Exploring Topic Modelling in The Domain of Integrated Water Resource Management

AKSHAY KOHLI

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE STUDIES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF ARTS

GRADUATE PROGRAM IN INFORMATION SYSTEMS AND TECHNOLOGY
YORK UNIVERSITY
TORONTO, ONTARIO

June 2021

Abstract

To successfully achieve the United Nations Sustainable Development goals, policy and decision making should include Integrated Environmental Assessment (IEA). Water resources and there utilization play an important role in achieving these goals at all levels from global to local. Sustainability of a water resource is of paramount importance for achieving United Nations' long-term development goals. Sustainability of a resource is governed by the interplay of inner natural processes, biological, economical and social systems, making management of a water resource a complex multidisciplinary problem which can be solved only by combining various approaches. The thesis explored application of text mining techniques, namely, topic modelling, to scientific publications in the sustainable water resource management domain with the goal to identify major research questions, practical problems and methodological approaches used to address these problems. Comparative analysis of approaches to building corpora and model performance evaluations were conducted.

Acknowledgement

Firstly, I would like to thank my supervisor, Dr. Marina Erechtchoukova for constantly motivating and guiding me towards the successful completion of my thesis. Professor Marina helped me with my queries, encouraged me to explore different research areas and constantly communicated with ways on how the work can be enhanced. Under her guidance, I got to learn a lot about Natural Language Processing and text mining. I would also like to thank the committee chair; Dr Aijun An and Dr Matthew Brzozowski for accepting the proposal to supervise the thesis defence and for their valuable time. Lastly, I would like to appreciate the support of my parents and friends without whom I would not have been able to perform to the best of my abilities.

Table of Contents

List of Tables

List of Figures

Key Terms

The following list describes the terminologies that have been consistently used within the body of the thesis:

C_V: Statistical measure to evaluate the quality of topics.

bi-gram: bi-gram is a continuous chain of two words in a document.

BoW: Bag-of-words (BoW) is a text representation scheme, where text is represented by the frequency/occurrence of words within a document.

Corpus: Collection of written texts/documents.

LDA: Latent Dirichlet Allocation

n-gram: n-gram is a continuous chain of n words in a document. *n* represents an integer.

TF-IDF: Term Frequency-Inverse Document Frequency

token/term: Token/term can be referred as a word in a document.

1 Introduction

Natural Language Processing (NLP) is a branch of Artificial Intelligence that automatically processes human language to understand meaning of speech or text and convey the conclusions. One of the research areas of NLP includes semantic/text mining (Figure 1). The main aim of text mining is to extract meaning from unstructured text. Topic Modelling, which is extensively used in NLP, is an automated process which discovers hidden themes, *e.g.* topics, and mine the semantic knowledge from the unstructured text. Topic modelling algorithm is able to capture the latent themes automatically because it works on the linguistic concept which states that the context in language is always represented by words which have some relationship with each other (De Saussure, 2011). For instance, a document is said to have the context of ‘finance’ only when words that are frequently used in financial discourse are present in the document. This shows that a coherent topic derived from the unstructured text, is made up of combination of words which tend to complement each other (Mohr and Bogdanov, 2013). A derived topic can be viewed as a collection of words which appear more frequently when that particular hidden theme is being discussed in the corpus. The topic modelling algorithms are able to record the co-occurrences of words and are immune to the other intricacies of the language. For instance, the narrative, placement or syntax have no effect on the functioning of these algorithms. This happens because the topic modelling algorithms do not handle the unstructured text in its original form but rather convert them into a bag of words (BoW) representation. A document consists of semantically coherent unstructured text and a corpus is a collection of such documents. The topic modelling algorithms analyze the corpus to recognize the pattern of co-occurrence of words and utilize these patterns to come up with an arrangement of words that produce coherent topics. The distribution of topics for each document is also identified in the corpus. This helps the researchers to understand whether the document had one dominating latent theme or multiple latent themes. Topic modelling becomes of paramount importance to a vast expanse of studies in diverse fields ranging from sociology, humanities to technology due to its capability to automatically determine document relevance to a subject of investigation and correct categorization of large collections of documents. The most unique characteristic of topic modelling algorithms is that they are generated automatically and can be used to represent each document via a meaningful cluster or clusters of words. These clusters of words are called topics.

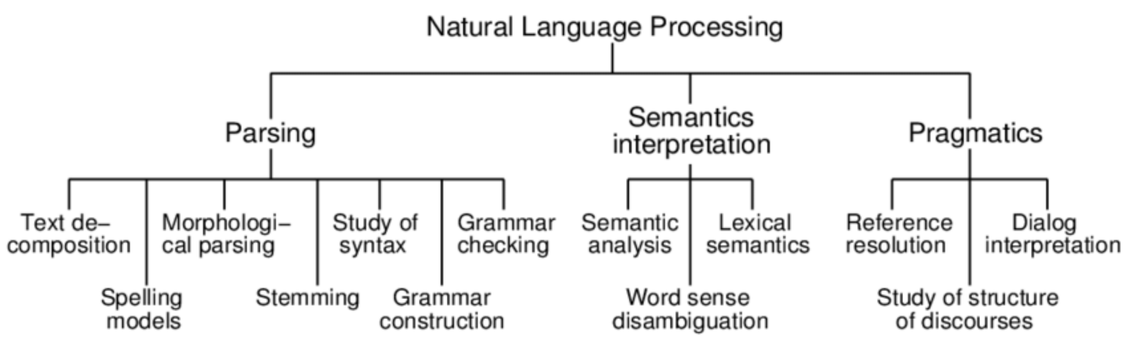


Figure 1: Research Areas in NLP (Lensu, 2002)

Topic modelling algorithms work with minimal human inputs which makes this technique even more beneficial for fields such as humanities, politics or sociology. This eliminates the need to come up with pre-defined concepts or classes for labelling each document in the corpus, which is usually needed while manually annotating each document with a label in the corpus. The researcher has to provide the corpus for topic modelling along with the number of topics to be identified as inputs to these algorithms. The output of topic modelling algorithms is generally a list of topics, where each topic is represented as a set of words and the likelihood of each word appearing in the topic. It also returns the topic distribution per document. The algorithms at times, can fail to produce semantically coherent topics. However, if applied with due diligence, the results are able to capture the latent themes in the corpus (Mohr and Bogdanov, 2013). In general the quality of identified topics depends on the data sets, *e.g.* collections of documents. Therefore, it is expected that domain specific topic modelling should provide more informative results (DiMaggio et al., 2013; Jockers and Mimno, 2013; Ghosh and Guha, 2013; Bonilla and Grimmer, 2013). Application of topic modelling to multidisciplinary areas is even more important because it helps to automate the process of selection of relevant documents and categorize the documents into themes for targeted analysis and knowledge extraction. Sustainable development of human society is one of the complex multidisciplinary issues whose importance cannot be overestimated. To successfully achieve the United Nations Sustainable Development goals, policy and decision making should follow an integrated approach which takes into account environmental, economic, and social aspects of undertakings. Water resources and their utilization play an important role in achieving these goals at all levels from global to local. Sustainability of a water resource is of paramount importance for the long-term development of human society. Sustainability of a water resource is governed by the interplay of inner natural processes;

including hydrological, biological and ecological, economical and social activities, making management of a water resource a complex multidisciplinary problem which can be solved only by combining knowledge and expertise from different problem domains. This makes topic modelling a very promising tool for sustainability assessment. However, use of topic modelling and text mining is relatively new in the domain of integrated environmental assessment and integrated water resource management. Li and Zhao (2015), conducted a bibliometric analysis and frequent keyword analysis on global environmental assessment to identify the patterns, trends and collaboration among the authors. There have been handful of papers (Wang et al., 2010; Jiang et al., 2016; Cheng et al., 2018; Daume et al., 2014; Minx et al., 2017; Marvuglia et al., 2020) which explores various branches of Environmental Sciences using topic modelling. This thesis makes the following contributions:

- Investigation of topic modelling techniques for qualitative analysis of domain-specific scientific publications.
- Modified topic modelling framework incorporating automatic corpus building and filtration.
- A utility was developed which automates the integration of full-text articles in their research.
- Corpus token filtration based on TF-IDF text representation.
- Application of the proposed framework to the multidisciplinary domain of Integrated

Water Resource Management to identify dominating topics and research gaps.

