operations described, culturing pearls, "Japanese pearl productions," status of pearl farming, production.
<narr> Narrative: Definitions of difference between natural, cultured, and imitation pearls is relevant. Description of how pearls are formed and discussion of mussel production for use in pearl preselection is relevant. "Pearl industry" with no further detail is NOT relevant. Rules for jewelers is relevant. Statistics on pearl production are relevant.
</top>
<top> <num> Number: 703
<title> U.S. against International Criminal Court
<desc> Description: What are the arguments the U.S. uses against joining the International Criminal Court?
<narr> Narrative: The title of an article alone is not sufficient to make a document relevant, nor are lists of articles relevant without descriptions of their content.
</top>
<top> <num> Number: 704
<title> Green party political views
<desc> Description: What are the goals and political views of the Green Party.
<narr> Narrative: Evidence that the Green Party is a recognized U.S. political party is relevant. Documents which reveal their goals, values, or political views are relevant. Attempts by green party members to stop or interrupt U.S. government efforts regarding certain environmental actions which they disagree with is not relevant. Any members names noted are considered relevant.
</top>
<top> <num> Number: 705
<title> Iraq foreign debt reduction
<desc> Description: Identify any efforts, proposed or undertaken, by world governments to seek reduction of Iraq's foreign debt.
<narr> Narrative: Documents noting this subject as a topic for discussion (e.g. at U.N. and G7) are relevant. Money pledged for reconstruction is irrelevant.
</top>
<top> <num> Number: 706
<title> Controlling type II diabetes
<desc> Description: What are methods used to control type II diabetes?
<narr> Narrative: Items containing such controls as determining blood sugar levels and keeping triglycerides, cholesterol and blood pressure in normal ranges are relevant. Mention of mild to moderate weight loss, regular exercise and learning new behaviors and attitudes, medications is relevant.
</top>
<top> <num> Number: 707
<title> Aspirin cancer prevention

<desc> Description: What evidence is there that aspirin may help prevent cancer?
<narr> Narrative: Relevant documents will state the type of cancer and the organizations which made the findings that aspirin was useful in prevention. Documents which appear to only link to the desired information, rather than having this information in the document, are not relevant.
</top>
<top> <num> Number: 708
<title> Decorative slate sources
<desc> Description: What are sources of slate stone for decorative use?
<narr> Narrative: Relevant documents will mention where slate can be obtained that is appropriate for decorative uses such as flooring, counter tops, or art. Quarries as well as distributors are relevant. "Slate belt" does not imply a source of slate.
</top>
<top> <num> Number: 709
<title> Horse racing jockey weight
<desc> Description: What are the limits and regulations concerning jockey weight in horse racing?
<narr> Narrative: Documents which discuss rules for horse racing relating to the weight of the jockey are relevant. Documents which give a minimum jockey weight are relevant. List of titles of sections with no text are NOT relevant.
</top>
<top> <num> Number: 710
<title> Prostate cancer treatments
<desc> Description: What are the various treatments for prostate cancer?
<narr> Narrative: Relevant cancer treatments include radiation therapy, radioactive pellets, hormonal therapy and surgery. "Watchful waiting" is also considered relevant.
</top>
<top> <num> Number: 711
<title> Train station security measures
<desc> Description: What security measures have been employed at train stations due to heightened security concerns?
<narr> Narrative: Use of national guard forces is considered relevant. Surveillance cameras, more police officers, K-9 units, and better ID checks are considered relevant.
</top>
<top> <num> Number: 712
<title> Pyramid scheme
<desc> Description: What are some actual examples of pyramid schemes?
<narr> Narrative: A relevant document must describe the actual pyramid scheme, not just the place it occurred or that it was a pyramid scheme. Theoretical description of a pyramid scheme is not relevant. Ponzi schemes are relevant (they are a type of pyramid).
</top>
<top> <num> Number: 713
<title> Chesapeake Bay Maryland clean
<desc> Description: What is the state of Maryland doing to clean up the Chesapeake Bay?
<narr> Narrative: Relevant documents will describe what Maryland in particular is doing to clean up the Bay. Documents that focus on other states that only mention Maryland in a group of cooperating states are not relevant.
</top>
<top> <num> Number: 714
<title> License restrictions older drivers
<desc> Description: What restrictions are placed on older persons renewing their drivers' licenses in the U.S.?
<narr> Narrative: Not relevant are documents on restrictions for persons with medical conditions, not necessarily related to age. Also not relevant are lists of publications on drivers licenses for older persons which give no restriction information.

</top>
<top> <num> Number: 715
<title> Schizophrenia drugs
<desc> Description: What organizations (private or governmental) are developing drugs to combat schizophrenia?
<narr> Narrative: Mention of pharmaceutical companies and university laboratories developing drugs is relevant. Also relevant are entities (such as NIMH) engaged in clinical trials of drugs which had been developed but were still under investigation.
</top>
<top> <num> Number: 716
<title> Spammer arrest sue
<desc> Description: Have any spammers been arrested or sued for sending unsolicited e-mail?
<narr> Narrative: Instances of arrests, prosecutions, convictions, and punishments of spammers, and lawsuits against them are relevant. Documents which describe laws to limit spam without giving details of lawsuits or criminal trials are not relevant.
</top>
<top> <num> Number: 717
<title> Gifted talented student programs
<desc> Description: What states or localities offer programs for gifted and talented students?
<narr> Narrative: Relevant is any document that shows a state or locality offers a program for gifted and talented children. Not relevant are documents on federal programs for gifted and talented children, nor information on organizations that promote programs for gifted and talented students.
</top>
<top> <num> Number: 718
<title> Controlling acid rain
<desc> Description: What methods are used to control acid rain and its effects?
<narr> Narrative: Documents pertaining to the decrease or eliminations of sulfur dioxide and nitrogen oxides into the air are relevant. Any reduction of the burning of fossil fuels is also relevant. Documents pertaining to the "Clean Air Act" are considered relevant.
</top>
<top> <num> Number: 719
<title> Cruise ship damage sea life
<desc> Description: What kinds of harm do cruise ships do to sea life such as coral reefs, and what is the extent of the damage?
<narr> Narrative: References to "large ships" or "shipping" are relevant unless a freighter or other type of non-cruise ship is specified or is otherwise apparent from name or picture.
</top>
<top> <num> Number: 720
<title> Federal welfare reform
<desc> Description: Find documents about Federal welfare reform legislation, regulation, and policy.
<narr> Narrative: Documents pertaining to welfare reform by a state are irrelevant. Resources such as pertinent legislation, regulations, policy and reports on the progress of Federal welfare reform are relevant.
</top>
<top> <num> Number: 721
<title> Census data applications
<desc> Description: What applications are there for U.S. decennial census data, and how is it used?
<narr> Narrative: Relevant documents must clearly describe how census data is used. Documents which only list types of available census data are not relevant.
</top>
<top> <num> Number: 722
<title> Iran terrorism

<desc> Description: In what ways does Iran support terrorism?
<narr> Narrative: Documents state only that Iran supports terrorism, without giving details, are not relevant. Relevant documents will identify terrorist groups sponsored by Iran, official links between Iran and terrorist attacks, or funding trails.
</top>
<top> <num> Number: 723
<title> Executive privilege
<desc> Description: What is the U.S. government's definition of "executive privilege?"
<narr> Narrative: Relevant documents will define the term "executive privilege", or describe an invocation of executive privilege which helps explain it. Application of executive privilege beyond the President to the executive branch is relevant.
</top>
<top> <num> Number: 724
<title> Iran Contra
<desc> Description: What was the Iran-Contra scandal and what were the consequences?
<narr> Narrative: Not relevant are documents that mention the Iran-Contra affair only peripherally in discussing some other subject.
</top>
<top> <num> Number: 725
<title> Low white blood cell count
<desc> Description: What would cause a lowered white blood cell count?
<narr> Narrative: A relevant document will describe a condition or disease that causes a lowered white blood cell count. Lowered white blood cell counts caused by HIV infection, bone marrow failure and chemotherapy are relevant. A low count caused by a treatment or medication would also be relevant.
</top>
<top> <num> Number: 726
<title> Hubble telescope repairs
<desc> Description: What repairs have been made on the Hubble telescope?
<narr> Narrative: Not relevant are documents such as lists of resources, inquiries or photos of or by Hubble that provide no information on repairs, unless the captions discuss or describe repairs.
</top>
<top> <num> Number: 727
<title> Church arson
<desc> Description: Identify any specific instances of church arson.
<narr> Narrative: Relevant documents must identify a specific arson attack and give the name of the church and/or its location. General references such as mention of histories or patterns of arsons at churches, or statistics from the church arson task force were considered irrelevant. Suspected arson and burning of synagogues are considered relevant.
</top>
<top> <num> Number: 728
<title> whales save endangered
<desc> Description: What's being done to save endangered whales?
<narr> Narrative: A relevant document will reveal any noted efforts to save whales. The classification of certain whales as an endangered species is an effort to save them, and therefore documents reflecting this are relevant. Proposals to limit whale hunting are not relevant if they are proposals not actual laws or regulations. Bans and moratoriums on whaling are relevant.
</top>
<top> <num> Number: 729
<title> Whistle blower department of defense
<desc> Description: What have been revelations of whistle blowers concerning the U.S. Department of Defense?

<narr> Narrative: Relevant documents will describe instances of kickbacks, fraud, or other illegal activity revealed by whistle blowers at the Department of Defense. Whistle blower protections and regulations are not relevant.
</top>
<top> <num> Number: 730
<title> Gastric bypass complications
<desc> Description: What are some of the possible complications and potential dangers of gastric bypass surgery?
<narr> Narrative: Mention of other bypass surgeries (e.g. heart, intestinal bypass, vertical banding, etc) are not considered relevant.
</top>
<top> <num> Number: 731
<title> Kurds history
<desc> Description: What is the history of the Kurds?
<narr> Narrative: Relevant are documents on long-time history of the Kurds, as well as documents on recent events, such as Saddam Hussein's 1988 poison gas attack on Kurdish residents of Halabja. Not relevant are documents on unrelated subjects that only briefly mention the Halabja attack.
</top>
<top> <num> Number: 732
<title> U.S. cheese production
<desc> Description: What cheese production is carried out in the U.S.?
<narr> Narrative: Relevant documents describe cheese production, including cheddar, swiss, gouda, and goat cheese. Products that include cheese are not relevant unless it is made clear that the cheese is made in the U.S. USDA regulations are not relevant unless they cite U.S.-produced cheese.
</top>
<top> <num> Number: 733
<title> Airline overbooking
<desc> Description: What are the regulations regarding airline overbooking?
<narr> Narrative: Relevant documents indicate regulations or industry rules which govern the practice of overbooking, selling more seats than are available on the plane. Schedules for reporting over-sales are relevant.
</top>
<top> <num> Number: 734
<title> Recycling successes
<desc> Description: What recycling projects have been successful?
<narr> Narrative: Guidelines by themselves are not relevant. Titles in a table of contents are relevant if they identify places or product programs which have had success. Must be declared successful or success should be clearly assumed from the description. Name of state identified as successful recycler is relevant. Listing of recycled products for sale are relevant.
</top>
<top> <num> Number: 735
<title> Afghan women condition
<desc> Description: Is the condition of Afghan women better under the new government than under the Taliban?
<narr> Narrative: Relevant documents must give information about Afghan women, post-Taliban, including whether they have more rights and freedom. Not relevant are documents from U.S. officials supporting women's rights or describing U.S. aid to Afghanistan aimed to help women.
</top>
<top> <num> Number: 736
<title> location BSE infections
<desc> Description: Where have animals infected with bovine spongiform encephalopathy (also known as BSE or Mad Cow disease) been found?

<narr> Narrative: Relevant documents should be specific, i.e. give a country or state instead of Europe or Asia. BSE should not be confused with TSE (Transmissible Spongiform Encephalopathy). Measures taken to limit or curb BSE infections are not relevant.
</top>
<top> <num> Number: 737
<title> Enron California energy crisis
<desc> Description: What allegations have been made about Enron's culpability in the California Energy crisis?
<narr> Narrative: Relevant documents will allege than Enron contributed to California's energy crisis. Mention of the failure of Enron to provide documents is not relevant.
</top>
<top> <num> Number: 738
<title> Anthrax hoaxes
<desc> Description: What are some examples of anthrax hoaxes?
<narr> Narrative: Only specific examples are relevant. Documents giving statistics such as "20 individuals arrested for hoaxes" are not relevant.
</top>
<top> <num> Number: 739
<title> Habitat for Humanity
<desc> Description: What is the organization "Habitat for Humanity", and what activities are they involved in?
<narr> Narrative: Documents that explain a specific Habitat for Humanity program are relevant. Locations of Habitat for Humanity chapters are relevant. Documents that reveal the purpose of Habitat for Humanity are relevant.
</top>
<top> <num> Number: 740
<title> regulate assisted living Maryland
<desc> Description: Who regulates assisted living facilities in Maryland?
<narr> Narrative: Relevant documents will indicate agencies which regulate assisted living facilities in Maryland. Documents which mention such an agency without indicating their regulatory responsibility are not relevant. Documents mentioning the agency and purpose, but not the state are not relevant.
</top>
<top> <num> Number: 741
<title> Artificial Intelligence
<desc> Description: What is artificial intelligence?
<narr> Narrative: Any documents pertaining to pattern recognition, problem-solving, natural language processing, and game playing, which are branches of artificial intelligence, are relevant.
</top>
<top> <num> Number: 742
<title> hedge funds fraud protection
<desc> Description: What protection do investors have against fraud by hedge funds?
<narr> Narrative: Relevant documents will give information on consumer protections against fraud by hedge funds. If different financial structures are involved, such as multiple kinds of funds or derivatives, the document is not relevant unless it makes clear that the fraud applies to hedge funds.
</top>
<top> <num> Number: 743
<title> Freighter ship registration
<desc> Description: What are the regulations and other considerations concerning registering a freighter in a country?
<narr> Narrative: Relevant documents will describe rules and regulations for registering a freighter under a particular flag or country. Anything relating to registration of freighters is relevant. Information about other types of ships is not relevant.