2 Literature Review

2.1 Topic Modelling

Topic modelling can be defined as a method to determine the latent themes in a large pack of documents. The basic idea underlying topic modelling is extracting main themes which can give an overview on the collection of documents. Every theme is characterized by a set of words which are coherent in nature and capable of capturing the context. Probability of presence of a theme can be estimated with the help of topic modelling algorithms. Similarly, probability of a particular word to be included in that theme can also be estimated. Majority of these algorithms are iterative in nature. Therefore, numerous iterations over the corpus are required to identify the specified number of topics, determine words explaining these topics and the correct proportion of topics presented in each document (Muita et al., 2020). The working of topic modelling algorithms is analogous to dimensionality reduction. Rather than describing each document in the n-dimensional space where *n* refers to the length of the vocabulary of the corpus, every document is represented in k-dimensions where *k* represents the number of topics identified by the algorithm. It is assumed that some words that capture the semantics of the latent theme would have high frequency in the document which covers that theme in a considerable amount as compared to other terms in the same document. For instance, a document based on tourism theme, words such as ‘meals’, ‘hotels’, ‘accommodation’ are expected to have high frequency in it. Similarly, a document that talks about transportation will have high frequency of words like ‘car’, ‘bicycle’ or ‘airplane’ in comparison to words like ‘hotel’. Therefore, discovery of such latent themes in multidisciplinary areas is even more important because it helps to automate the process of selection of relevant documents and categorize the documents into themes for targeted analysis and knowledge extraction. There is a considerable number of studies reported in scientific literature that employ topic modelling to extract concise meaning from large textual data. DiMaggio et al. (2013) utilized the Latent Dirichlet Allocation (LDA) algorithm to understand the relationship of US government with the art organizations and museums. The study was based on 8000 documents which were collected over the period of 1986 and 1997. 12 topics were obtained, and the authors also captured the interconnection between topics and integral concepts involved in study of culture like heteroglossia, polysemy *etc.* The study highlighted how topic modelling can be implemented to gauge framing & nuances of meaning. Jockers and Mimno (2013) extracted themes of the novels published in the 19th century. The novels were shortlisted from different regions such as American fiction, Irish fiction and British fiction. The authors analyzed the impact of the external factors by utilizing the obtained topics and their word representation. Ghosh and Guha (2013) applied topic modelling to collected tweets which discussed health, fitness and obesity related issues. After preprocessing of the tweets and representing corpus in BoW format, LDA was applied on the tweets while spatial analysis for prevalence of obesity was conducted using Geographic Information Analysis. This study highlighted how topic modelling could be used to obtain salient themes from short-text data. Bonilla and Grimmer (2013) examined how the alerts issued by the government influence the public response in the light of a possible terrorist attack. They utilized LDA to obtain the topics which were able to capture the responsiveness of the media to such alerts. The topics were derived after applying topic modelling on over 50,000 news articles which were obtained when such alerts were issued. The authors followed basic preprocessing strategies and presented the BoW corpus to the topic modelling algorithm (LDA). The notable finding of the study was that such frequent alerts often affected the economy and resulted in the negative movement of the stock market. These studies capture the importance of techniques such as unsupervised text mining and also accentuate the breadth of applicability of topic modelling. It appeared that, application of topic modelling and text mining to the domain of integrated environmental assessment and, specifically, integrated water resource management is relatively new. Li and Zhao (2015) conducted a bibliometric analysis and frequent keyword analysis on global environmental assessment to identify the patterns, trends and collaboration among the authors. They concluded that the strategic environmental assessment would supersede project environmental impact assessment. They also discovered hot-spots in the domain of environmental assessment, for example the terms ‘climate change’ and ‘biodiversity’ has been drawing attention and this trend will continue in the future as well. (Wang et al., 2010) explored the *Water Research* journal’s publications to perform BoW analysis on the corpus to identify the high frequency words. In research papers published between 1967 and 2008, he observed that ‘activated sludge’ was the top occurring word while ‘adsorption’ was at second and ‘drinking water’ was at third. Jiang et al. (2016) performed topic modelling on papers published in hydropower research area from 1994 to 2013. They observed ‘English’ was the preferred language for publication and the hot-spots of hydropower research were ‘sediment’, ‘climate’, ‘fish’, ‘emission’, ‘lake’, ‘sediment’, ‘Turkey’ *etc.* Since, ‘Turkey’ was the only country that appeared in the high frequency words, the authors concluded that Turkey had been constantly investing in developing hydropower as compared to other countries. The entire knowledge graph of the 1726 articles was captured by a 29-topic model. Cheng et al. (2018) applied topic modelling for the first time to the field of Ecology, Environment and Poverty by considering 4335 articles published in this field between 1981 to 2017. The articles were shortlisted using search terms ‘Ecology and Environment and Poverty (EEP)’. The integrated framework of ecology, environment and poverty was contained by a 9-topic model. This helped the authors to gain better understanding of the research related to EEP nexus and how it can be approached to achieve sustainable development. Daume et al. (2014) wanted to examine the role of participatory forest monitoring in effective surveillance over the development and conservation of the forest. The authors believed that the ecosystem of the forests is constantly threatened due to the social and economic demands of urbanization. Prudent monitoring of the forests can be an efficient way to prevent or detect the ecological alterations that may take place due to the constant exploitation of these resources. However, monitoring vast forested areas is a difficult task and the resources are often limited. Therefore, the authors wanted to determine if citizens can expand the reach of forest monitoring. To construct a framework to quantify and discover the latent themes covered in publications centred around ‘forest monitoring’ and ‘citizen science’, authors performed topic modelling on 1015 publication abstracts. Topic modelling was performed using the Mallet’s implementation of LDA and 100 topics were obtained. The hyper-parameter optimisation was enabled to automatically tune the Dirichlet parameters α and η. They were able to discover the shared topics between the two concepts *i.e.* 'forest monitoring' and 'citizen science'. However, the authors also revealed that the domain specific topics were missing in the obtained topics. The obtained topics were either indicating 'forest monitoring' or 'citizen science'. The authors concluded that topic modelling in general is an efficient and highly scalable technique to perform text mining. At the same time, the whole analysis could have delivered better results if, it was conducted on the full-length publications rather than only abstracts. Minx et al. (2017) explored the publications using scientometric and topic modelling techniques in which the concept of negative emission technologies (NETs) had been discussed. Eliminating the carbon dioxide from the atmosphere is labelled as a negative emission and such negative emission is a crucial driving factor to achieve the long-term climate goals. There had been rapid development in the field of NET which made the qualitative analysis of the field a cumbersome process. Abstracts of 2900 articles published from 1991 to 2016 were extracted from Web of Science (WoS). 19-topic model delivered the best results, because the obtained themes were insightful, relevant and covered a wide array of themes relevant to the domain. The topics revealed primarily three latent themes: 'energy systems', 'forestry' and 'land based methods'. The theme discussing the impact and the sustainability analysis of integrated NET portfolios was missing which highlighted that NETs were still in its early days. The authors concluded with the help of topic modelling techniques that the NETs research was yet to be developed compared to the broader climate change discourse. Marvuglia et al. (2020) conducted the qualitative as well as quantitative analysis of the publications on the topic of ‘urban sustainability’. The qualitative analysis was conducted on 21 papers which captured the obstacles and the variety of topics present in the field of urban sustainability. The quantitative analysis was conducted by performing bibliometric analysis on the publications in the field of urban sustainability, selected from WoS, and several quantitative measures were calculated to evaluate the outcomes of the research. Overall, 2305 documents were shortlisted for the bibliometric analysis and topic modelling. Application of LDA algorithm resulted in a 15 topic model. The authors also displayed the journals with respect to the topic distributions with the aim to help new researchers to find relevant journals for future partnerships. The analysis was able to group the collection of documents into different clusters where each cluster represented a common central theme. With the help of the qualitative and quantitative analysis the authors concluded the high correlation between energy utilization, harmful emissions and fight for the limited resources. Syed and Spruit (2017) explored the domain of Fisheries to discover the latent themes. This study was exhaustive as compared to other studies mentioned above. In addition to identification of latent themes, the authors also examined the effect of the document structure on the quality of the topics. Two datasets were investigated, one dataset consisted of 4417 research articles published in Canadian Journal of Fisheries and Aquatic Sciences from 1996 to 2016 while 15000 articles from 12 journals were included in the second dataset. Both datasets had the abstract as well as full-text. LDA was the chosen topic modelling technique and C_V was the chosen coherence metric. It was observed that dataset with fewer articles showed higher coherence score for full-text articles as compared to only abstracts though this difference didn’t exist with the increase in the size of the dataset. The BoW approach dominated other text representation approaches in the reviewed studies. To summarize, topic models capture the latent themes in the corpus on the basis of word statistics observed in the collection of documents in a mathematical framework (Muita et al., 2020). The topic modelling workflow is shown in Figure 2.

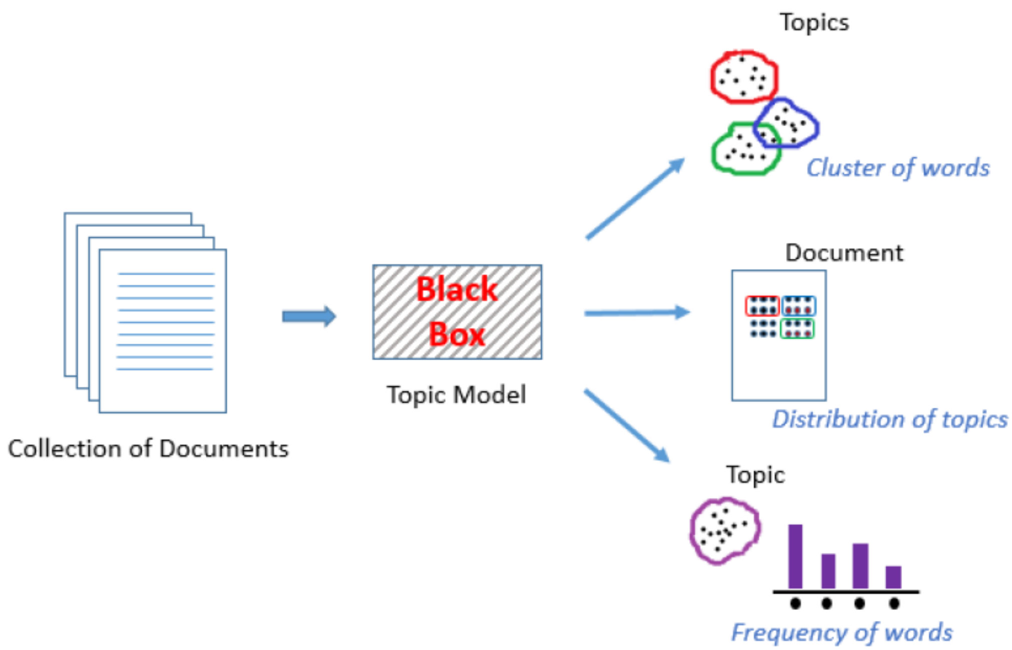


Figure 2: Topic Modelling Process (Joshi, 2020)

2.1.1 Feature Extraction

No text mining algorithm can handle text in its original form. All text mining algorithms require text to be transformed into a numerical format. There are several techniques that can be employed to transform text into numerical representation. These techniques are: BoW approach, term frequency-inverse document frequency (TF-IDF) approach and word-embeddings. Word embedding is a numerical representation scheme where each word is represented as a dense vector. Words with similar meanings tend to have similar vector representations, for e. g ‘man’ and ‘king’. These vector representations are obtained, using deep learning techniques. Traditionally, a text document is represented using BoW technique. The BoW representation is a trivial technique in which the corpus is represented in the form of a sparse matrix. Each row of a matrix represents a unique document in the corpus and each column is a unique word that occurred in the corpus. Each cell either represents the frequency of the occurrence of that particular word or represents the presence or absence of that word. The representation where each cell represents either the absence or presence of the word is also referred to Boolean representation. The BoW conversion ignores all underlying relationships between words in the text which results in information loss. Despite this loss, BoW approach performs sufficiently well in binary classification tasks (Aggarwal, 2018). Text analysis which employ BoW representation give more importance to the terms having high frequency as compared to words having low frequency. This tends to have a negative impact on the accuracy of the classification tasks due to the induced bias for high frequency words. Thus, to ensure equal representation of less frequent terms and also penalize the high frequency words Inverse Document Frequency (IDF) method was proposed by Jones (1988). IDF is used in combination with term frequency (TF) and together they form the term frequency- inverse document (TF-IDF) representation. The corpus is first converted to the BoW format and after that the TF-IDF representation is derived. The mathematical equation that defines the TF-IDF representation is given in Equation 1:

$$W(d,t) = TF(d,t) * log(N/df(t)) \tag{1}$$

where *W(d,t)* represents the TF-IDF weight for a term *t* in document *d*, *N* represents the number of documents in corpus *D*, *d* is an individual document such that $d \in D$, *TF(d,t)* represents the number of times *t* appears in document *d*, *df(t)* represents the quantity of documents having the term *t*.

Narke (2017) experimented with different preprocessing techniques with the aim to identify feature extraction techniques which work better in classification problems. The experiment included BoW approach and the TF-IDF approach. Both approaches were applied to BBC news data and the performance of 10 classifiers was evaluated. The classifiers were compared against evaluation metrics: f-score, accuracy and recall. The results showed that TF-IDF yielded the highest performance compared to BoW approach. Zin et al. (2017) investigated the impact of the pre-processing strategies on performance of classifiers trained on online movie reviews. Feature extraction from the original reviews was done using BoW approach and TF-IDF approach. The effect of BoW representation and TF-IDF representation on classifiers' performance was evaluated using accuracy, precision, recall and F-score. In general, pre-processing steps: stop word removal, dropping illogical words, numbers *etc.* improved the performance of classifiers. In that set of experiments, TF-IDF representation outperformed BoW approach.