</top>
<top> <num> Number: 744
<title> Counterfeit ID punishments
<desc> Description: What punishments or sentences have been given in the U.S. for making or selling counterfeit IDs?
<narr> Narrative: Relevant documents will describe punishments for manufacturing or selling counterfeit identification, such as drivers licenses, passports, social security cards, etc. Fake professional certifications and fake credit cards are relevant. Counterfeit goods or auto serial numbers not relevant. Counterfeit checks are not relevant. "Counterfeiting" with no indication of type is relevant.
</top>
<top> <num> Number: 745
<title> Doomsday cults
<desc> Description: Identify any doomsday cult, their name, and location throughout the world.
<narr> Narrative: To be relevant a document must reflect that the cult is a doomsday or apocalyptic cult. Any document which indicates the location (by country) and name of the doomsday cult is relevant. References to cults other than doomsday cults is not relevant.
</top>
<top> <num> Number: 746
<title> Outsource job India
<desc> Description: What jobs have been outsourced to India?
<narr> Narrative: Relevant documents will identify jobs which used to be in the U.S. but have been outsourced to India. When the word "outsourced" is used not used but implied it is still relevant. Outsourcing to other countries besides India is not relevant.
</top>
<top> <num> Number: 747
<title> Library computer oversight
<desc> Description: What control or oversight is there over computer use in public libraries?
<narr> Narrative: Relevant documents will describe specific controls or oversight mechanisms that currently are in use. Bills that propose controls but have not been passed are not relevant. Statements regarding a lack of control/monitoring are not relevant. The fact that a Librarian must load CDs for patrons is relevant. Fees for computer print outs are not relevant. Statement of illegal activities that are prohibited is not relevant. Restrictions to legal research is relevant.
</top>
<top> <num> Number: 748
<title> Nuclear reactor types
<desc> Description: Name the types of nuclear reactor power plants in operation in the United States.
<narr> Narrative: Relevant document will contain information which identifies the type of commercial nuclear reactors in use in the United States. Documents indicating the difference in how these reactors work is relevant. General information on nuclear reactors is not relevant. Future changes in nuclear reactors is not relevant. Total numbers of each type of reactor is of relevance.
</top>
<top> <num> Number: 749
<title> Puerto Rico state
<desc> Description: Do people in Puerto Rico want for it to become a U.S. State?
<narr> Narrative: Relevant documents will show that the Puerto Rican people want or do not want United States statehood. Relevant documents can contain information indicating other choices for Puerto Ricans such as to remain a commonwealth of the U.S. or to become an independent nation. All other documents containing Puerto Rico commercial, geographical, economic, etc. information is not relevant.
</top>
<top> <num> Number: 750

<title> John Edwards womens issues
<desc> Description: What are Senator John Edwards' positions on women's issues such as pay equity, abortion, Title IX and violence against women.
<narr> Narrative: Relevant documents will indicate Senator John Edwards' stand on issues concerning women, such as pay parity, abortion rights, Title IX, and violence against women. Lists of press releases are relevant when the headlines show he is voting for or against bills on women's issues. Not relevant are Edwards' positions on issues not exclusively concerning women.
</top>
<top> <num> Number: 751
<title> Scrabble Players
<desc> Description: Give information on Scrabble players, when and where Scrabble is played, and how popular it has been.
<narr> Narrative: Give information on the social aspects of the game Scrabble. Scrabble players may be named or described as a group. Both real and fictional players are relevant. Mention of a scheduled Scrabble game is relevant. Scrabble's popularity is relevant. An account of a particular game is relevant. Descriptions of variants on the Scrabble game are not relevant. Use of Scrabble tiles for other purposes are not relevant. Scrabble software is not relevant unless there is mention of its users. Titles of Scrabble-related books (dictionaries, glossaries, rulebooks) are not relevant.
</top>
<top> <num> Number: 752
<title> Dam removal
<desc> Description: Where have dams been removed and what has been the environmental impact?
<narr> Narrative: Discussions of dam removal in general are relevant. Identification of specific places where dams have been removed is relevant, as is discussion of the environmental impact of removal. Applications for removal are not relevant unless reasons are given.
</top>
<top> <num> Number: 753
<title> bullying prevention programs
<desc> Description: What programs have been used in schools to prevent bullying of students?
<narr> Narrative: Relevant documents must have details about actual programs developed for and implemented in a school or schools with the goal of preventing bullying. This would include details such as classes, assemblies, discipline, mediation or projects for students and staff. Advice on how to develop such a program is not relevant. Listings or titles of programs without details as to how they are implemented are not relevant.
</top>
<top> <num> Number: 754
<title> domestic adoption laws
<desc> Description: Provide any legal information about domestic human adoption.
<narr> Narrative: Relevant documents must describe any laws covering adoption of minor children within the United States, to include laws of any specific states. Description of a law without specific name or identification of that law is acceptable. Reference to a law and its impact are acceptable. A reference to a law without any textual description of that law is not acceptable .(e.g. "Joe Bloggs was responsible for passage of adoption law Nr#####.") Documents containing only tax-related legalities, such as tax allowances, are not acceptable.
</top>
<top> <num> Number: 755
<title> Scottish Highland Games
<desc> Description: What is the history and location of Scottish highland games in the United States.
<narr> Narrative: Description of Scottish and Highland Games performed in the United States, their history and their geographic location are relevant. Mere mention of the games, time and place without description is considered not relevant.
</top>

<top> <num> Number: 756
<title> Volcanic Activity
<desc> Description: Locations of volcanic activity which occurred within the present day boundaries of the U.S. and its territories.
<narr> Narrative: Relevant information would include when volcanic activity took place, even millions of years ago, or, on the contrary, if it is a possible future event.
</top>
<top> <num> Number: 757
<title> Murals
<desc> Description: Show examples of murals.
<narr> Narrative: A picture of a mural must be present. Murals from any time period, works in progress, and portions of murals are relevant. Rothko's mural paintings are relevant. Friezes (bands of wall decoration) are relevant.
</top>
<top> <num> Number: 758
<title> Embryonic stem cells
<desc> Description: What are embryonic stem cells, and what restrictions are placed on their use in research?
<narr> Narrative: Explanation of the nature of embryonic stem cells is relevant. Their usefulness in research is relevant. Sources for them and restrictions on them also are relevant.
</top>
<top> <num> Number: 759
<title> civil war battle reenactments
<desc> Description: When and where are Civil War battle reenactments held?
<narr> Narrative: Reenactments of actual Civil War battles which are scheduled to be or have previously been performed are relevant. Regularly scheduled, or annual events are relevant. General reenactments of a soldier's life, dress, encampments, and demonstrations not connected to a specific battle are not relevant.
</top>
<top> <num> Number: 760
<title> american muslim mosques schools
<desc> Description: Statistics regarding American Muslims, mosques, and schools.
<narr> Narrative: Relevant documents should provide some count or proportion of mosques, Muslim-affiliated schools, or population. With regard to population, specific age groupings, sexes, or other categorizations are acceptable. The statistics can be pertinent to a specific geographic area, such as Fulton County, the state of California, or the Northeast. There is no restriction as to time period (for example 2005 versus 1987).
</top>
<top> <num> Number: 761
<title> Problems of Hmong Immigrants
<desc> Description: Describe the problems faced by Hmong immigrants to the United States
<narr> Narrative: Problems faced by Hmong immigrants and programs or legislation to assist with these problems are relevant. Legislation to grant citizenship to Hmong with an inadequate knowledge of English is judged relevant. Listings of materials and services available in the Hmong language are not relevant.
</top>
<top> <num> Number: 762
<title> History of Physicians in America
<desc> Description: Who have been considered "doctors" since the first European settlement in America?
<narr> Narrative: The history of physicians in America is relevant. Native American practitioners would not be relevant. Women who knew herbs and practiced midwifery would be relevant. Mention of "doctor" would be relevant if he/she is considered to be a doctor by other people.
</top>
<top> <num> Number: 763

<title> Hunting deaths
<desc> Description: Give information on human deaths associated with hunting for game.
<narr> Narrative: Accidental deaths, murders, and suicides are relevant. Deaths can be from any cause. Fatalities of people not in the hunting party are relevant, but the deaths must be connected with hunting. Relevant hunting must be for live prey. Deaths related to submarine hunting are not relevant.
</top>
<top> <num> Number: 764
<title> Increase mass transit use
<desc> Description: What is being done to increase mass transit use?
<narr> Narrative: Relevant documents should describe measures taken in municipalities to increase mass transit use. Actions taken by city officials or transit managers to increase rider ship of mass transit are relevant. Factors preventing use of mass transit are relevant. General statements of the value of mass transit are not relevant.
</top>
<top> <num> Number: 765
<title> ephedra ma huang deaths
<desc> Description: How many deaths have been attributed to the drug ephedra, also known as the herbal ingredient ma huang?
<narr> Narrative: Information of specific individuals' deaths or numbers of deaths actually attributed to the use of ephedra or ma huang is relevant. Mixed statistics which conflate injuries and deaths are not relevant. General statements referring to death as a side effect of the drug's use are not relevant.
</top>
<top> <num> Number: 766
<title> diamond smuggling
<desc> Description: Illicit activity involving diamonds, to include diamond smuggling.
<narr> Narrative: Relevant documents should include information about any illicit diamond activity. Smuggling of diamonds may be stated literally but other activities where illegal export (smuggling) is implied are acceptable. Such terms as "conflict diamonds" and "blood diamonds" are pertinent. The diamond terminology without text (such as a bibliography or document-less table of contents) is not acceptable. Any geographic area or relatively current (post 1950) timeframe is relevant.
</top>
<top> <num> Number: 767
<title> Pharmacist License requirements
<desc> Description: What are the requirements for a pharmacist's license in the U.S.?
<narr> Narrative: Laws governing time issuance or renewal of a pharmacist's license in the various states are relevant. Requirements to qualify for a license and to perform under the license are relevant. Listings of pharmacies, pharmacists or agencies regulating or concerned with pharmacy are not relevant.
</top>
<top> <num> Number: 768
<title> Women in state legislatures
<desc> Description: What is the number of women legislators or what percentage of the total legislators in any given state are women?
<narr> Narrative: A relevant document gives the exact number or the percentage of the total number of legislators that are women in any given state of the 50 states or gives a total for the entire United States. A document talking about only one woman legislator is not relevant.
</top>
<top> <num> Number: 769
<title> Kroll Associates Employees
<desc> Description: Identify employees of Kroll Associates.

<narr> Narrative: Employees must be named. Past and present Kroll employees are relevant. Kroll Associates is an international investigative and security firm whose regional branches are known as Kroll Associates Inc. or Kroll Associates Ltd. CK Kroll & Associates is an unrelated company and not relevant.
</top>
<top> <num> Number: 770
<title> Kyrgyzstan-United States relations
<desc> Description: What is the state of Kyrgyzstan-United States relations?
<narr> Narrative: Relevant documents should discuss interactions between Kyrgyzstan and the United States. Teaching guides for U.S. schools and other documents on Kyrgyzstan are not relevant unless they include information on relations between the two nations.
</top>
<top> <num> Number: 771
<title> deformed leopard frogs
<desc> Description: What deformities have been found in leopard frogs?
<narr> Narrative: Relevant documents must contain information pertaining to deformities specifically in leopard frogs. Details such as where these deformed frogs have been found and the types and/or causes of their deformities are relevant.
</top>
<top> <num> Number: 772
<title> flag display rules
<desc> Description: What are the rules or guidelines for display of the United States flag?
<narr> Narrative: Relevant documents must provide any regulations or guidelines for display of the U.S. flag. Not acceptable are bibliographies, tables of contents, or links to websites containing such information. Documents limited to flag abuse or desecration are not acceptable.
</top>
<top> <num> Number: 773
<title> Pennsylvania slot machine gambling
<desc> Description: What is the legal status of slot machine gambling in Pennsylvania?
<narr> Narrative: Legislation dealing with slot machine gambling in Pennsylvania is relevant as are count reports of cases prosecuted under existing law. Proposals to change existing laws in Pennsylvania relating to slot machine gambling outside of Pennsylvania are not relevant.
</top>
<top> <num> Number: 774
<title> Causes of Homelessness
<desc> Description: What are some of the causes of homelessness?
<narr> Narrative: Relevant causes do not have to be the actions of the homeless person/s themselves. Causes can come from outside the homeless person/s, such as being a veteran, or lack of available, lost-cost housing. Relevant causes can also be the actions of the homeless person/s such as alcoholism or drug addictions. Other relevant reasons are a persons desire to be "free" or mental illness.
</top>
<top> <num> Number: 775
<title> Commercial candy makers
<desc> Description: Identify commercial candy makers and give information concerning them.
<narr> Narrative: A candy manufacturer or brand must be named. Information on the company and its products, methods, markets, history, problems, successes, and popularity are relevant. Lawsuits and regulatory actions are relevant if a candy company or brand is named in the discussion.
</top>
<top> <num> Number: 776
<title> Magnet schools success
<desc> Description: Are magnet schools considered successful in districts where they have been created?

<narr> Narrative: Documents which describe the way magnet schools operate in general or in specific locations are relevant if they relate whether or not they are thought to be successful. Success can be judged in terms of reduced racial segregation or improved academic achievement or both. Documents which only state plans or aims are not relevant.
</top>
<top> <num> Number: 777
<title> hybrid alternative fuel cars
<desc> Description: What hybrid or alternative fuel passenger cars are auto manufacturers now marketing or developing for future sales?
<narr> Narrative: Relevant documents must contain some detail as to the manufacturer and model or type of car. Lists of manufactures and car models are relevant. SUV's and mini vans for private, not commercial, use are relevant. Commercial vehicles, such as buses or trucks are not relevant.
</top>
<top> <num> Number: 778
<title> golden ratio
<desc> Description: Golden ratio formula, description, or examples.
<narr> Narrative: Documents must contain the Golden Ratio formula or a description of its development. References to works of art, design, or architecture are acceptable. Examples of the Golden Ratio appearing in nature are acceptable. Other terms, such as "Golden Mean" or "Golden Proportion", are acceptable and can be substituted for Golden Ratio. References to Fibonacci without specifying the Golden Ratio or its formula are not acceptable. Documents limited to web links, bibliographies, or tables of contents (without accompanying documents) are not acceptable.
</top>
<top> <num> Number: 779
<title> Javelinas range and description
<desc> Description: Describe the Javelina or collared peccary and its geographic range.
<narr> Narrative: Physical description, habits, habitat, and range are all relevant. Photographs identified as "javelina" or "peccary" are relevant. Biographical listings or listings of 200 exhibits including the species' name are not considered relevant.
</top>
<top> <num> Number: 780
<title> Arable land
<desc> Description: How much of planet Earth is arable at present? Area must have plenty of water, sun and soil to support plant life.
<narr> Narrative: The mention of the percentage of land in any given state or country that will support plant life is relevant. The land does not have to be presently used for crops to be relevant.
</top>
<top> <num> Number: 781
<title> Squirrel control and protections
<desc> Description: Give information on steps to manage, control, or protect squirrels.
<narr> Narrative: Real or abstract measures are relevant. Petitions and decisions concerning squirrels are relevant. Dangers to squirrels and threats to their survival are relevant. Mentions of squirrel hunting are relevant. Measures to protect from damage by squirrels are relevant. Simple descriptions of squirrels are not relevant.
</top>
<top> <num> Number: 782
<title> Orange varieties seasons
<desc> Description: What are the varieties of oranges and when is each in season?
<narr> Narrative: Descriptions of orange varieties and the seasons in which they ripen are relevant. Lists are relevant if they name varieties of oranges. General information about oranges is not relevant.
</top>

<top> <num> Number: 783
<title> school mercury poisoning
<desc> Description: How have mercury poisonings of children occurred in schools and what measures are being taken to prevent such incidents?
<narr> Narrative: Relevant documents must have information on the ways mercury poisonings have actually occurred in schools and/or concrete measures that have been taken to prevent or reduce mercury exposure in schools. Relevant preventative measures would include specific regulations or policies enacted or the establishment of mercury-free zones. General information of how a poisoning might occur or advice on precautions in handling mercury is not relevant.
</top>
<top> <num> Number: 784
<title> mersenne primes
<desc> Description: Give a definition or description of Mersenne prime numbers.
<narr> Narrative: The document should have the formula for Mersenne primes. However, the exact formula may be omitted if there is a narrative description sufficiently clear for the reader to render it into the desired equation. A discussion of Mersenne numbers without reference to Mersenne primes is not acceptable! If the equation appears to be incorrect it will be acceptable anyway.
</top>
<top> <num> Number: 785
<title> Ivory-billed woodpecker
<desc> Description: What is the history and present status of the Ivory-billed Woodpecker?
<narr> Narrative: References to places where the Ivory-billed woodpecker was previously reported are relevant whether or not its present status has been determined. Statements that its status is "extirpated" or "extinct" are relevant descriptions of programs to locate, preserve, encourage or reintroduce this species are relevant. Reports of sightings and last sightings are relevant.
</top>
<top> <num> Number: 786
<title> Yew trees
<desc> Description: Where do yew trees grow anywhere on the globe?
<narr> Narrative: To be relevant the document must mention a place where the yew tree grows. The name of a forest or park or river is sufficient. The country's name does not have to be listed for it to be relevant.
</top>
<top> <num> Number: 787
<title> Sunflower Cultivation
<desc> Description: Give information on the cultivation of sunflowers.
<narr> Narrative: Problems in growing sunflowers and tips for success are relevant. Descriptions of how sunflowers grow are relevant. Identifications of places where sunflowers are cultivated are relevant. Information on the Ox-eye Sunflower and plants in the sunflower family are relevant.
</top>
<top> <num> Number: 788
<title> Reverse mortgages
<desc> Description: What are reverse mortgages and how do they work?
<narr> Narrative: Relevant documents will define reverse mortgages, describe requirements for qualifying, and explain how homeowners make use of them. Discussions on their advantages and disadvantages are relevant. Listings of links are relevant if they give information about reverse mortgages other than the name.
</top>
<top> <num> Number: 789
<title> abandoned mine reclamation
<desc> Description: Find information on abandoned mine reclamation projects.

<narr> Narrative: Relevant documents must have information regarding specific mine reclamation projects planned, underway, or completed. General information or generalized statistics about mine reclamation are not relevant. Also not relevant are remining projects or mining refuse reclamation.
</top>
<top> <num> Number: 790
<title> women's rights in Saudi Arabia
<desc> Description: Provide any description of laws or restrictions affecting Saudi Arabian women's rights.
<narr> Narrative: Acceptable documents must provide a narrative description of Saudi laws or restrictions affecting Saudi women. Relevance to women must be specifically mentioned; general laws affecting women only as a subset of the whole Population is not acceptable. Documents that allude to women's rights and restrictions are not acceptable unless they give an example, however vague. References to changes in the laws, improvements, etc. are all acceptable. The time period should be relatively current (past 50 years).
</top>
<top> <num> Number: 791
<title> Gullah geechee language culture
<desc> Description: Describe the historical background and present status of Gullah-Geechee language and culture
<narr> Narrative: Information on the origin and development of the Gullah-Geechee people, language and culture is relevant. Statements on relationships to other groups, languages and cultures are relevant. Present geographical limits and geographical origins are relevant. Simple identification of an individual as Gullah-Geechee is not considered relevant.
</top>
<top> <num> Number: 792
<title> Social Security means test
<desc> Description: Does Social Security use a means test?
<narr> Narrative: Relevant documents will define "means test" and describe how it is used in figuring Social Security benefits. Discussions on the way in which means tests have changed over the years are also relevant.
</top>
<top> <num> Number: 793
<title> Bagpipe Bands
<desc> Description: Give information on, and examples of, bagpipe bands.
<narr> Narrative: References to, or descriptions of, actual or fictional bagpipe bands are relevant. Abstract discussions of bagpipe bands are relevant. Bands don't have to be named. Bagpipes must make up a major part of the band to be relevant. Bagpipe-and-drum bands and bagpipe-and-accordion bands are relevant. Bands that include a nominal bagpipe or two are not relevant. A single bagpiper playing alone is not relevant, but two or more bagpipers playing together are relevant.
</top>
<top> <num> Number: 794
<title> pet therapy
<desc> Description: How are pets or animals used in therapy for humans and what are the benefits?
<narr> Narrative: Relevant documents must include details of how pet- or animal-assisted therapy is or has been used. Relevant details include information about pet therapy programs, descriptions of the circumstances in which pet therapy is used, the benefits of this type of therapy, the degree of success of this therapy, and any laws or regulations governing it.
</top>
<top> <num> Number: 795
<title> notable cocker spaniels
<desc> Description: Provide any reference to notable cockers or other spaniels.
<narr> Narrative: Relevant documents must identify a spaniel that is notable by fame (Lady), association with a famous person (Checkers), or newsworthy activity (bomb-sniffing airport security duty). Although cocker spaniels are preferred, other spaniels (such as water, Brittany, clumber, etc.) are accepted. A spaniel mentioned as a pet on

someone's Web biography will be acceptable as long as the owner has some degree of notoriety. For example, the dog of a state governor is relevant; some company website's middle manager's dog is not. A newsworthy activity can include any special talent, trick, or service. Lists of books about dogs are not acceptable unless there is some identification of the dog.
</top>
<top> <num> Number: 796
<title> Blue Grass Music Festival history
<desc> Description: Describe the history of bluegrass music and give location of bluegrass festivals.
<narr> Narrative: Any facts relating to the history of bluegrass music are relevant. Music festivals including bluegrass are relevant whether titled "bluegrass" blue grass or not. Geographical location of festivals is relevant while date or time of year is optional.
</top>
<top> <num> Number: 797
<title> reintroduction of gray wolves
<desc> Description: Where in the US have gray wolves been reintroduced in the wild?
<narr> Narrative: Relevant documents must have information about planned or accomplished reintroductions of gray wolves and include some detail as to number or location. Reintroductions of any type of gray wolf are relevant. Any other wolf species is not relevant. Proposals for, or discussions and debates about gray wolf reintroduction is not relevant.
</top>
<top> <num> Number: 798
<title> Massachusetts textile mills
<desc> Description: History, development, and locations of textile mills in Massachusetts
<narr> Narrative: Relevant documents must provide narrative information about the Massachusetts textile mills. Descriptions of preserved mill-related sites open to the public are acceptable if they contain historical information. Sites included only as part of a tour itinerary are not acceptable. Lesson plans with activities relating to mills are not acceptable. Mill references appearing only as Web links, bibliographies, or tables of contents are not acceptable.
</top>
<top> <num> Number: 799
<title> Animals in Alzheimer's research
<desc> Description: What animals have been used in Alzheimer's research?
<narr> Narrative: Animals that are or were effectively used in Alzheimer's research are relevant. Animals that were tried and found not effective in Alzheimer's research are also relevant.
</top>
<top> <num> Number: 800
<title> Ovarian Cancer Treatment
<desc> Description: The remedies and treatments given to lesson or stop effects of ovarian cancer.
<narr> Narrative: Relevant documents must include names of chemicals or medicines used to fight ovarian cancer. Studies of new treatments that are being tried are valid, even if they have not reached a conclusion as to effectiveness.
</top>
<top> <num> Number: 801
<title> Kudzu Pueraria lobata
<desc> Description: Describe the origin, nature, extent of spread and means of controlling kudzu.
<narr> Narrative: Identification of kudzu as an invasive species with description of how it spreads and grows is relevant. A document which is simply a list headed "invasive species" or "noxious weeds" including kudzu is not relevant. A statement that kudzu is present in a specific location is not relevant unless it relates to its spread. Features of kudzu such as its use as a treatment for alcoholism or its function as a haven for plant pathogens describe its nature and are relevant.
</top>
<top> <num> Number: 802

<title> Volcano eruptions global temperature
<desc> Description: What is the impact of volcano eruptions on global temperature?
<narr> Narrative: Relevant documents discuss the scientific bases for volcanic eruptions' causing warming or cooling, and may give examples of volcanoes that did. Documents which show photos or videos of volcanic eruptions without tying the volcano to global temperature, or make the connection only in a causal way are not relevant.
</top>
<top> <num> Number: 803
<title> May Day
<desc> Description: Give the history of this holiday and the various ways of celebrating May Day.
<narr> Narrative: The document should have either a discussion of the history of this holiday or a discussion of various ways of celebrating the holiday (for example, a discussion of the Maypole). A page which states that Mayday is the International Worker's Day, or that Mayday is May 1, or discusses some unrelated event that takes place around this time will not be considered relevant (e.g. Louisiana Wetlands May Day).
</top>
<top> <num> Number: 804
<title> ban on human cloning
<desc> Description: Describe resolutions proposed and legislation passed to ban the cloning of humans and the rationale for the bans.
<narr> Narrative: Relevant documents will describe the resolutions and legislation enacted to ban human cloning. Also relevant are the rationales for the bans. Documents citing arguments for or against banning of human cloning that are not in the context of a resolution or legislation are not relevant. Documents containing references to a ban on human cloning without additional information are not relevant.
</top>
<top> <num> Number: 805
<title> Identity Theft Passport
<desc> Description: Describe the Identify Theft Passport issued to identity theft victims to show to creditors and law enforcement officers questioning their credit worthiness or innocence.
<narr> Narrative: Relevant documents will describe the efforts to establish an Identity Theft Passport, the uses of the Passport, and how it helps theft identify victims. Documents with references to identity theft outside the context of the identity theft passport are not relevant.
</top>
<top> <num> Number: 806
<title> Doctors Without Borders
<desc> Description: What is Doctors Without Borders/Medecins Sans Frontieres and what do they do?
<narr> Narrative: Documents must contain a reference to some fact about Doctors Without Borders, for example, that it is a charitable organization or that it brings relief to crisis areas. A discussion of an event or operation which assumes prior familiarization with DWB is not acceptable. There is no assumption that the requester knows that the name "Medecins Sans Frontieres" is the same organization. If the two terms appear conjoined in such a way that it is clear they are synonymous, then any further reference to MSF can be assessed for relevance to the requested information. The organization name on a list of "humanitarian organizations" qualifies because at least the terminology tells us that it is an organization and that it does do humanitarian works. The name listed solely under the heading "charity" does not qualify.
</top>
<top> <num> Number: 807
<title> Sugar tariff-rate quotas
<desc> Description: Describe the nature and history of sugar tariff-rate quotas in the United States.
<narr> Narrative: Documents describing the system, its history and how it works are relevant. Proposed changes to the system or new agreements explaining how it works are relevant. Listings of current allocations are not relevant.
</top>
<top> <num> Number: 808

<title> North Korean Counterfeiting
<desc> Description: What information is available on the involvement of the North Korean Government in counterfeiting of US currency.
<narr> Narrative: A document should provide information on the counterfeiting or the distribution of counterfeit US currency by the North Korean Government. A page which provides evidence for, or quotes government officials claims that this is happening will be considered relevant. However, a page that simply states this, with no supporting evidence, will not be considered relevant.
</top>
<top> <num> Number: 809
<title> wetlands wastewater treatment
<desc> Description: Identify wastewater treatment projects that involve constructed or natural wetlands.
<narr> Narrative: Wetlands wastewater treatment projects purposely integrate wetlands to act as final filters for wastewater. The project must be named or geographically located, more precisely than simply the state in which it resides.
</top>
<top> <num> Number: 810
<title> timeshare resales
<desc> Description: Provide information regarding timeshare resales.
<narr> Narrative: Relevant documents will include those describing the prospects of reselling a timeshare and the pitfalls one should be aware of when selling a timeshare. Real estate legislature regarding the resale of timeshares is not relevant.
</top>
<top> <num> Number: 811
<title> handwriting recognition
<desc> Description: What is the state of recognizing handwritten inputs to computers?
<narr> Narrative: Relevant documents describe the methodology in carrying out handwriting recognition. Patent descriptions may be relevant if they describe the methodology used in recognizing handwritten input, such as pen-driven analysis to identify individual characters. Also relevant are documents which discuss techniques which are not character based. Relevant also is discussion of any databases being collected to study the problem of handwriting recognition as a means to inputting handwriting documents into a computer in digitized form.
</top>
<top> <num> Number: 812
<title> total knee replacement surgery
<desc> Description: What conditions lead doctors to recommend total knee replacement surgery and what complications can result from such surgery?
<narr> Narrative: Documents simply describing the surgery are not relevant. Documents concerning court cases in which doctors testify as to the need for total knee replacement or claimants charge complications from such surgery are relevant.
</top>
<top> <num> Number: 813
<title> Atlantic Intracoastal Waterway
<desc> Description: What is the Atlantic Intracoastal Waterway?
<narr> Narrative: Relevant documents describe the Atlantic Intracoastal Waterway in any of the States on the Atlantic seaboard. Relevant documents may discuss the extent to which the AIWW is used for economic and/or recreational purposes. They also may discuss the extent to which the AIWW is being maintained by federal and State governments. Documents reporting draw bridge schedules and waterway closures in emergency events such as hurricanes are not relevant.
</top>
<top> <num> Number: 814
<title> Johnstown flood