2.1.2 Workflow for Topic modelling

Assembling a set of text documents to form a corpus is the first step in topic modelling. Alongside the documents, the number of topics to be extracted from the corpus is also passed as a parameter to the topic modelling algorithm. Let k be the number of topics. The algorithm performs the necessary computations to infer topics as set of representative words. By analyzing the words generated for each topic a user can identify the concept being conveyed in the topic. The model also yields a weight for every word in a topic, this weight is an indication of the importance of the word for that particular topic (Sathi and Ramanujapura, 2016). Examination of the articles shown in Figure 3, reveals that if topic modelling is performed on these articles with the number of topics set to 3, then the latent themes discussed in the articles can be widely categorized as 'Foreign relations', 'Sports' and 'Technology'. If article 'A-4' shown in Figure 3 is observed, it can be deduced that the article is entirely about sports. Likewise, its contribution to the topic of sports would also be high. Meanwhile article 'A-3' is a combination of sports and foreign relations. Thus, its contribution would be equally divided in foreign relations and sports. Article 'A-5' touches the topic of sports, foreign relations and technology in varying degrees as it contains specific keywords for example 'cricket' is an indication for sports, rivalry between 'India' and 'Pakistan' touches upon the foreign relations aspect and lastly 'social media' highlights the technology theme present in the article (Sathi and Ramanujapura, 2016). The examples listed in Figure 3 have high interpretability and therefore, decoding these latent themes is easy. Another example of articles, listed in Figure 4, were used for training a topic modelling algorithm with keeping the number of topics to 3. Analysis of the obtained topics shown in Table 1 showed that the topics had low interpretability.

Figure 3: Topic Modelling Scenario 1 (Sathi and Ramanujapura, 2016)

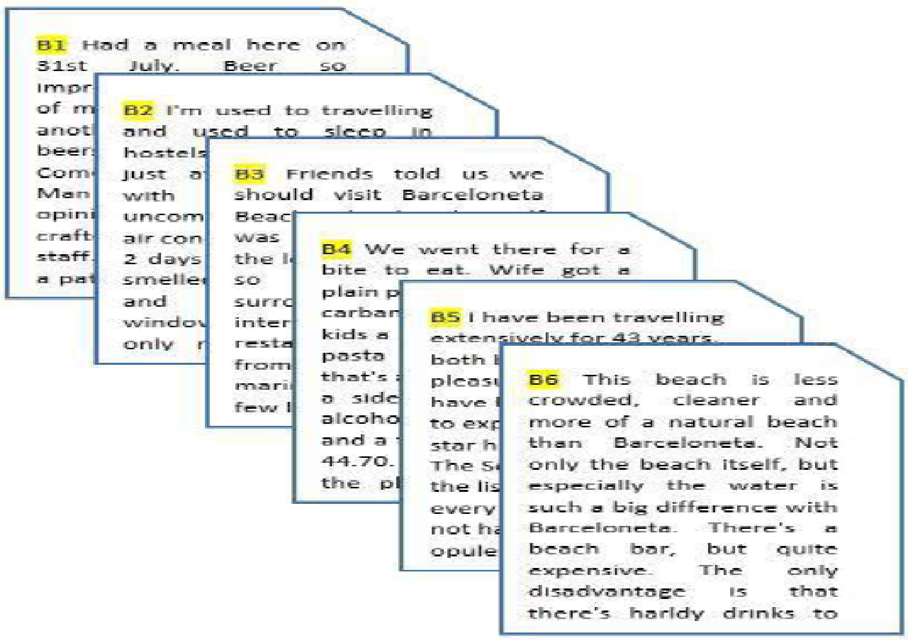


Figure 4: Topic Modelling Scenario 2 (Sathi and Ramanujapura, 2016)

Table 1: Topics with low interpretability (Sathi and Ramanujapura, 2016)

Topic 1	Topic 2	Topic 3
Spaghetti	Stay	Beach
Meal	Make	Walk
Plain	People	Beer
Hot	Cheap	Food
Staff	Senior	Barcelonata

The words ‘spaghetti’, ‘restaurant’, ‘meal’ and ‘staff’ indicates that the cluster is a strong representation for a topic centred around ‘Food’. Topic 2 is ambiguous as it can either be about accommodation in a hotel or it can also refer to a task that was done sincerely by the staff. Topic 3 could be labelled as enjoying on the beach but still it does not give complete clarity. It is worth noting that the word ‘Barcelonata’ does not convey whether it is about ‘Barcelona’ or it is a location on the beach. There is a high probability that the derived topics could have been more descriptive, had the number of topics would have been different from 3. Therefore, the right value of the hyperparameter k which governs the number of topics is essential to obtain topics with high interpretability (Sathi and Ramanujapura, 2016).

2.2 Classification of Topic Modelling Algorithms

The technique of topic modelling can be described as a novel process where a large volume of text is represented in a low dimensional numerical format which can be utilized to find the latent thematic structures or hidden patterns in the text. Reducing high dimensions of the numerical text representations was primarily viewed with an algebraic problem where the initial matrix was decomposed into factor matrix. However, with the advancement in research, an algebraic approach was complemented by the probabilistic one. Thus, topic modelling techniques can be classified into two categories *Non Probabilistic Approach (Algebraic)* and *Probabilistic Approach* depending on the type of dimensionality reduction they perform (Kherwa and Bansal, 2020). Non-Negative Matrix Factorization (NNMF) and Latent Semantic Analysis (LSA) identify the underlying themes by performing dimensionality reduction using algebraic methods (Deerwester et al., 1990; Paatero and Tapper, 1994; Paatero, 1997; Lee and Seung, 1999). NNMF and LSA utilize the BoW methodology in which the text is converted into the term document matrix and the sequence in which the words appeared in the document is neglected and only the term frequency is considered. Probabilistic topic modelling based on the concept of probability considers document as an output of a generative process, in which the parameters of the process can be controlled (Blei, 2012). Topic modelling algorithms can also be derived following supervised or unsupervised approach. Initially the technique of topic modelling had been attributed to unsupervised learning. PLSA and LDA are topic models which help to discover the latent themes in an unsupervised approach. However, with continuous efforts of the researchers LDA could be implemented in a supervised manner (Blei and McAuliffe, 2010). Finally, topic modelling algorithms can be segregated on the basis of whether the sequence of words is preserved during topic modelling or not. Traditionally, LDA has been implemented with using the BoW text representation where the term frequency is considered, but the order of terms is not. However, Wallach (2006) trained a Hierarchical Dirichlet Bi-gram model by preserving the sequence of words using n-gram methodology (Kontoghiorghes, 2005). It was demonstrated that the Hierarchical Dirichlet Bi-gram model outperformed the model trained using BoW approach. Despite the success of considering sequence of words by integrating bi-grams or n-grams while training the model, the BoW approach is generally preferred and yields high quality topics (Kherwa and Bansal, 2020). The diagrammatic representation of classification of topic modelling approaches is shown in Figure 5.