<desc> Description: Provide information about the Johnstown Flood in Johnstown, Pennsylvania.
<narr> Narrative: Documents must include any factual information about the Johnstown Flood, such as number of lives lost, high water statistics, route, towns affected, causes, etc. Tourist information sites are acceptable if they provide historical facts. Blurbs about library books are acceptable provided they contain some fact(s) about the flood and are not listed as fictional accounts. Information about the year alone (for example, "The Johnstown Flood of 1889") is not enough.
</top>
<top> <num> Number: 815
<title> Coast Guard rescues
<desc> Description: Find accounts of actual Coast Guard rescues.
<narr> Narrative: Detailed descriptions of rescues are relevant, as well as headlines, article titles or picture captions if they give sufficient detail to know what happened. Rescues involving deaths of the rescuees are relevant. Medical evacuations are relevant. Simulated rescues are NOT relevant. Searches without rescues are NOT relevant. Mere statistics on numbers of rescues are NOT relevant.
</top>
<top> <num> Number: 816
<title> USAID assistance to Galapagos
<desc> Description: Describe efforts made by USAID to protect the biodiversity in the Galapagos Islands in Ecuador.
<narr> Narrative: Relevant documents will include activities undertaken by USAID as well as USAID planned activities. Activities undertaken and funded by USAID as well as activities undertaken by USAID and funded by another organization are relevant. Documents that reference USAID assistance to the Galapagos but provide no additional information are not relevant.
</top>
<top> <num> Number: 817
<title> sports stadium naming rights
<desc> Description: How are naming rights to sports stadiums acquired?
<narr> Narrative: Documents stating how naming rights to any sports stadium were acquired are relevant. Text of any plans, laws or agreements stating how naming rights are to be acquired are also relevant.
</top>
<top> <num> Number: 818
<title> Chaco Culture National Park
<desc> Description: What is known about the culture and history of the Chaco people from features of the Chaco Culture National Historic Park?
<narr> Narrative: Documents containing description and explanation of the major archaeological findings in the park are relevant. Vague references to "wanderers" and "ruins" without specifics about the people's culture or type of ruins are not relevant. Photographs of artifacts identified simply as "Chaco" or "Chaco Anasazi" with no other elaboration are not relevant.
</top>
<top> <num> Number: 819
<title> 1890 Census
<desc> Description: What is known about the 1890 U.S. Census?
<narr> Narrative: Relevant documents include information from the 1890 U.S. Census, including generalized data like population totals for any locales, as well as any information about the conduct of the census itself. Relevant documents also may include information about the present status of the original documents of the 1890 U.S. Census.
</top>
<top> <num> Number: 820
<title> imported fire ants
<desc> Description: What are imported fire ants, and how can they be controlled?

<narr> Narrative: Relevant documents describe what imported fire ants are and where they came from. Relevant documents describe the impact of imported fire ants on the inhabitants and economy of the areas they infest. Relevant documents discuss possible methods of controlling imported fire ants, as well as the effectiveness of such methods.
</top>
<top> <num> Number: 821
<title> Internet work-at-home scams
<desc> Description: Describe the work-at-home scams that are promoted over the Internet.
<narr> Narrative: Relevant documents will include those that describe work-at-home scams and the extent of the scam problem. Documents that describe what is being done to identify and punish the fraud promoters of work-at-home scams are not relevant.
</top>
<top> <num> Number: 822
<title> Custer's Last Stand
<desc> Description: Give the history of the Battle of the Little Big Horn, June 25 and 26, 1876, also referred to as Custer's Last Stand.
<narr> Narrative: The documentation should contain information about the events leading up to the battle, discussion of the actual battle, including Reno's and Benteen's battalions, or discussion of the events immediately after the battle, including the escape of Sitting Bull to Canada, or the evacuation of Reno's wounded to Bismark. Pages discussing the monument, the burial or reburial of Custer's troops, or the battle of Washita are not relevant. Also, documents which just give the date of the battle, just give one simple fact about the battle (e.g. number killed), or are a biography of Custer which only touches on the battle are not relevant.
</top>
<top> <num> Number: 823
<title> Continuing care retirement communities
<desc> Description: What features and services are provided by continuing care retirement communities (CCRC's)?
<narr> Narrative: Documents describing facilities and services provided by CCRC's are relevant. Documents concerning licensing and requirements for operating a CCRC are not relevant.
</top>
<top> <num> Number: 824
<title> Civil Air Patrol
<desc> Description: What is the current role of the Civil Air Patrol and what training do participants receive?
<narr> Narrative: Civil Air Patrol (CAP) mission statements are relevant. CAP involvement in emergency services, communications, and aerospace education is relevant.
</top>
<top> <num> Number: 825
<title> National Guard Involvement in Iraq
<desc> Description: Describe the deployment of National Guard units to Iraq.
<narr> Narrative: Relevant documents will include those describing the deployment of specific National Guard units to Iraq. Documents describing the role of the National Guard in Iraq in general terms, such as "playing a vital role," and describing the deployment of National Guard troops to Iraq in general are not relevant.
</top>
<top> <num> Number: 826
<title> Florida Seminole Indians
<desc> Description: What is the relationship between the U.S. and the Seminole Indians of Florida?
<narr> Narrative: Relevant documents describe the history of the Seminole Indian tribes which are now now located in Florida, and their interaction with the United States. Relevant documents also describe the current relationship between the U.S. and those tribes.
</top>
<top> <num> Number: 827
<title> Hidden Markov Modeling HMM

<desc> Description: Give a definition of and/or a description of an application for the Hidden Markov Modeling algorithm.
<narr> Narrative: The documentation should either have a definition or explanation of Hidden Markov Models, or have a detailed description of an application of Hidden Markov Models to a specific problem. A discussion of the Baum-Welch algorithm, and how it relates to HMM would be acceptable. However, a document with only a statement of the form, "this software package is based on a generalized HMM.", or a title of a document (even if the title gives an application), or nothing more than a state transition diagram will not be considered relevant.
</top>
<top> <num> Number: 828
<title> secret shoppers
<desc> Description: What companies or organizations use secret or mystery shoppers?
<narr> Narrative: "Secret shoppers" or "mystery shoppers" are people who pose as customers and report on the service they received. The same function under a different name is relevant. A relevant document must clearly identify the user/employer of the secret shopper.
</top>
<top> <num> Number: 829
<title> Spanish Civil War support
<desc> Description: Provide information on all kinds of material international support provided to either side in the Spanish Civil War.
<narr> Narrative: Given that the requester knows the terminology for the various Spanish factions on each side in the Spanish Civil War, provide information on international support for either side, in the form of anything concrete, e.g. arms, aircraft, money, medicine, volunteers (such as fighters or medical personnel), etc. The recipient of such aid (i.e., which Spanish faction) should be specified. Library blurbs are acceptable as long as they mention a foreign involvement and the side taken. Verbal expressions of support from foreign leaders are not acceptable. Artistic support does not count - unless the artist actually went to Spain!
</top>
<top> <num> Number: 830
<title> model railroads
<desc> Description: Locate past or present model railroad layouts.
<narr> Narrative: Model trains in any scale and in any country are relevant. A simple exhibit of miniature train cars is not relevant; a layout with track for the trains to run on must be present.
</top>
<top> <num> Number: 831
<title> Dulles Airport security
<desc> Description: Describe the security measures at Dulles International Airport in Virginia.
<narr> Narrative: Relevant documents contain information on security measures at Dulles International Airport in Virginia and the problems with security measures that have been taken. Documents that pose questions about Dulles airport security without providing answers are not relevant.
</top>
<top> <num> Number: 832
<title> labor union activity
<desc> Description: What activity involving U.S. labor unions has taken place since 1980?
<narr> Narrative: Labor union membership figures are relevant. Strikes, recruiting, lobbying, negotiating, and support for members are all relevant. Actions taken concerning unions are relevant. Union manuals or theoretical descriptions of procedures are not relevant.
</top>
<top> <num> Number: 833
<title> Iceland government
<desc> Description: Provide information about the government of Iceland.

<narr> Narrative: Provide any information about the form of national government in Iceland. This may include any reference to the type or composition of the government or its historical development. Documents containing information on local governments alone are not acceptable. References to the activities, action, or position of the government (to include any ministries or governmental agencies) on any given topic are not adequate.
</top>
<top> <num> Number: 834
<title> Global positioning system earthquakes
<desc> Description: How is the global positioning system (GPS) used for research and monitoring of earthquakes?
<narr> Narrative: All documents referring to earthquake monitoring or research making use of a global positioning system are relevant. Documents referring to such monitoring and/or research without specific mention of the role of GPS are not relevant.
</top>
<top> <num> Number: 835
<title> Big Dig pork
<desc> Description: Why is Boston's Central Artery project, also known as "The Big Dig", characterized as "pork"?
<narr> Narrative: Relevant documents discuss the Big Dig project, Boston's Central Artery Highway project, as being a big rip-off to American taxpayers or refer to the project as "pork". Not relevant are documents which report fraudulent acts by individual contractors. Also not relevant are reports of cost-overruns on their own.
</top>
<top> <num> Number: 836
<title> illegal immigrant wages
<desc> Description: What level of wages are paid to illegal immigrants?
<narr> Narrative: The simple statement that illegals depress wages is not relevant. "Sub minimum wage" is relevant because it implies less than the legally established minimum wage. A dollar amount per hour, day, week, month, or year is an ideal answer. A percent comparison of wages paid to documented workers is relevant. Wage levels of illegal immigrants in any country are relevant.
</top>
<top> <num> Number: 837
<title> Eskimo History
<desc> Description: Provide information on the pre-1500 history of the Eskimo (Inuit) people.
<narr> Narrative: The document should have information on the history and culture of the Inuit people, previous to 1500. Pages discussing when and from where they migrated to North America are acceptable. Pages discussing the Inuit people outside of North America are acceptable, as are pages about the distribution of the Inuit people when first contacted by Europeans (as this reflects the distribution 300 years earlier). However, a statement that some aspect of Inuit culture dates dates back more than 4,000 years, does not by itself make a page relevant.
</top>
<top> <num> Number: 838
<title> urban suburban coyotes
<desc> Description: How have humans responded and how should they respond to the appearance of coyotes in urban and suburban areas?
<narr> Narrative: Documents reporting the appearance of coyotes in urban and suburban areas are relevant as are those describing human reactions or prescribing how to deal with the situation.
</top>
<top> <num> Number: 839
<title> textile dyeing techniques
<desc> Description: Explain various techniques used in dyeing textiles and identify their advantages or disadvantages.
<narr> Narrative: Abstracts describing techniques are relevant. Processes directly related to dyeing (for example, pre-treatment of material to be dyed, wastewater treatment) are relevant. Lists of techniques without explanation are not relevant.

</top>
<top> <num> Number: 840
<title> Geysers
<desc> Description: Give the definition, locations, or characteristics of geysers.
<narr> Narrative: The document should contain what makes a geyser a geyser, where geysers are found, or the names, location, and characteristics of a specific geyser. A page discussing man made geysers, just photos of geysers, geysers on other planets, or list of geysers even if associated with a particular basin will not be considered relevant.
</top>
<top> <num> Number: 841
<title> camel North America
<desc> Description: Provide information on camels in North America in both prehistoric and modern times.
<narr> Narrative: Documents may include mention and/or photos of fossils, bones, etc. Documents with photos alone, as long as they are labeled, are relevant. A casual mention of a camel presence in Late Miocene trackways is not relevant. Modern-day importation and use of camels such as that by the US military/cavalry for western desert transport is acceptable. Non-relevant documents include ones which are limited to discussions of camel-hair (imported vice domestic, with regard to NAFTA) or camel pox. Documents containing laws governing the importation of camels into the US are not relevant unless they also contain reference to specific camels being brought in. Articles referring to camels in the entertainment field such as zoos or cinema are acceptable. Fictional camels such as Joe Camel are not acceptable.
</top>
<top> <num> Number: 842
<title> David McCullough
<desc> Description: Give information about David McCullough, author, his life, works, and/or awards.
<narr> Narrative: The documentation should provide some information about David McCullough, the author of "Harry S. Truman". Pages on some other David McCullough (even if an author) will not be considered relevant. A quote from one of his books, or a complete citation of one of his books, including publisher will be considered relevant. However, the observation that someone talked with historian and writer David McCullough about history does not make the page relevant.
</top>
<top> <num> Number: 843
<title> Pol Pot
<desc> Description: Who was Pol Pot and what did he do?
<narr> Narrative: Documents must include the role of Pol Pot. It should be indicated that he was the leader of the Khmer Rouge, a Cambodian leader, responsible for mass killings, etc. Any discussion of the Khmer Rouge that alludes only to the "Pol Pot regime" or the "days of Pol Pot" is not acceptable. Bibliographies or library listings are not acceptable.
</top>
<top> <num> Number: 844
<title> segmental duplications
<desc> Description: Give information about segmental duplications in genomes.
<narr> Narrative: The documentation should either give an explanation of what is a segmental duplication, how much of the human genome is involved in segmental duplications, or what are some of the consequences of (diseases caused by or desirable effects of) miscopying a segmental duplication. Pages which contain the title of a document with this term, or the term with no indication of its meaning will not be considered relevant.
</top>
<top> <num> Number: 845
<title> New Jersey tomato
<desc> Description: Provide information about tomato farming and production in New Jersey.
<narr> Narrative: Any discussion of New Jersey tomato production, volume, percentage of US tomato production, percentage of farmland devoted to tomatoes, etc., is relevant. Articles about specific tomato farms are relevant as

long as the farms are significant for something such as size or production method or volume. Agriculture Department rulings or guidelines are acceptable as long as they specifically include the NJ tomato crop. If NJ is included with other states in production figures the article is relevant even if the NJ share of the production volume is indeterminate. (For example, "New Jersey, Maryland, and California together produce 75tables alone are not relevant. A picture of a NJ tomato or tomato farm is not relevant unless there is some associated significant data. Bibliographies are not acceptable.
</top>
<top> <num> Number: 846
<title> heredity and obesity
<desc> Description: Describe evidence that heredity does or does not play a role in obesity.
<narr> Narrative: Relevant documents will include those that provide evidence for or against the position that heredity plays a role in obesity. Documents citing planned research regarding a link between obesity and heredity are not relevant.
</top>
<top> <num> Number: 847
<title> Portugal World War II
<desc> Description: What was the role of Portugal in World War II?
<narr> Narrative: Give evidence of the role or importance of Portugal on either side (or not at all) in World War II. Documents can include any reference to gold being stored by or for Portugal, use of Portuguese ports as escape routes from Nazi Europe, Portuguese provision of arms materials to Axis countries, any reference to Portuguese status as a neutral nation, etc.
</top>
<top> <num> Number: 848
<title> radio station call letters
<desc> Description: Identify radio stations by their call letters and location or ownership.
<narr> Narrative: Call letters must be present for a document to be relevant. Either geographic location or personal or organizational ownership is relevant. Stations no longer in existence are relevant. Explanations of call letter allocation systems are not relevant unless they include actual stations with identifying information.
</top>
<top> <num> Number: 849
<title> Scalable Vector Graphics
<desc> Description: What is "scalable vector graphics"?
<narr> Narrative: Relevant documents describe the concept "scalable vector graphics", what it is used for, and what special advantages the technique has, especially with respect to portability and zooming features. Documents which merely mention the term in a list of, e.g., system requirements are not relevant.
</top>
<top> <num> Number: 850
<title> Mississippi River flood
<desc> Description: How frequently does the Mississippi River flood its banks?
<narr> Narrative: Flooding is a relative term which implies water overflowing its container and causing damage to the surrounding ares. Documents are relevant if they describe Mississippi River events which are commonly considered to be floods. Relevant documents may also show how such events have led to the introduction of controls to lessen the frequency of damaging floods of this river. Relevant documents include different levels of flooding, not only the major ones. Documents are not relevant if they are essentially forecasts or routine reports of water levels. They are also not relevant if they are purely bibliographies or lists of sources for relevant documents. Photos and videos of floods alone are not relevant.
</top>

B.7 Appendix: Topics in Blog06

<top> <num> 851 </num>
<title> "March of the Penguins" </title>
<desc> Provide opinion of the film documentary "March of the Penguins". </desc>
<narr> Relevant documents should include opinions concerning the film documentary "March of the Penguins". Articles or comments about penguins outside the context of this film documentary are not relevant. </narr>
</top>
<top> <num> 852 </num>
<title> larry summers </title>
<desc> Find opinions on Harvard President Larry Summers' comments on gender differences in aptitude for mathematics and science. </desc>
<narr> Statements of opinion on Summers' comments are relevant. Quotations of Summers without comment or references to Summers' statements without discussion of their content are not relevant. Opinions on innate gender differences without reference to Summers' statements are not relevant. </narr>
</top>
<top> <num> 853 </num>
<title> state of the union </title>
<desc> Find opinions on President Bush's 2006 State of the Union address. </desc>
<narr> All statements of opinion on the address are relevant. Descriptions of the address, quotes from the address without comment, and comedians' jokes about the address are not relevant unless there is a clear statement of opinion. Announcements that the address will take place or has taken place are not relevant. Schedules of events or discussion groups to support or oppose the address are not relevant. Predictions of what will be in the address are not relevant. </narr>
</top>
<top> <num> 854 </num>
<title> "Ann Coulter" </title>
<desc> What opinions do readers have of Ann Coulter? </desc>
<narr> Statements showing like or dislike of Ann Coulter are relevant. Any characterizations of her political orientation are also relevant. </narr>
</top>
<top> <num> 855 </num>
<title> abramoff bush </title>
<desc> Find opinions on the significance of the relationship between lobbyist Jack Abramoff and President George W. Bush. </desc>
<narr> Statements of belief, disbelief, approval or disapproval concerning the relationship between Bush and Abramoff are relevant. Quoted statements from Abramoff or Bush without other comment are not relevant. News reports which do not express an opinion are not relevant. Polls showing public opinion on the relationship are not relevant unless such a poll is cited in such a way on the blog as to show the poster's opinion. </narr>
</top>
<top> <num> 856 </num>
<title> macbook pro </title>
<desc> What has been the reaction to the Macbook Pro laptop computer? </desc>
<narr> General statements of liking or disliking the Macbook Pro are relevant. Value comparisons to earlier versions of Macintosh laptops or to other companies laptops are relevant. Product reviews are relevant if they contain opinions. Speculation about unreleased laptops is not relevant. </narr>
</top>
<top> <num> 857 </num>
<title> jon stewart </title>
<desc> Find opinions of Jon Stewart of Comedy Central's "The Daily Show". </desc>

<narr> Opinions on Stewart's performance, effectiveness and content of "The Daily Show" are relevant. Direct quotations or retelling of segments from the show without comment are not relevant. Expressions of praise or disapproval of particular jokes or lines are relevant. </narr>
</top>
<top> <num> 858 </num>
<title> "super bowl ads" </title>
<desc> What is the opinion of advertisements shown during the Super Bowl? </desc>
<narr> Documents should reveal whether the ads during the Super Bowl were enjoyed by viewers. The documents may mention which ads were liked better than others. Documents describing the ads and providing web sites where they may be viewed are not relevant. </narr>
</top>
<top> <num> 859 </num>
<title> "letting india into the club?" </title>
<desc> Find pro and con views on accepting India as a nuclear power. </desc>
<narr> Documents that discuss whether India should be allowed to have nuclear weapons and nuclear technology are relevant. Opinions on the US-brokered agreement to which India agrees to assume the same responsibilities as other nuclear weapons countries, particularly in relation to IAEA safeguards, are also relevant. </narr>
</top>
<top> <num> 860 </num>
<title> "arrested development" </title>
<desc> What is the opinion of the TV program "Arrested Development"? </desc>
<narr> Documents should reveal whether Arrested Development was a favorite or not and the reasons for watching it or not. The documents may specify programs that were enjoyed and those that were not. Documents mentioning Fox and whether the station was liked or not when Arrested Development was canceled are not relevant. </narr>
</top>
<top> <num> 861 </num>
<title> mardi gras </title>
<desc> Provide opinions concerning the festival of Mardi Gras. </desc>
<narr> Relevant documents should include statements of opinion concerning the festival of Mardi Gras. If news media reports included either editorial or quoted opinion, the document is relevant. Items of a purely commercial nature are not relevant. </narr>
</top>
<top> <num> 862 </num>
<title> blackberry </title>
<desc> How does the Blackberry cellphone compare in performance to other cellphones. </desc>
<narr> A document should reveal the functions of push email, phone capability and web browsing of the blackberry and how this compares in usage and performance to other cellphones. The document should indicate the reason for either purchasing or using this cellphone and why it was chosen over others. Documents which mention the product without any opinion as to how it is liked or not are considered not relevant. </narr>
</top>
<top> <num> 863 </num>
<title> netflix </title>
<desc> Identify documents that show customer opinions of Netflix. </desc>
<narr> A relevant document will indicate subscriber satisfaction with Netflix. Opinions about the Netflix DVD allocation system, promptness or delay in mailings are relevant. Indications of having been or intent to become a Netflix subscriber that do not state an opinion are not relevant. </narr>
</top>
<top> <num> 864 </num>
<title> colbert report </title>
<desc> Find opinions of the Comedy Central show "The Colbert Report." </desc>

<narr> Descriptions of the show or its star in a positive or negative vein are relevant. Reporting on the nature of the show without opinion is not relevant. Direct quotes from the show without comment are not relevant. Complaints about missing the show, comparisons of The Colbert Report to other shows, and petitions to bring the show to Canadian TV are all relevant opinions. </narr>
</top>
<top> <num> 865 </num>
<title> basque </title>
<desc> What are the reactions worldwide to Spain's Basque separatist movements? </desc>
<narr> Documents which convey official and unofficial positive or negative opinions about Basque violence, terrorism or other displays of separatist activities are relevant. Documents describing the Basque region without reference to Basque separatists are not relevant. </narr>
</top>
<top> <num> 866 </num>
<title> "Whole Foods" </title>
<desc> Find opinions on the quality, expense, and value of purchases at Whole Foods stores. </desc>
<narr> All opinions on the quality, expense and value of Whole Foods purchases are relevant. Comments on business and labor practices or Whole Foods as a stock investment are not relevant. Statements of produce and other merchandise carried by Whole Foods without comment are not relevant. </narr>
</top>
<top> <num> 867 </num>
<title> cheney hunting </title>
<desc> Find opinions on Vice President Cheney's quail-hunting trip on which a companion was wounded. </desc>
<narr> Only opinions on the particular quail-hunting trip on which the accident occurred are relevant. Opinions about quail hunting in general or Cheney's quail hunting in general are not relevant. Quotes of official news releases are not relevant; only opinions expressed on blogs. Comical parodies on what other newsworthy people might say about the trip are not relevant. Jokes about the trip and accident that don't state an opinion are not relevant, but if the joke is from a blogger, not a comedian, and is heavily sarcastic in one direction or another it represents an opinion and is relevant. </narr>
</top>
<top> <num> 868 </num>
<title> "joint strike fighter" </title>
<desc> Provide opinion concerning the aviation defense program "Joint Strike Fighter". </desc>
<narr> Relevant documents should include opinions concerning the "Joint Strike Fighter" aviation defense program. News media reports are relevant when editorial or quoted opinions concerning the program are included. Articles concerning the computer game "JSF-Joint Strike Fighter" are not relevant. </narr>
</top>
<top> <num> 869 </num>
<title> muhammad cartoon </title>
<desc> Find opinions worldwide to the cartoons depicting the Muslim prophet Muhammad printed in a Danish newspaper. </desc>
<narr> Relevant documents contain a direct opinion on the cartoons. Descriptions of protest demonstrations and reactions by public officials are not relevant. Quotes from or links to news reports are not relevant unless they contain an opinion about the cartoons specifically. </narr>
</top>
<top> <num> 870 </num>
<title> "barry bonds" </title>
<desc> Find opinions on alleged use of steroids by baseball player Barry Bonds. </desc>
<narr> Relevant documents will refer to Barry Bonds' use of steroids. Comments about Bonds personally or as a baseball player are not relevant unless they mention allegations of drug use. </narr>
</top>

<top> <num> 871 </num>
<title> cindy sheehan </title>
<desc> What has been the reaction to Cindy Sheehan and the demonstrations she has been involved in? </desc>
<narr> Any favorable or unfavorable opinions of Cindy Sheehan are relevant. Reactions to the anti-war demonstrations she has organized or participated in are also relevant. </narr>
</top>
<top> <num> 872 </num>
<title> brokeback mountain </title>
<desc> What opinions have been posted on the movie Brokeback Mountain? </desc>
<narr> Relevant documents should indicate that the movie was liked or disliked. Judgments of the movie's content, quality of presentation, or timeliness are also relevant. A description of the movie without review or assessment is not relevant. </narr>
</top>
<top> <num> 873 </num>
<title> "bruce bartlett" </title>
<desc> What opinions do readers and viewers have of columnist and commentator Bruce Bartlett. </desc>
<narr> Any opinion of Bruce Bartlett and his writing is relevant. Also relevant are characterizations of his political assessments and of his political orientation. </narr>
</top>
<top> <num> 874 </num>
<title> coretta scott king </title>
<desc> Find opinions of speeches and eulogies delivered at the funeral of Coretta Scott King. </desc>
<narr> Opinions on the speeches delivered at the funeral of Coretta Scott King are relevant. Straight reporting on the speeches is not relevant. Vague references to "lack of civility", "politicizing" or "spreading propaganda" at the funeral with no indication of the content of the speeches are not relevant. </narr>
</top>
<top> <num> 875 </num>
<title> american idol </title>
<desc> What is the public's opinion on the TV program "American Idol"? </desc>
<narr> Relevant documents present an opinion of the show "American Idol". Those which refer to the popularity and quality of the program, and the behavior and fairness of the judges are relevant. Opinions on the individual performers are not relevant. </narr>
</top>
<top> <num> 876 </num>
<title> "life on mars" </title>
<desc> Provide opinion of the BBC television series "Life on Mars". </desc>
<narr> Relevant documents should include statements of opinion concerning the BBC television series "Life on Mars". Any item not specifically about the TV show, such as a science article about the planet Mars, is not relevant. </narr>
</top>
<top> <num> 877 </num>
<title> sonic food industry </title>
<desc> How well are Sonic fast-food restaurants liked by those eating at fast food establishments. </desc>
<narr> A document should reveal the reason why Sonic fast food is enjoyed and what foods are liked best. The document may include why Sonic is chosen and what consistently brings the person back there. Documents which mention Sonic games are not relevant. </narr>
</top>
<top> <num> 878 </num>
<title> jihad </title>
<desc> Provide opinion concerning the jihad movement. </desc>

<narr> Relevant documents should include statements of opinion concerning the jihad movement. Jihad is generally defined as holy war dedicated to the expansion of Islam. News media reports are relevant when they include editorial or quoted opinions concerning the jihad movement, or in reports about individual factions involved in the jihad movement. Articles having tag lines such as "women's jihad" or "cartoon jihad" are only relevant when opinions are in the context of holy war and/or Islam, and not when emphasizing women's rights movements or internal strife in certain countries. </narr>
</top>
<top> <num> 879 </num>
<title> "hybrid car" </title>
<desc> Find reviews and ratings of hybrid cars in current production. </desc>
<narr> Relevant documents are those which rate hybrid cars in terms of performance and prospects for future success. A document which mentions a hybrid car without giving any opinion of it is not relevant. </narr>
</top>
<top> <num> 880 </num>
<title> "natalie portman" </title>
<desc> Provide opinion concerning the actress Natalie Portman. </desc>
<narr> Relevant documents should include opinion concerning the actress Natalie Portman. Articles or comments with opinions about her work, appearance, or personal history are relevant. News media reports are relevant when editorial or quoted opinions are present. </narr>
</top>
<top> <num> 881 </num>
<title> "Fox News Report" </title>
<desc> Is the Fox News Report perceived to be fair or biased in selecting and reporting news topics? </desc>
<narr> Relevant documents include opinions which would indicate bias or fairness in reporting. Also relevant are opinions that some topics in the Fox News Report receive unwarranted emphasis. Opinions regarding presentation such as logos and studio sets are not relevant. </narr>
</top>
<top> <num> 882 </num>
<title> seahawks </title>
<desc> What are fans' opinions regarding the Seattle Seahawks football team? </desc>
<narr> Relevant documents should describe whether the team is liked or not. Document containing emotional reactions to the Seahawks are relevant. The sentiments of the person may reveal their reasons for either a positive or negative response to the team. Documents describing team members or how a game was played without expressing an opinion are not relevant. </narr>
</top>
<top> <num> 883 </num>
<title> heineken </title>
<desc> Find opinions on the taste, quality, health effects and marketing of Heineken's Beer. </desc>
<narr> Opinions to Heineken's beers such as "good", "terrible", "like water", etc. are relevant. Statistics rating favorite beers and advertisements stating that Heineken's is "favorite", "number 1", etc. are judged relevant. Pictures featuring Heineken's or advice to try Heineken's without comment are not relevant. </narr>
</top>
<top> <num> 884 </num>
<title> Qualcomm </title>
<desc> What are opinions of Qualcomm, its business practices, marketing and products? </desc>
<narr> Any opinion of Qualcomm's business practices, marketing procedures or products are relevant. Statements of mergers or takeovers are relevant only if they serve to alter the opinion of the company. Any reference to Qualcomm Stadium used by the San Diego Chargers is not relevant. </narr>
</top>
<top> <num> 885 </num>