2.3 Different Types of Topic Modelling Algorithm

2.3.1 Latent Semantic Analysis (LSA)/Latent Semantic Indexing (LSI)

LSA is a non-probabilistic technique which uses singular value decomposition (SVD) as the underlying principle. SVD transforms the high dimensions of the text to lower dimensions ensuring that the semantics of the document is preserved (Deerwester et al., 1990). SVD has been regularly applied to mine text for information retrieval. The concept behind LSA is that words that have similar connotations are used in the language with similar contexts. The corpus which is the collection of documents is utilized to obtain the

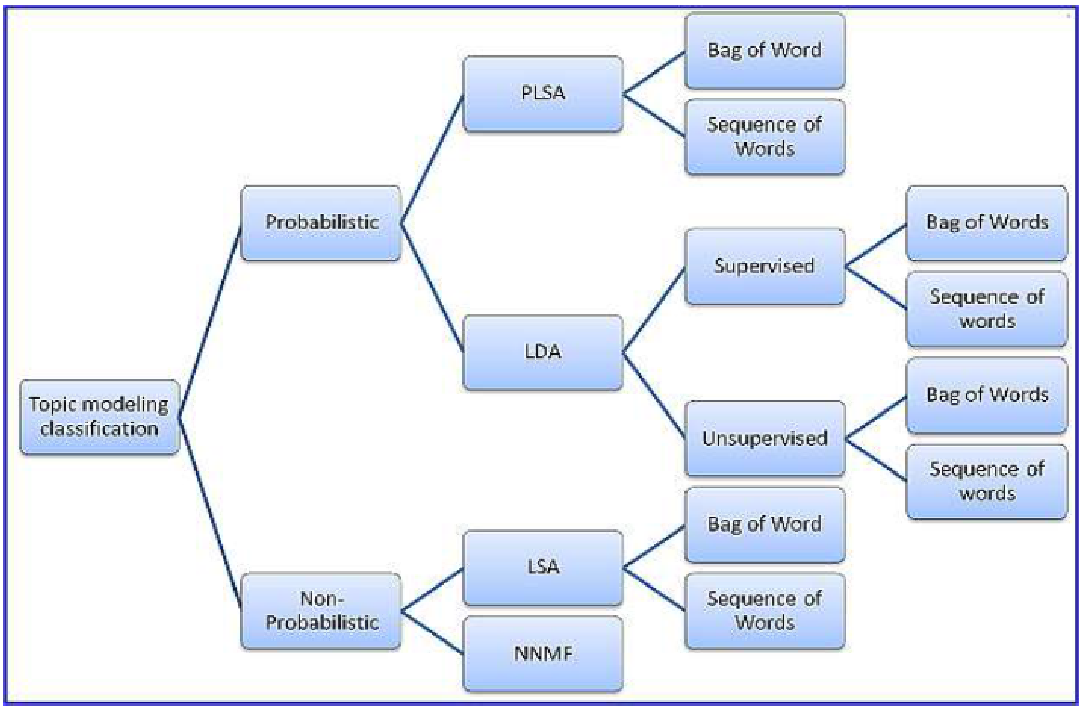


Figure 5: Classification of Topic Models (Kherwa and Bansal, 2020)

context. The corpus is transformed to calculate the closeness between the documents, to figure out the words that have high affinity with each other and finally to represent the collection of documents into clusters grouped on the basis of semantics. LSA has been regularly employed in text mining tasks such as analysis of the social networks, retrieving relevant information and to present the summarized version of the text (Kherwa and Bansal, 2020). Singular Value Decomposition (SVD) transforms a given matrix *A* into a matrix multiplication of three unique matrices as shown in Equation 2:

$$A = U\Sigma V^T \tag{2}$$

where *A* represents the term-document matrix, *U* and *V* are semi-orthogonal matrices, Σ represents a diagonal matrix with singular values on the diagonal and *T* represents the transpose of a matrix

The corpus is firstly represented in the form of a term $\times$ document matrix. The term $\times$ document matrix captures the count of occurrence of each term by each document. The term document matrix *A* has *M* number of rows corresponding to the number of tokens in the corpus and *N* number of columns corresponding to the number of documents on the corpus, as shown in Figure 6. Every single term in matrix *A* is multiplied by a function which preserves the term importance. High frequency terms, that appear in high number of documents, are scaled to lower values while the terms that appear in few documents are associated with a larger weight Buckley et al. (1994). SVD is then applied to the term document matrix *A* which discovers the co-relation that exists between the terms across the corpus.

$$N\left\{\overbrace{\begin{bmatrix} & T_1 & T_2 & \cdots & T_t \\ D_1 & w_{11} & w_{21} & \cdots & w_{t1} \\ D_2 & w_{12} & w_{22} & \cdots & w_{t2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ D_n & w_{1n} & w_{2n} & \cdots & w_{tn} \end{bmatrix}}^{M}\right.$$

Figure 6: Term-Document Matrix (Muñoz, 2021)

$$A = U_k \Sigma_{k \times k} V_k^T \quad (3)$$

Matrices U_k and V_k, shown in Equation 3 are semi-orthogonal matrices. To retrieve the latent themes from the term document matrix *A*, the obtained matrices are then reduced to *k* dimensions as shown in Equation 3 where *k* represents the latent themes in the corpus. This reduction in dimensions removes the noise and also helps to retain the semantic relationship that exists between the documents (Kwon et al., 2017). Kherwa and Bansal (2017) utilized the Singular Value Decomposition to reduce the collection of documents from high dimensions to low dimensions. By using SVD they reduced a collection of research papers on natural language processing into coherent topics where the extracted terms in the topics captured the semantic structure in lower dimensional space.

2.3.2 Non-Negative Matrix Factorization (NNMF)

Non Negative Matrix Factorization approaches dimensionality reduction in a way similar to LSA i.e. the term document matrix is decomposed into product of matrices but it also adds a restriction to the elements of the matrices such that all matrix coefficients are positive. NNMF has been successfully applied on research articles in the environmental domain to extract the latent concepts (Paatero and Tapper, 1994). The rationale behind putting a constraint on negative values is driven by the notion that negative values in many applications confutes the real life values. The term document matrix *A* has *M* number of rows corresponding to the number of tokens in the corpus and *N* number of columns corresponding to the number of documents on the corpus. The term document matrix *A* is transformed into positive vectors as shown in Equation 4:

$$A = VH + C; V \geq 0, H \geq 0 \tag{4}$$

where $V = [v_{mk}] \in R^{M \times K}$ and $H = [v_{kn}] \in R^{K \times N}$ are two non negative matrices and C is the matrix of residuals. The disintegration of *A* in its corresponding factors is not accurate but rather an approximation (Kherwa and Bansal, 2020). NNMF is also referred as Positive Matrix Factorization (PMF). PMF technique is capable of handling sparse data in numerous applications (Cai et al., 2008; Kreutz-Delgado et al., 2003). It was observed that the accuracy improved when clustering was performed in combination with NNMF as compared to clustering alone (Bryan et al., 2006). The versatility and flexibility of NNMF makes it a good choice in many industrial use cases such as computer vision, language models and pattern recognition (Kherwa and Bansal, 2020)

2.3.3 Probabilistic Latent Semantic Analysis

Probabilistic Latent Semantic Analysis (pLSA) is predominantly used in information retrieval, text mining *etc.* pLSA preserves the semantics in the text by modelling the co-occurrence statistics under a probabilistic framework using independent multinomial distributions. It can be seen that both LSA and pLSA works on similar grounds by transforming a document term matrix into product of several matrices. However, the method adopted to reduce number of dimensions is different. LSA performs factorization by implementing Singular Value Decomposition which is entirely based on algebraic rules whereas pLSA adopts a probabilistic route to factorize the matrix (Hofmann, 2013). The probability of obtaining a word for a document *d* is given by Equation 5:

$$P(d,w) = P(d)\sum_{z}\theta_d(z)\tau_z(w) \tag{5}$$

The three functions which govern pLSA are τ_z, θ, ρ. τ_z represents a probability distribution of words over the hidden themes *z* given by $p(w|z)$, θ_d gives the probability distribution of topics in document *d* and $P(d)$ represents the probability of choosing a document *d* (Kherwa and Bansal, 2020).