<title> shimano </title>
<desc> Provide opinions on equipment using the brand name Shimano. </desc>
<narr> Relevant documents should include opinions concerning either bicycle or fishing equipment using the brand name Shimano. An advertisement quoting a consumer opinion or review is relevant. Ad copy using arbitrary comparative terms (cheaper, more solid, stronger, etc) without opinion is not relevant. </narr>
</top>
<top> <num> 886 </num>
<title> "west wing" </title>
<desc> Provide opinion concerning the television series West Wing. </desc>
<narr> Relevant documents should include opinion or reviews concerning history, actors or production information for the television series "West Wing". Simple news updates or media reports are relevant only when they include editorial or quoted opinions. Articles or comments about the real-world White House west wing are not relevant. </narr>
</top>
<top> <num> 887 </num>
<title> World Trade Organization </title>
<desc> Find reaction worldwide to the results of the World Trade Organization meetings in the years 2004 and 2005. </desc>
<narr> Relevant documents are those which provide commentary on decisions about changes in trade relations, marketing and investment. Coverage or opinions about protests of the WTO are not relevant. </narr>
</top>
<top> <num> 888 </num>
<title> audi </title>
<desc> What are opinions of the Audi automobiles and the company that produces them? </desc>
<narr> A relevant document must indicate an opinion of the Audi automobile and company. Statements regarding introduction of new models or upgrading capabilities of an existing line of Audis are not relevant unless used to form or support an opinion. </narr>
</top>
<top> <num> 889 </num>
<title> scientology </title>
<desc> Provide opinion concerning the Scientology religion. </desc>
<narr> Relevant documents should include opinion concerning either the history, beliefs, recruiting practices, or other information concerning Scientology. Articles or comments simply reporting an individual's membership in the Scientology religion without including opinions in one or more of the above Scientology-related contexts are not relevant. Reports by news media are relevant when editorial or quoted opinions are present. </narr>
</top>
<top> <num> 890 </num>
<title> olympics </title>
<desc> Find opinions on the overall appeal and impression of the Winter and Summer Olympics. </desc>
<narr> Relevant documents are those pertaining to overall impressions of the Olympic games. Documents dealing only with specific subjects such as particular ceremonies, events or individual participants are not relevant. </narr>
</top>
<top> <num> 891 </num>
<title> intel </title>
<desc> What is the opinion of Intel processors in personal computers? </desc>
<narr> A document should reveal consumers' sentiments regarding Intel processors in PC's. The documents may include the attitude towards introducing Intel processors into Apple computers. Documents which merely discuss the new logo introduced by Intel are considered not relevant. </narr>
</top>
<top> <num> 892 </num>

<title> "Jim Moran" </title>
<desc> Find opinions of positions, statements, votes and behavior of Congressman Jim Moran (D-Va). </desc>
<narr> Statements of approval, disapproval or neutrality towards specific activities of the Congressman are relevant. References to his sponsorship of "town meetings" or reports on his activities without comment or obvious bias are not relevant. </narr>
</top>
<top> <num> 893 </num>
<title> zyrtec </title>
<desc> Is Zyrtec considered an effective medication and does it have side effects? </desc>
<narr> Opinions as to the effectiveness of Zyrtec and identification of any side effects are relevant. The mere fact of someone taking Zyrtec where no indication of its effect is given is not relevant. </narr>
</top>
<top> <num> 894 </num>
<title> board chess </title>
<desc> Provide opinion concerning the board game Chess. </desc>
<narr> Relevant documents should include statements of opinion concerning the board game Chess. The term board chess is generally defined to be the traditional game of chess using 32 pieces and played on a board having 64 black and white squares. All articles or comments having editorial or quoted opinions concerning individual games, tournaments, books, rules, membership, or general information about the game of chess are relevant. Articles or comments concerning non-traditional games such as literary chess, card chess, battle chess, etc., are not relevant. </narr>
</top>
<top> <num> 895 </num>
<title> Oprah </title>
<desc> Find opinions about Oprah Winfrey's TV show. </desc>
<narr> Relevant documents are those which refer to her TV program itself. Opinions of Oprah's other activities or about Oprah as a person are not relevant, unless they come in the context of her TV show. </narr>
</top>
<top> <num> 896 </num>
<title> global warming </title>
<desc> What is the opinion about global warming? </desc>
<narr> A document should reveal concern pertaining to global warming and the effects there is to the environment. The concerns of scientists about the effects of climate change, for example an increase in the number of hurricanes, is relevant. The documents should describe the critical environmental challenge which is faced concerning global warming. </narr>
</top>
<top> <num> 897 </num>
<title> ariel sharon </title>
<desc> What is the opinion regarding Ariel Sharon? </desc>
<narr> Documents should reveal the sentiment toward the leadership of Ariel Sharon. The documents may also mention his illness and how this affects the current stance toward him. Documents relating to his biographical history are not relevant. </narr>
</top>
<top> <num> 898 </num>
<title> Business Intelligence Resources </title>
<desc> What is the opinion of available business intelligence resources? </desc>
<narr> Relevant documents should indicate current resources available for business intelligence which will enhance business performance. The documents should indicate the usefulness of the resources in the current business arena. Not relevant are documents that only mention the topic of business intelligence resources. </narr>
</top>

<top> <num> 899 </num>
<title> cholesterol </title>
<desc> What is the concern regarding cholesterol? </desc>
<narr> Documents should reveal the attitude towards cholesterol and how the available medications have affected lowering the cholesterol. The documents may contain how lifestyle affects the individual's cholesterol. Documents describing good and bad cholesterol are not relevant. </narr>
</top>
<top> <num> 900 </num>
<title> mcdonalds </title>
<desc> Find opinions regarding the food at McDonald's restaurants. </desc>
<narr> Relevant documents give an opinion about the food in McDonald's restaurants. Documents which refer to dietary problems, nutrition, and health issues resulting from eating Mcdonalds foods are relevant. Opinions of McDonald's as a company are not relevant. </narr>
</top>
<top> <num> 901 </num>
<title> jstor </title>
<desc> Find opinions on JSTOR, the system developed to make scholarly journals available from a digital archive. </desc>
<narr> Reports of difficulty or ease in using JSTOR are relevant opinions. A statement that one is lucky to have access or wishes to have access to JSTOR is a relevant opinion. A statement that information is available in JSTOR is not an opinion. Simply citing JSTOR as a reference for a document is not an opinion. </narr>
</top>
<top> <num> 902 </num>
<title> lactose gas </title>
<desc> Find opinions about lactose gas, related symptoms and remedies. </desc>
<narr> All opinions about lactose gas, its related symptoms and remedies are relevant. </narr>
</top>
<top> <num> 903 </num>
<title> "Steve jobs" </title>
<desc> Find documents stating opinions about Apple CEO Steve Jobs. </desc>
<narr> Relevant documents will state opinions about Steve Jobs, the head of Apple Computer. Documents will include comments on his great success with the iPod, his management style, and his unusual keynote presentations he gives at the introduction of new products. </narr>
</top>
<top> <num> 904 </num>
<title> alterman </title>
<desc> Find opinions about author and columnist Eric Alterman. </desc>
<narr> Any opinions about author/columnist and blogger Eric Alterman are relevant. Also pertinent are comments on positions Alterman has taken on controversial subjects. </narr>
</top>
<top> <num> 905 </num>
<title> king funeral </title>
<desc> Find documents that show opinions on the funeral of Coretta Scott King. </desc>
<narr> Opinions on the political nature of speeches given by top government officials and others at the funeral of the civil rights icon, Coretta Scott King, are relevant. Opinions on other aspects of the funeral are also relevant. Not relevant are mention of the King funeral in lists of news topics. </narr>
</top>
<top> <num> 906 </num>
<title> davos </title>
<desc> Find opinions on the World Economic forum in Davos, Switzerland. </desc>

<narr> Relevant documents will include opinions about the World Economic Forum held annually in Davos, Switzerland, and the topics or people at the most recent Forum. Opinions on economic topics not connected with Davos are not relevant. </narr>
</top>
<top> <num> 907 </num>
<title> brrreeeport </title>
<desc> Find opinions about the Brrreeeport experiment. </desc>
<narr> Robert Scoble requested bloggers to attach the coined word brrreeeport to their blogs. The purpose was to determine how soon that word would show up on various search engines. Opinions of the experiment and its author are relevant. Opinions of the Microsoft company, where Scoble worked, are not relevant. Opinions regarding search engine reactions to brrreeeport are not relevant. </narr>
</top>
<top> <num> 908 </num>
<title> "carrie underwood" </title>
<desc> Find opinions of singer Carrie Underwood. </desc>
<narr> Opinions of singer Carrie Underwood, personally and of her singing ability are relevant. Opinions of songs she has written are also relevant. Statements regarding her appearance in any particular venue can be relevant but do not necessarily imply an opinion. </narr>
</top>
<top> <num> 909 </num>
<title> Barilla </title>
<desc> Find opinions on Barilla brand pasta or other foods. </desc>
<narr> Indication of Barilla as one's brand of choice is a relevant opinion. The simple statement that Barilla was used is not a relevant opinion because Barilla may have been all that was available. </narr>
</top>
<top> <num> 910 </num>
<title> "Aperto Networks" </title>
<desc> Find opinions about either the company Aperto Networks or its products. </desc>
<narr> All opinions about either the company itself or the products produced by Aperto Networks are relevant. </narr>
</top>
<top> <num> 911 </num>
<title> SCI FI CHANNEL </title>
<desc> Find opinions regarding the cable television Sci-Fi Channel. </desc>
<narr> All statements of opinion regarding the Sci-Fi Channel are relevant. All statements regarding programming, such as show titles or reviews, are relevant provided the name Sci-Fi Channel is included. </narr>
</top>
<top> <num> 912 </num>
<title> nasa </title>
<desc> Find opinions about NASA, the National Aeronautics and Space Administration. </desc>
<narr> All statements of opinion regarding the National Aeronautics and Space Administration (NASA) are relevant. Statements of opinion about NASA projects or programs are relevant provided that NASA is mentioned. </narr>
</top>
<top> <num> 913 </num>
<title> sag awards </title>
<desc> Find opinions regarding the Screen Actors Guild Awards (SAG). </desc>
<narr> All statements of opinion regarding the Screen Actors Guild Awards are relevant. Statements of opinion regarding actors or other participants or film productions are relevant provided the Screen Actors Guild Awards or SAG are mentioned. </narr>
</top>

<top> <num> 914 </num>
<title> northernvoice </title>
<desc> Find opinions about the Canadian blogging conference "NorthernVoice." </desc>
<narr> All opinions expressed about the NorthernVoice blogging conference are relevant, including comments by attendees and conference organizers. </narr>
</top>
<top> <num> 915 </num>
<title> "allianz" </title>
<desc> Find opinions of the international insurance company Allianz. </desc>
<narr> Opinions of the international insurance company Allianz and its business and personnel practices are relevant. Also, any opinions about Allianz subsidiary companies or the process creating them are pertinent. Opinions about the Allianz Stadium are not relevant. </narr>
</top>
<top> <num> 916 </num>
<title> dice.com </title>
<desc> Find opinions concerning dice.com, an on-line job search site. </desc>
<narr> Opinions on dice.com's effectiveness are relevant. Mention of its problems is relevant. Recounting an experience using dice.com is relevant. Simply mentioning it as a possible tool is not relevant. </narr>
</top>
<top> <num> 917 </num>
<title> snopes </title>
<desc> Find opinions on the urban legends reference site snopes.com. </desc>
<narr> Recommendations to check snopes.com are relevant opinions. Opinions on the effectiveness of Snopes or the quality of its advice are relevant. A statement that Snopes confirms or debunks something is not relevant by itself. Opinions of stories or myths found on Snopes are not relevant. Report of use of Snopes is not relevant. </narr>
</top>
<top> <num> 918 </num>
<title> varanasi </title>
<desc> Find opinions about the city of Varanasi, India. </desc>
<narr> Relevant opinions should be about the city of Varanasi and its people, either currently or historically. </narr>
</top>
<top> <num> 919 </num>
<title> pfizer </title>
<desc> Find opinions of the drug company Pfizer and its products. </desc>
<narr> Any opinion of the drug company Pfizer is relevant. Opinions of Pfizer products are also relevant. Statements that a Pfizer product is being used are not relevant unless accompanied by an opinion of like or dislike or the effectiveness of the product. </narr>
</top>
<top> <num> 920 </num>
<title> "andrew coyne" </title>
<desc> Find opinions about the Canadian journalist Andrew Coyne. </desc>
<narr> Opinions about the journalist Andrew Coyne himself and about his writings are all relevant. </narr>
</top>
<top> <num> 921 </num>
<title> "Christianity Today" </title>
<desc> Find opinions on the Evangelical Protestant magazine "Christianity Today". </desc>
<narr> Statements of opinions regarding content from or descriptions of the Evangelical Protestant magazine "Christianity Today" are relevant. </narr>
</top>

<top> <num> 922 </num>
<title> "howard stern" </title>
<desc> Find opinions of shock jock Howard Stern. </desc>
<narr> Opinions of Howard Stern and his radio program are relevant. Comments and opinions expressed by Stern are not relevant. Opinions expressed by guests on the program are relevant only if about Stern or the program but not if about the program's topic. </narr>
</top>
<top> <num> 923 </num>
<title> challenger </title>
<desc> Find opinions about the Challenger space shuttle disaster. </desc>
<narr> Relevant opinions should be specifically about the space shuttle Challenger disaster event, not about other Challenger missions, crews, or the shuttle program generally. </narr>
</top>
<top> <num> 924 </num>
<title> "mark driscoll" </title>
<desc> Find opinions about the evangelist Mark Driscoll. </desc>
<narr> Opinions about Mark Driscoll include value judgments expressed about his appearance, behavior or views (observed, inferred, or imputed). Apparent distortions of fact about him constitute relevant opinions. </narr>
</top>
<top> <num> 925 </num>
<title> mashup camp </title>
<desc> Find opinions about the first Mashup Camp event. </desc>
<narr> Relevant opinions should be about the first real Mashup Camp event, other events that used that same name, and similar events focused on the mashup approach to software-based services development. References to the mashup programming approach, per se, are not relevant. </narr>
</top>
<top> <num> 926 </num>
<title> hawthorne heights </title>
<desc> Find opinions on the band Hawthorne Heights, its music, or its members. </desc>
<narr> Opinions on activities of the band's members are relevant. The presence of the band's music in a playlist is not a relevant opinion. A decision to listen to the band's music or see the band in concert is not an opinion. </narr>
</top>
<top> <num> 927 </num>
<title> oscar fashion </title>
<desc> Find opinions on dresses or other garments worn at the Academy Awards, or fashions that are designed for Oscar night. </desc>
<narr> Predictions of what will be worn or what designers will be represented are relevant opinions. A statement that a garment is appropriate for the Oscars is a relevant opinion. Hairstyles are not relevant. A statement that a gown has been rejected, with no other information about it, is not relevant. Fashions for other award ceremonies are not relevant. "Oscar" or "Academy Award" must be mentioned in connection with clothing for a document to be relevant. </narr>
</top>
<top> <num> 928 </num>
<title> "big love" </title>
<desc> Find opinions regarding the HBO television show "Big Love". </desc>
<narr> All statements of opinion regarding the HBO production "Big Love" are relevant. Statements of opinion about HBO or actors in the show are relevant provided that "Big Love" is mentioned. </narr>
</top>
<top> <num> 929 </num>
<title> "brand manager" </title>

<desc> Find opinions about individual brand managers. </desc>
<narr> Relevant opinions should be about individuals who are brand managers identified by their name, product, brand or company, not about the brand manager process generally. </narr>
</top>
<top> <num> 930 </num>
<title> ikea </title>
<desc> Find opinions on Ikea or its products. </desc>
<narr> Recommendations to shop at Ikea are relevant opinions. Recommendations of Ikea products are relevant opinions. Pictures on an Ikea-related site that are not related to the store or its products are not relevant. </narr>
</top>
<top> <num> 931 </num>
<title> fort mcmurray </title>
<desc> Find opinions on the Canadian oil boom town of Fort McMurray. </desc>
<narr> Opinions about the people who make up the community are relevant. Opinions of Fort McMurray residents are not relevant unless the opinion concerns or reflects on the town. Descriptions of the town that show bias are relevant. </narr>
</top>
<top> <num> 932 </num>
<title> goobuntu </title>
<desc> Find opinions expressed about the rumor that Google is planning a desktop operating system named "Goobuntu." </desc>
<narr> All opinions expressed about "Goobuntu" are relevant, regardless of the fact that there was really no operating system by that name. </narr>
</top>
<top> <num> 933 </num>
<title> "winter olympics" </title>
<desc> Find opinions on the Winter Olympics. </desc>
<narr> All statements of opinion about the Winter Olympic games are relevant. Statements about the Summer Olympics are not relevant. </narr>
</top>
<top> <num> 934 </num>
<title> cointreau </title>
<desc> Find opinions of the liqueur Cointreau and drinks or food containing Cointreau. </desc>
<narr> Statements regarding like or dislike of Cointreau are pertinent as are opinions about its addition to drinks or food items. Opinions regarding the business practices of the company that owns Cointreau are not relevant. </narr>
</top>
<top> <num> 935 </num>
<title> mozart </title>
<desc> Find opinions regarding the composer Wolfgang Amadeus Mozart. </desc>
<narr> All statements of opinion regarding the composer Wolfgang Amadeus Mozart are relevant. All statements of opinion regarding music authored by Mozart are relevant. Statements of opinions regarding events, festivals, or publications using Mozart's name are relevant. </narr>
</top>
<top> <num> 936 </num>
<title> grammys </title>
<desc> Find opinions regarding the Grammy awards. </desc>
<narr> All statements of opinion regarding the annual television Grammy awards are relevant. Statements of opinions of contestants or television productions are relevant provided the names "Grammy" or "Grammys" are mentioned in the statement. </narr>
</top>

<top> <num> 937 </num>
<title> LexisNexis </title>
<desc> Find opinions about the information service LexisNexis. </desc>
<narr> Relevant documents will provide opinions about the information service LexisNexis. Documents that are obviously sponsored by LexisNexis are considered to be spam and not relevant. </narr>
</top>
<top> <num> 938 </num>
<title> "plug awards" </title>
<desc> Search for opinions about the "Plug Awards" annual award show. </desc>
<narr> All opinions are relevant, including those of promoters or attendees. The "Plug Awards" show pits deserving independent and underground artists against each other. </narr>
</top>
<top> <num> 939 </num>
<title> "Beggin Strips" </title>
<desc> Find opinions about Beggin' Strips brand dog snacks. </desc>
<narr> All opinions about Beggin' Strips dog treats are relevant. </narr>
</top>
<top> <num> 940 </num>
<title> Lance Armstrong </title>
<desc> Find opinions about the seven-time Tour de France cyclist Lance Armstrong. </desc>
<narr> All opinions about Lance Armstrong are relevant, whether related to the Tour de France race or not. </narr>
</top>
<top> <num> 941 </num>
<title> "teri hatcher" </title>
<desc> Find opinions about the actress Teri Hatcher. </desc>
<narr> All statements of opinion regarding the persona or work of film and television actress Teri Hatcher are relevant. </narr>
</top>
<top> <num> 942 </num>
<title> lawful access </title>
<desc> Find documents providing opinions about lawful access by the government to private files. </desc>
<narr> Relevant documents will state opinions on whether or not government agents had lawful access to private files they have examined in recent years. These files included e-mails, library and phone records, etc., of private U.S. citizens as well as aliens. Documents about actions of both Canadian and government agents are relevant. Documents on Internet access are not relevant. </narr>
</top>
<top> <num> 943 </num>
<title> censure </title>
<desc> Find opinions on bills in the U.S. Congress to censure President Bush and Vice-president Cheney. </desc>
<narr> Relevant documents should include opinions on bills in the U.S. Congress, sponsored by Representative Conyers, to censure President Bush and Vice President Cheney. Not relevant are actions to censure other people or other countries, i.e. Iran in the UN. </narr>
</top>
<top> <num> 944 </num>
<title> "Opera Software" OR "Opera Browser" OR "Opera Mobile" OR "Opera Mini" </title>
<desc> Find opinions about the "Opera" web browser. </desc>
<narr> General information about the Opera Browser and its features does not constitute an opinion. Simple links (e.g., "Download the best and fastest...") with no supporting text are also not relevant. All other opinions about

either the company or the Opera web browser product are relevant. For example, how does it compare with other browsers? Is it faster? More reliable? Better features? Liked or disliked? </narr>
</top>
<top> <num> 945 </num>
<title> bolivia </title>
<desc> Find documents that show opinions about Bolivia. </desc>
<narr> Relevant documents will state opinions about Bolivia and the government headed by its new indigenous president, Evo Morales. Documents that provide opinions about Bolivia other than its government can be relevant, i.e. opinions about new businesses and changed aspects of a city. </narr>
</top>
<top> <num> 946 </num>
<title> tivo </title>
<desc> Find opinions about TIVO brand digital video recorders </desc>
<narr> Relevant opinions should be about TIVO technology, functionality, marketing, benefits to users, and comparisons with similar products. </narr>
</top>
<top> <num> 947 </num>
<title> sasha cohen </title>
<desc> Find opinions about Olympic figure skater Sasha Cohen. </desc>
<narr> All opinions expressed about figure skater Sasha Cohen are relevant. She should not be confused with Sacha Baron Cohen. </narr>
</top>
<top> <num> 948 </num>
<title> sorbonne </title>
<desc> Find opinions about the Sorbonne. </desc>
<narr> Documents that state opinions about the Sorbonne University in Paris are relevant. Mention of books written by professors currently at the Sorbonne, or notations about persons having studied or given speeches there are not relevant. </narr>
</top>
<top> <num> 949 </num>
<title> ford bell </title>
<desc> Find opinions of Minnesota veterinarian and political candidate Ford Bell. </desc>
<narr> Opinions of Ford Bell's character and his viability as a political candidate are relevant. Opinions of him as a veterinarian are also relevant. Opinions of the actions of other political elements that affected Ford's candidacy are not relevant. </narr>
</top>
<top> <num> 950 </num>
<title> Hitachi Data Systems </title>
<desc> Locate opinions about either the Hitachi Data Systems company or its products. </desc>
<narr> All opinions about Hitachi Data Systems are relevant, including those expressed in advertising by the company itself, with the exception of spam. </narr>
</top> <top> <num> 1001 </num> <title> Carmax </title>
<desc> Find opinions of people who have sold a car, purchased a car, or both, through Carmax. </desc>
<narr> Relevant documents will include experiences from people who have bought or sold a car through Carmax and expressed an opinion about the experience. Do not include posts where people obtain estimates from Carmax but do not buy or sell an auto with Carmax. </narr>
</top>
<top> <num> 1002 </num> <title> Wikipedia primary source </title>
<desc> Find positive and negative opinions of the use of Wikipedia as a primary source of information by students. </desc>

<narr> Student use of Wikipedia as a resource for term papers, etc. is controversial among educators and librarians. Opinions expressed about Wikipedia as a general or secondary research source are not considered relevant. </narr>
</top>
<top> <num> 1003 </num> <title> Jiffy Lube </title>
<desc> What is the opinion of the services provided by Jiffy Lube? </desc>
<narr> Jiffy Lube is a franchise company that provides automobile maintenance services, particularly oil changes. Documents should indicate whether the service provided by Jiffy Lube is viewed as good or poor. Comments about help provided by people at Jiffy Lube, even if not part of their usual services, will be considered as a positive opinion. </narr>
</top>
<top> <num> 1004 </num> <title> Starbucks </title>
<desc> What do people think about the Starbucks chain of coffee shops? </desc>
<narr> Any opinion of Starbucks and their products and services is relevant. Opinions of Starbucks' business practices, their ubiquity, etc are also relevant. </narr>
</top>
<top> <num> 1005 </num> <title> Windows Vista </title>
<desc> Find opinions about the Microsoft Windows Vista operating system or any of its features. </desc>
<narr> Relevant documents should express views that refer specifically to the Windows Vista operating system, either generally, or any features included in the operating system. </narr>
</top>
<top> <num> 1006 </num> <title> Mark Warner for President </title>
<desc> Find positive and/or negative opinions of Mark Warner as a potential presidential candidate. </desc>
<narr> Blog posts should express an opinion about Mark Warner as a presidential candidate. Posts that contain opinions of Mark Warner's tenure and/or policies as Governor of Virginia without reference to his presidential candidacy are not relevant. </narr>
</top>
<top> <num> 1007 </num> <title> women in Saudi Arabia </title>
<desc> What is the opinion of the treatment of women in Saudi Arabia? </desc>
<narr> Relevant documents should show the views of people regarding Saudi Arabian attitudes toward women and restrictions placed upon women in Saudi society. Documents containing opinions of the conservative Wahabi branch of Islam, the dominant form of Islam in Saudi Arabia, are relevant provided they concern the role and rights of women. Also relevant are posts that describe how Saudi attitudes and society are changing. Comments about who might be required to wear abayas, or that generally these views need to change, will be considered negative opinions. </narr>
</top>
<top> <num> 1008 </num> <title> UN Commission on Human Rights </title>
<desc> Is the UN Commission on Human Rights viewed as effective? </desc>
<narr> The United Nations Commission on Human Rights (UNCHR) was the UN's principal mechanism and international forum concerned with the promotion and protection of human rights. The Commission held its final meeting on March 27, 2006. Relevant documents will indicate whether the actions of UNCHR are considered effective. Documents that report on presentations or planned presentations before UNCHR will be considered as not relevant. </narr>
</top>
<top> <num> 1009 </num> <title> Frank Gehry architecture </title>
<desc> What opinions have been made by critics and the public about the architecture of Frank Gehry? </desc>
<narr> Relevant documents will praise or criticize structures designed by Frank Gehry. Criticisms and compliments from other architects would be most appropriate but comments by the general public are also relevant. </narr>
</top>
<top> <num> 1010 </num> <title> Picasa </title>
<desc> What do people think about Picasa, the photo management software suite. </desc>