2.3.4 Latent Dirichlet Allocation

LDA is a topic modelling technique which is probabilistic in nature. It is a generative process which identifies the semantically coherent latent themes underlying the corpus *D*. Each document in corpus *D* is assumed to have different proportion of each topic. LDA can be explained as a theoretical random procedure established on probabilistic sampling rules capable of generating the documents present in the corpus. In LDA the only observed variable are words in the corpus so LDA algorithm has to learn the hidden patterns such as a semantic structure of the topics and their contribution to each document. This learning takes place through different inference techniques. The LDA algorithm implements an inference procedure by determining how likely it is for a document to be associated with a set of topics. With a sufficient number of iterations over the corpus, the LDA algorithm is able to determine the posterior distribution which determines how the collection of documents, *i.e.* the corpus, can be represented (Syed and Spruit, 2017). Firstly, let the number of topics range from 1 to *K*. Here *K* is provided as an input to the

LDA algorithm. The vocabulary V is the collection of unique words present in the corpus. For every topic k, a distribution β_k is sampled over the vocabulary V such that $\beta_k \sim \text{Dir}(\eta)$ where η is the topic hyperparameter. For each document d in the corpus, topic distribution θ_d is sampled from the Dirichlet distribution such that $\theta_d \sim \text{Dir}(\alpha)$ where α is the parameter of the Dirichlet function. For every word $w_{d,n}$ in the document, topic $z_{d,n}$ is sampled from the multinomial distribution over θ_d. Then for the sampled topic, a word $w_{d,n}$ is sampled from the multinomial distribution over $\beta_{z_{d,n}}$. $w_{d,n} \in \{1,. . . ,V\}$. The plate notation for LDA is shown in Figure 7.

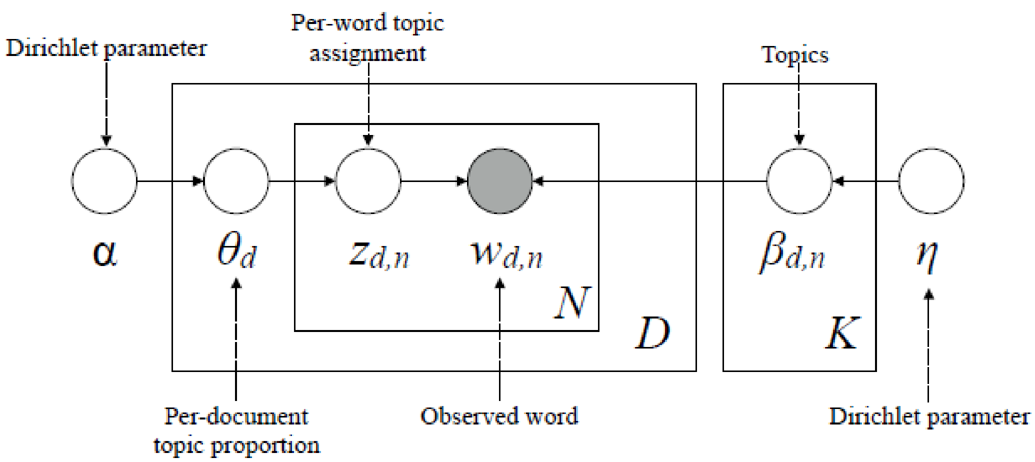


Figure 7: Representation of LDA (Syed and Spruit, 2017)

It is assumed that each document d is made up of different proportions of topics. Both α and η are the hyperparameters in the model and controls the shape of the distribution. α is responsible for the smoothening of topics while η controls the distribution of words within the topics. The distributions β_k and θ_d are approximations of the probability distributions evaluated based on the analyzed texts, this is typically done using statistical inference techniques. Therefore, this whole process can be referred as an inversion of the generative process Syed and Spruit (2017). The aim of LDA is to calculate the conditional probability also known as the posterior distribution, or posterior shown in Equation 6:

$$p(\beta_K, \theta_D, z_D | w_D) = p(\beta_K, \theta_D, z_D, w_D) / p(w_D) \tag{6}$$

The value of the posterior can not be calculated due to the high computations involved but can be approximated using statistical inference techniques (Blei, 2012). Primarily there are two types of inference procedures: (1) sampling based algorithms (Newman et al., 2007) and variational-based algorithms (Blei et al., 2006). Both approaches result in similar performance and are widely used in the available libraries for topic modelling.

2.3.5 Evaluation of Topic Coherence for LDA

Once the topics are obtained using the LDA algorithm, the next question is how semantically coherent the derived topics are? Every topic represents a multinomial distribution over the vocabulary *V*. Each word in the corpus is associated with each topic *k* such that $k \in (1, .., K)$ topics, however, the probability of the word to occur in each topic is different which ensures that the topics are differentiating in nature and are able to capture the latent themes present in the corpus. For every topic distribution the set of words having high probability have a high tendency to co-occur. These sets of high probability words where the size of the set of words usually ranges from 10 to 15 are used to extract the hidden meaning and associate a label that reflects the context associated with those sets of high probability words. An important thing to note is the effect of changing K, *i.e.* the number of topics. A very small value of K might result in topics that are too broad and the LDA algorithm is likely to fail to convey the correct information. If the number of topics are too high, then it often yields topics which are difficult to understand. Hence, selecting the optimum value of K is essential for evaluating the performance of LDA algorithm (Syed and Spruit, 2017). There are several ways to evaluate the quality of the generated topics. One of them is the predictive likelihood of held-out data also referred as perplexity (Wallach et al., 2009). This measure of coherence is often associated with low correlation with human ranked topics and at times negatively correlated too (Chang et al., 2009). As a result, the LDA algorithm with low perplexity score for the held out data might not be relevant to the real life applications due to negative correlation with human interpretation of topics. Aletras and Stevenson (2013); Newman et al. (2010) proposed another topic coherence methodology that is able to quantify the coherence of generated topics. Coherence score is used to report the quality of the topics obtained by the LDA algorithms. Higher the coherence score, more is the observed co-occurrence of the words in the corpus. The foundation of the approach is laid on the distributional hypothesis of linguistics which states that words with almost identical meanings often exist in identical contexts (Harris, 1954). The topic is said to be coherent if majority of the words or top N words that form the topic have high correlation with each other in the corpus. Röder et al. (2015) conducted extensive experiments on topic coherence and evaluated the measures with respect to the human ranked data. The aim of the study was to identify the coherence measure which has the highest correlation with human ranked topics for various datasets belonging to different domains. The authors compared the existing coherence measures following the framework shown in Figure 8.