<narr> Find blogs containing opinions about any facet of Picasa. Opinions in advertisements such as those contained in ads by Google, freeware sites etc. are not relevant. </narr>
</top>
<top> <num> 1011 </num> <title> Chipotle Restaurant </title>
<desc> What is the opinion of Chipotle restaurants? </desc>
<narr> Blog posts describe how people feel about Chipotle Restaurants and the reasons they feel that way. How people feel about specific menu items is not relevant, but their opinions of Chipotle burritos in general as well as opinions on Chipotle food in general are relevant. </narr>
</top>
<top> <num> 1012 </num> <title> Ed Norton </title>
<desc> Find opinions on movie actor Ed Norton. </desc>
<narr> Ed Norton's career as a leading movie actor is uncertain. He is mostly known for his supporting roles in movies. Documents should emphasize whether or not people liked/disliked him and/or his movies. Comments, both positive/negative, regarding his capabilities are relevant. </narr>
</top>
<top> <num> 1013 </num> <title> Iceland European Union </title>
<desc> Find opinions on Iceland's relations with the European Union. </desc>
<narr> Documents about Iceland's joining the European Union are relevant. Opinions from non-residents of Iceland, as well as from residents, are all relevant. Documents about European Union relations with other nations are not relevant. </narr>
</top>
<top> <num> 1014 </num> <title> tax break for hybrid automobiles </title>
<desc> Find opinions on the Federal tax break for purchasers of gasoline- electric hybrid automobiles. </desc>
<narr> The Federal program offers graduated (based on gas mileage) incentives to purchasers of new, efficient, mostly hybrid automobiles. Positive and negative opinions on the tax incentive concern changes in the tax incentive from a deduction to a credit, limits on the incentive's application (phasing out after 60,000 units are sold), positive and negative economic impacts, and competitive advantages are all relevant. </narr>
</top>
<top> <num> 1015 </num> <title> Whole Foods wind energy </title>
<desc> Find opinions about Whole Foods' announcement that all its stores would get their electricity from wind power. </desc>
<narr> Relevant documents will respond to the announcement that Whole Foods would be paying a premium on its energy costs to buy wind power in the form of credits to be applied to present and future wind power production. Opinions on Whole Foods' role in encouraging the use of wind power in the United States are also relevant. </narr>
</top>
<top> <num> 1016 </num> <title> Papa John's Pizza </title>
<desc> What do people think about Papa John's Pizza? </desc>
<narr> Opinions about Papa John's Pizza, not limited to food items but including any facet of the chain are relevant. Upper-level business practices, ownership, or stock are not relevant. Blogs consisting solely of advertisements are not relevant. </narr>
</top>
<top> <num> 1017 </num> <title> Mahmoud Ahmadinejad </title>
<desc> What is the opinion of Mahmoud Ahmadinejad, the president of Iran? </desc>
<narr> Mahmoud Ahmadinejad, the president of Iran, has made a number of statements recently, and has promoted a number of programs in Iran that have created controversy. Documents should indicate how Mahmoud Ahmadinejad, including his recent actions or statements, is viewed. Documents discussing the views of other Iranians of Ahmadinejad will be considered relevant. </narr>
</top>
<top> <num> 1018 </num> <title> MythBusters </title>
<desc> Who likes The MythBusters television show on the Discovery Channel, and why? </desc>

<narr> Relevant documents are those blogs which express comments about the MythBusters show on the Discovery Channel. Documents about other mythbusters than this specific television show are not relevant. </narr>
</top>
<top> <num> 1019 </num> <title> China one child law </title>
<desc> Find opinions about China's law that restricts families to only one child. </desc>
<narr> Relevant documents will include opinions about China's one child per family law, and why these opinions are favorable or not. Discussion of abortions of female fetuses is not relevant unless a connection is made to the one-child law. </narr>
</top>
<top> <num> 1020 </num> <title> intelligent design </title>
<desc> Find opinions about "intelligent design", the proposition that the world and its inhabitants did not evolve but instead were created by an entity or force possessing great intelligence. </desc>
<narr> Reasons both for and against the proposition of intelligent design are relevant. Equally important is citing individuals who have taken a position on the subject. Preferably these individuals should be well known and have some claim to credibility with the public. </narr>
</top>
<top> <num> 1021 </num> <title> Sheep and Wool Festival </title>
<desc> This topic seeks views about attending a Sheep and Wool Festival. </desc>
<narr> Relevant documents are those which express personal views about attending any Sheep and Wool Festival. Documents about sheep, wool, or festivals which don't express opinions about attending a Sheep and Wool Festival are not relevant. Also, documents about other kinds of festivals are not relevant. </narr>
</top>
<top> <num> 1022 </num> <title> Subway Sandwiches </title>
<desc> What are peoples' opinions about Subway Sandwiches? </desc>
<narr> Relevant blog posts will describe how people feel about Subway Sandwiches and why they feel that way. How Subway views itself is not relevant, how others perceive them is relevant. Also not relevant is how people feel about specific sandwiches or other Subway sandwich menu items. How people feel about Subway food in general is relevant. </narr>
</top>
<top> <num> 1023 </num> <title> Yojimbo </title>
<desc> Find opinions on Yojimbo information organization software for Mac OS. </desc>
<narr> The release of Yojimbo, a program designed to help capture and classify disparate items found on computers and the World Wide Web, generated opinions from Mac OS user blogs and comments. Relevant opinions are from bloggers and comments that are using or testing Yojimbo. </narr>
</top>
<top> <num> 1024 </num> <title> Zillow </title>
<desc> Find opinions about Zillow, a website that provides estimates of home values. </desc>
<narr> Relevant blogs will contain opinions from people who have used Zillow to produce home value information. Opinions based on multiple visits to zillow and comparisons of resulting differing values for the same property are also relevant. May include posts made by Realtors. Not relevant are general discussions of home and real estate values, the role of realtors, or the role of State or local tax authorities. </narr>
</top>
<top> <num> 1025 </num> <title> Nancy Grace </title>
<desc> Find opinions on Nancy Grace, a former prosecutor who is now a journalist on CNN. </desc>
<narr> Nancy Grace began her TV career on Court TV and later moved on to CNN as a legal analyst. Documents should indicate whether people like or dislike Nancy Grace. The intensity of how people dislike Nancy Grace is relevant. </narr>
</top>
<top> <num> 1026 </num> <title> flag burning </title>

<desc> What positions are held by politicians and the public about criminalizing flag burning in the United States. </desc>
<narr> Relevant documents will contain pro or con positions on the subject of criminalizing burning of the United States flag, preferably with reasons for their position. Comments from public figures and the general public are equally relevant. </narr>
</top>
<top> <num> 1027 </num> <title> NAFTA </title>
<desc> Find views about the North American Free Trade Agreement, known by the acronym NAFTA. </desc>
<narr> Relevant documents will show the feelings of people about how well NAFTA is serving those countries. Documents which discuss views about other free trade agreements and do not discuss the NAFTA experiences per se are not relevant. </narr>
</top>
<top> <num> 1028 </num> <title> Oregon Death with Dignity Act </title>
<desc> What arguments have been made for and against Oregon's Death with Dignity Act, which was ruled as legal by the Supreme Court in February 2006? </desc>
<narr> Relevant documents will name people or organizations which have come out either supporting or opposing the Act. Preferably, these documents will also detail reasons for their positions. </narr>
</top>
<top> <num> 1029 </num> <title> Morgan Freeman </title>
<desc> What do people think about Morgan Freeman? </desc>
<narr> Blog posts or comments containing opinions about Morgan Freeman or his works are relevant. Fan fiction is not relevant. </narr>
</top>
<top> <num> 1030 </num> <title> System of a Down </title>
<desc> What do people think about the metal band System of a Down and their music? </desc>
<narr> Any positive or negative comment about System of a Down, their music, albums, or songs is relevant. Opinions concerning the performers' personal lives or endorsements are not relevant. </narr>
</top>
<top> <num> 1031 </num> <title> Sew Fast Sew Easy </title>
<desc> Find opinions about Sew Fast Sew Easy, an organization that provides materials and services on sewing and knitting. </desc>
<narr> Sew Fast Sew Easy objected to the use of a trademarked statement and forbade its use by local sewing and knitting groups. Documents should indicate whether people like or dislike Sew Fast Sew Easy, and why individuals and groups agree or disagree with the legal tactics used by Sew Fast Sew Easy. Historical disclaimers showing prior use of the effective date of the trademark statement are highly relevant. </narr>
</top>
<top> <num> 1032 </num> <title> I Walk the Line </title>
<desc> What is the opinion of the movie "I Walk the Line"? </desc>
<narr> Relevant blog posts will describe how people feel about the movie "I Walk the Line" and why they feel that way. Relevant opinions include how people feel about the movie, including the performance of the two lead actors. Posts describing award nominations and awards received by the movie are not relevant. </narr>
</top>
<top> <num> 1033 </num> <title> World Bank </title>
<desc> Find opinions about the World Bank, its president, staff or programs. </desc>
<narr> Documents about the World Bank, its president, staff or programs are relevant. Opinions about world monetary policy that that do not specifically mention the World Bank are not relevant. </narr>
</top>
<top> <num> 1034 </num> <title> Ruth Rendell </title>
<desc> What do people think about Ruth Rendell? </desc>

<narr> Any opinion about the author Ruth Rendell or about any of her books is relevant. (It must be evident that a given book is indeed by this author.) The author or a book by her appearing on a book list is not relevant unless there is an accompanying opinion, score, or any other indication of judgment. </narr>
</top>
<top> <num> 1035 </num> <title> Mayo Clinic </title>
<desc> Find opinions about the Mayo Clinic or its medical practice. </desc>
<narr> Relevant documents include opinions on the quality of the Mayo Clinic's medical care, research, programs, personnel or facilities. Opinions on research expressed by Mayo researchers themselves are not relevant. </narr>
</top>
<top> <num> 1036 </num> <title> Project Runway </title>
<desc> Find opinions of the reality show "Project Runway", where designs by clothing designers are judged until there is a single winner. </desc>
<narr> Project Runway is a reality television program with clothing designer contestants of various personalities and egos. Documents should indicate what the viewers liked or disliked about the program and contestants. A basic statement from a viewer that they liked or disliked the program was considered relevant. </narr>
</top>
<top> <num> 1037 </num> <title> New York Philharmonic Orchestra </title>
<desc> Find opinions about the New York Philharmonic Orchestra or its performances. </desc>
<narr> Relevant documents include opinions about the New York Philharmonic in general, its management and operations or its individual performances. Listings of upcoming performances are not relevant. </narr>
</top>
<top> <num> 1038 </num> <title> israeli government </title>
<desc> Find opinions about the government of Israel. </desc>
<narr> Relevant documents will include the feelings of government employees, other citizens, or noted non-Israelis who approve or disapprove of Israeli government policies. Positive or negative statements about individuals in the Israeli government are highly relevant. </narr>
</top>
<top> <num> 1039 </num> <title> The Geek Squad </title>
<desc> What is the opinion of the service provided by the Geek Squad? </desc>
<narr> The Geek Squad is a computer support subsidiary of Best Buy. Documents should indicate whether the service provided by the Geek Squad is viewed as good or poor. Recommendations that someone with a problem should consult Geek Squad will be considered a positive opinion of their services; where statements expressing a significant period of time that Geek Squad worked on their machines will be considered a negative opinion. </narr>
</top>
<top> <num> 1040 </num> <title> TomTom </title>
<desc> What do people think about the TomTom GPS navigation system? </desc>
<narr> How well does the TomTom GPS navigation system meets the needs of its users? Discussion of innovative features of the system, whether designed by the manufacturer or adapted by the users, are relevant. </narr>
</top>
<top> <num> 1041 </num> <title> federal shield law </title>
<desc> What do people think of the proposed Federal Shield Law which would allow journalists to not divulge their sources. </desc>
<narr> Relevant documents will name individual legislators and journalists who are proponents of the law and reasons for their support. Also relevant will be opponents of the proposed law and reasons offered why the law should not be passed. Many bloggers wish to be included in the law and any reasons pro and con for this would be relevant. The First Amendment to the Constitution is critical to the subject and words clarifying this connection would be relevant. </narr>
</top>
<top> <num> 1042 </num> <title> David Irving </title>

<desc> Find opinions on the arrest and trial of historian David Irving in Austria in violation of that country's law against denying the Holocaust. </desc>
<narr> Relevant documents will comment on historian David Irving's arrest and trial. Opinions regarding limits to free speech versus "hate speech," David Irving's motives for traveling to Austria, and inconsistent enforcement of Austria's law on Holocaust denial are all relevant Opinions on whether or not such prosecutions are useful from a strategic standpoint, that is, Calling attention to Holocaust deniers versus letting them disappear into obscurity, are also relevant. </narr>
</top>
<top> <num> 1043 </num> <title> A Million Little Pieces </title>
<desc> Find opinions of James Frey's book "A Million Little Pieces". </desc>
<narr> Relevant blog posts should describe how the people who read James Frey's book "A Million Little Pieces" felt about the book, or how people felt about James Frey after the revelation that his book was not biographical. The opinions of the book and of Frey are relevant. </narr>
</top>
<top> <num> 1044 </num> <title> talk show hosts </title>
<desc> Find opinions of radio and television talk show hosts. </desc>
<narr> Relevant documents will name those talk show hosts who are either liked or disliked by viewers or listeners of their programs. Reasons for why the host is liked or disliked are relevant. </narr>
</top>
<top> <num> 1045 </num> <title> Women on Numb3rs </title>
<desc> Find opinions about the TV show Numb3rs with regard to women. </desc>
<narr> Find opinions about the female characters in the TV show Numb3rs but not about the actresses that play them. I will accept generic comments about women on the show or roles on the show as relating to women. The spelling of the show title as "Numbers" is also acceptable. </narr>
</top>
<top> <num> 1046 </num> <title> universal health care </title>
<desc> Find opinions on whether the United States should provide a universal health care system that offers care to all citizens and residents. </desc>
<narr> Relevant documents will name people and organizations both supportive and opposed to universal health care, as well as the rationales offered for their positions. Relevant comments should be limited only to universal health care in the United States, although comments about the success or failure of such systems in other countries would be appropriate. </narr>
</top>
<top> <num> 1047 </num> <title> Trader Joe's </title>
<desc> What are peoples' opinions about Trader Joe's? </desc>
<narr> Relevant blog posts will describe how people feel about Trader Joe's and why they feel the way they do. How Trader Joe's views itself is not relevant. How people feel about specific Trader Joe's items is not relevant, but how they feel about a category of items or a food department (such as the produce department) is relevant. Peoples' requests or expressions of desire to have a Trader Joe's near their home are not relevant. </narr>
</top>
<top> <num> 1048 </num> <title> Sopranos </title>
<desc> Find opinions about "The Sopranos", a very popular, long running television program. </desc>
<narr> The Sopranos is the story of a mob leader and how he balances his life with his personal family life. Opinions as to whether mob violence should have been glorified on television are relevant. Documents expressing the attraction people had for watching the show are relevant. Positive comments regarding how well the male lead performs his dual life are very relevant. </narr>
</top>
<top> <num> 1049 </num> <title> YouTube </title>
<desc> Find views about the YouTube video-sharing website. </desc>

<narr> The YouTube video-sharing website provides internet users with a relatively new way to share videos. Documents which express views about how well it succeeds in meeting the needs of users are relevant. </narr>
</top>
<top> <num> 1050 </num> <title> George Clooney </title>
<desc> What are peoples' opinions of George Clooney? </desc>
<narr> Relevant blog posts will describe how people feel about George Clooney and the reasons they feel that way. Opinions about his performance in a specific movie are not relevant, but opinions of him as an actor are relevant. Opinions of his personality, strengths, and weaknesses are also relevant. </narr>
</top>