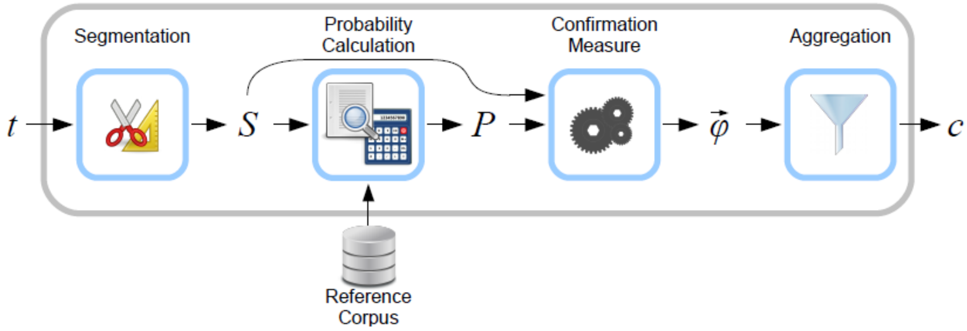


Figure 8: General Framework to Calculate Coherence Measure (Röder et al., 2015)

The coherence measure C_{UCI} is given by Equation 7. C_{UCI} considers K top words of a topic and calculates the average of the pointwise mutual information (PMI) over each of the word pairs. The value of K usually ranges from 5 to 20. PMI for a word pair can be calculated by Equation 8. Probability for a single word $P(w_i)$ or the joint probability of a word pair $P(w_i, w_j)$ is calculated by dividing the number of documents in which the word or the word pair occurred by the total number of documents in the corpus.

$$C_{UCI} = \frac{2}{K \times (K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} PMI(w_i, w_j) \tag{7}$$

$$PMI(w_i, w_j) = \log \frac{P(w_i, w_j) + \varepsilon}{P(w_i) \cdot P(w_j)} \tag{8}$$

Coherence measure C_{UMass} is given by Equation 9. C_{UMass} considers K top words of a topic and calculates the average of the smoothed conditional probability between the top word pairs of the topic (Röder et al., 2015). Coherence measure C_{UMass} also considers the word order present in the top K words for a topic. Word probabilities $P(w_i), P(w_j)$ in Equation 8 and Equation 9 are estimated based on document frequencies of the original documents used for learning the topics.

$$C_{UMass} = \frac{2}{K \times (K-1)} \sum_{i=2}^{K} \sum_{j=1}^{i-1} \log \frac{P(w_i, w_j) + \varepsilon}{P(w_j)} \tag{9}$$

Röder et al. (2015), proposed another coherence metric C_V which outperformed the existing coherence measures mentioned in Equation 7 and Equation 9. C_V had the highest co-relation with the manually labelled topics. The value of C_V can be calculated by following the framework as shown in Figure 8. A series of steps are followed to calculate C_V (Syed and Spruit, 2017). Suppose W be the set of the topic's most likely top K words such that $W = \{w_1, w_2 ..., w_k\}$. The first step is segmentation which segments the entire set W into pairs of word sets W' and W^*. Let S_i represent a pair, where word set W' is paired with word set W^* such that $W' \in W$ and $W^* = W$. Let S be the entire set of each of the pair S_i such that $S = \{(W', W^*) | W' = \{w_i\}; w_i \in W; W^* = W\}$. For instance, let $W = \{w_1, w_2, w_3\}$, then as per the segmentation technique, one of the segmented pair S_i is given by $S_i = \big((W' = w_1), (W^* = w_1, w_2, w_3)\big)$ (Syed and Spruit, 2017). The second step involves the calculation of probability of the individual word $P(w_i)$ and joint probability of a word pair $P(w_i, w_j)$. $P(w_i)$ or $P(w_i, w_j)$ is calculated by dividing the number of documents in which the word or the word pair occurred by the total number of documents in the corpus. As this calculation ignores whether the pair occurred in close vicinity to each other in the document, the concept of sliding window was introduced to overcome this issue. The sliding window divides the document into numerous virtual documents depending on the size of the window. The probability of the occurrence of the words or word pairs is then calculated using these virtual documents and as a result this calculation estimates how frequently the words occur in close proximity. A confirmation measure ϕ is calculated to observe the semantic similarity of the pair $S_i = \left(W^{'}, W^{*}\right)$. This confirmation measure gives an indication of how firmly W^* supports W'. There are two forms of confirmation *i.e.* direct and indirect (Röder et al., 2015). C_{UCI} and C_{UMass} are an example of direct confirmation measure whereas C_V is an example of indirect confirmation. In direct confirmation measures the similarity is only observed between the word sets. In indirect confirmation measure, the word-set W' and W^* are represented as context vectors. Each context vector captures the semantic similarity of the word-set with all the other words in W. Let $\vec{u}(W')$ shown in Equation 10, be the context vector that is formed by pairing W' with all the other words in W. Similarly, let $\vec{w}(W^*)$ shown in Equation 11 be the context vector that is formed by pairing W^* with all the other words in W. Each element of the context vector represents the normalized pointwise mutual information (NPMI) calculated between the individual words w_i and w_j. NPMI is given by Equation 12:

$$\vec{u}\left(W^{'}\right) = \left\{ \sum_{w_i \in W'} NPMI\left(w_i, w_j\right)^{\gamma} \right\}_{j=1,\ldots,|W|} \tag{10}$$

$$\vec{w}\left(W^{*}\right) = \left\{ \sum_{w_i \in W^{*}} NPMI\left(w_i, w_j\right)^{\gamma} \right\}_{j=1,\ldots,|W|} \tag{11}$$

$$NPMI\left(w_i, w_j\right)^{\gamma} = \left(\frac{\log \frac{P\left(w_i, w_j\right) + \varepsilon}{P(w_i).P(w_j)}}{-\log\left(P(w_i, w_j) + \varepsilon\right)} \right)^{\gamma} \tag{12}$$

In Equation 12, ε ensures that logarithm of zero is prevented and γ is used to give more weight to larger NPMI values. NPMI tends to outperform pointwise mutual information (PMI) with regards to the correlation to manually labelled topics (Bouma, 2009). The confirmation measure ϕ for the pair S_i is evaluated using the cosine similarity between the context vectors $\vec{u}$ and $\vec{w}$, as shown in Equation 13:

$$\phi_{S_i}(\vec{u},\vec{w}) = \frac{\sum_{i=1}^{|W|} u_i.w_i}{||\vec{u}||_2.||\vec{w}||_2} \tag{13}$$

Finally, aggregation is performed by combining the confirmation values obtained for each pair. The aggregation can be done by computing the mean, mode, median *etc.* For C_V, generally the arithmetic mean is calculated.

2.4 Research Questions Formulation

The conducted review of scientific publications on topic modelling techniques and their applications allowed to determine popular topic modelling techniques and research gaps in domain specific studies providing a road map for further investigations. Majority of the studies implementing topic modelling of papers from a problem domain used only abstracts of the papers. Given that the quality of identified latent topics depends on quality of the corpus used in the analysis, it is necessary to investigate the effect of full-text of articles on the depth and breadth of the identified topics. Majority of the studies only used basic preprocessing steps. Some domain specific words along with the stopwords would be removed and the remaining corpus would be transformed to their BoW representation.

Comparative analysis of topic modelling techniques showed that in many cases LDA approach outperformed other modelling tools in various problems domains. Since, LDA is a generative tool and works on the frequency of the terms, the assumption that the non-descriptive terms can have high frequency in the corpus and lead to degrading the quality of the topics is reasonable and should be investigated further. It is necessary to examine the effect of TF-IDF filtration on the quality of the topics. Therefore, the thesis aims to investigate how the quality of the corpus affects the quality of identified topics. There is no study which qualitatively or quantitatively analyses scientific publications in the domain of Integrated Water Resource Management (IWRM) using topic modelling. The investigation was conducted on IWRM publications with the aim to discover dominating topics, research trends and possible gaps in the area of IWRM.

3 Methodology

The commonly accepted framework for topic modelling can be represented by a 7-step process: identification of problem domain, building corpus, extracting text, preprocessing the corpus, applying topic models, visualizing the obtained topics and finally analyzing them, as shown in Figure 9.

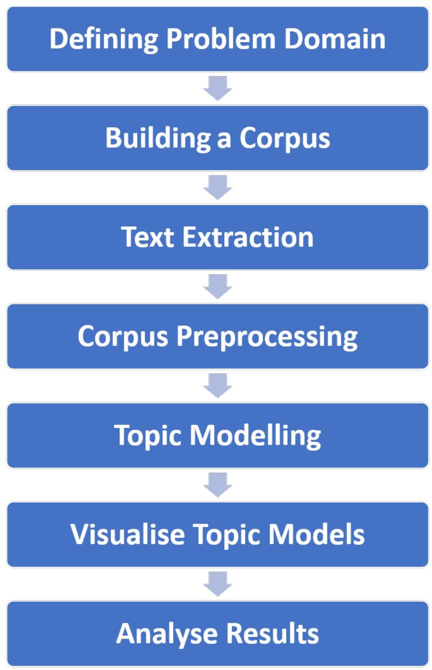


Figure 9: Conventional Framework for Topic Modelling

Traditionally, this framework is used to analyze short text documents, *e.g.* online tweets or brief reviews. Rarely, it is applied to process abstracts of research publications which length does not exceed 500 words. It can be explained by technical limitations of standard computing and the necessity to process matrices of high dimensionality. Increasing size of a text corpus causes the increase in dimensionality of the resulting matrices. One of the goals of the study includes the analysis of full-length journal articles. Therefore, dimensionality reduction is also included into consideration. The modified framework as shown in Figure 10 is further explained in the subsequent section and it can be utilized for future studies employing topic modelling.

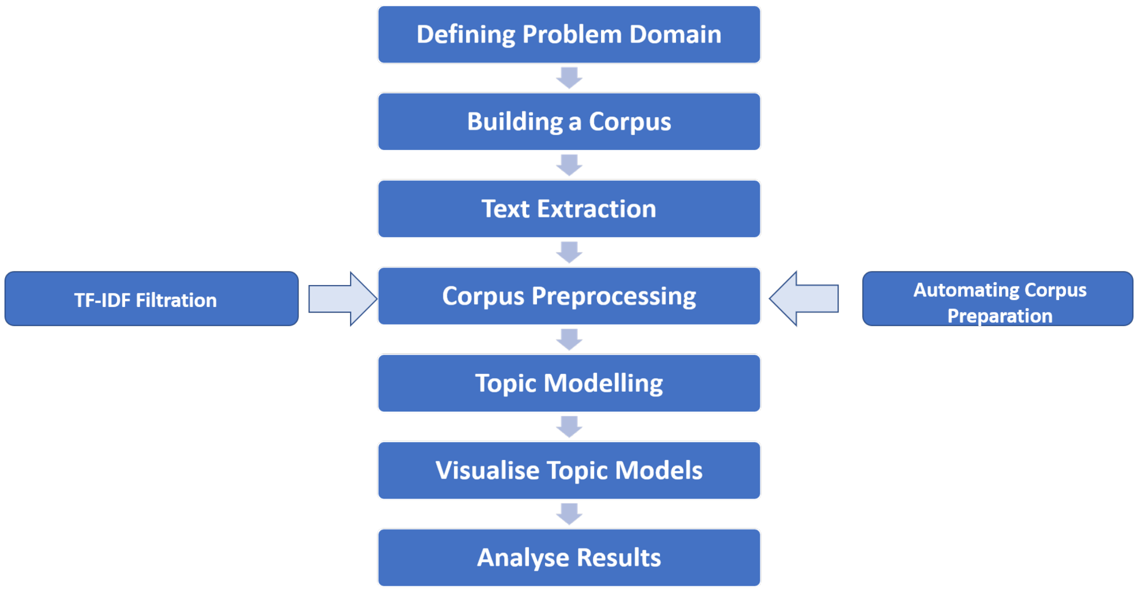


Figure 10: Modified Framework for Topic Modelling

3.1 Defining a Problem Domain

Identification of a problem domain is the first step of the framework. The selected domain must be sufficiently diverse to leverage the power of topic modelling algorithms to uncover domain specific aspects. Well-defined research questions are essential for a high-quality research and help the authors to explore the problem under consideration at a higher granularity level.

3.2 Building Corpus

3.2.1 Sources of Textual Data

The data governs the depth of the topics which the topic modelling algorithms would be able to derive. If the source of data does not cover the domain in a holistic manner, there are chances of information loss.

3.2.2 Keywords and Journal Selection

In order to fetch the research papers about the problem under investigation, it is important to ensure only relevant documents are used for the process of topic modelling. Filtering text beforehand becomes a crucial step, which can influence the quality of the topics obtained. The selection of the research papers about a specific domain should be done under the guidance of domain experts. Random selection of research papers will often yield low quality topics and would defeat the research purpose.

3.2.3 Time frame

Time frame is an essential component of research as different themes emerge and become dominating asynchronously. Hence, if the topic modelling framework is applied to a collection of research papers, substantial timeframe must be taken into consideration to observe major developments in the research domain under study.

3.3 Text Extraction

With the successful completion of the previous steps, the required documents can be selected from scientific journals available online. If the researcher is interested in performing bibliometric analysis along with topic modelling, the metadata about the publications must also be downloaded. The metadata contains important information about the article such as author name, year of publication, type of document, journal in which the article was published *etc.* If the analysis requires full-length research papers, then collection of those full-text papers must be done either by manually downloading them or automating the collection process.

3.4 Corpus Preprocessing

3.4.1 Tokenization

The process of segmenting text into distinct words is called tokenization. Textual data is made up of continuous strings of characters whereas any text mining process is based on words. Thus, the first step of text preprocessing is splitting the text into tokens. There are various open source tokenizers available for this purpose (Ramasubramanian and Ramya, 2013)

3.4.2 Removal of Stopwords

It has been observed that high frequency terms such as - ‘the’, ‘a’, ‘an’ *etc.* do not contain any relevant information. Thus, removing such words usually helps in cleaning the data as well as reducing the size of the vocabulary that forms the corpus. It is a common practice to integrate the bi-grams into the corpus after removing the stopwords.

3.4.3 Stemming and Lemmatization

Stemming is a text preprocessing technique that helps in preparing the data for tasks such as information retrieval, topic modelling *etc.* Stemming reduces the size of the corpus by replacing different morphological forms (nouns, verbs, adjective *etc.*) to its root form without considering the part of speech and the semantics of the word in the sentence (Jivani et al., 2011). For example 'studies' and 'studying' would be reduced to its root form *i.e.* 'studi' (Figure 11). Stemmers perform the task of stemming and the most widely used stemmer for English language is Porter's Stemmer (Singh and Gupta, 2016). Lemmatization is also similar to stemming, however, the process of conversion is different. Lemmatization reduces the word to its root form after taking into consideration the context and the part of speech of the word in the sentence (Jivani et al., 2011). For example the words 'studies' and 'studying' would be reduced to 'study'. The step wise process of corpus preprocessing is displayed in Figure 12

	original word	stemmed
0	studies	studi
1	studying	studi

	original word	lemmatized
0	studies	study
1	studying	study

Figure 11: Stemming vs Lemmatization

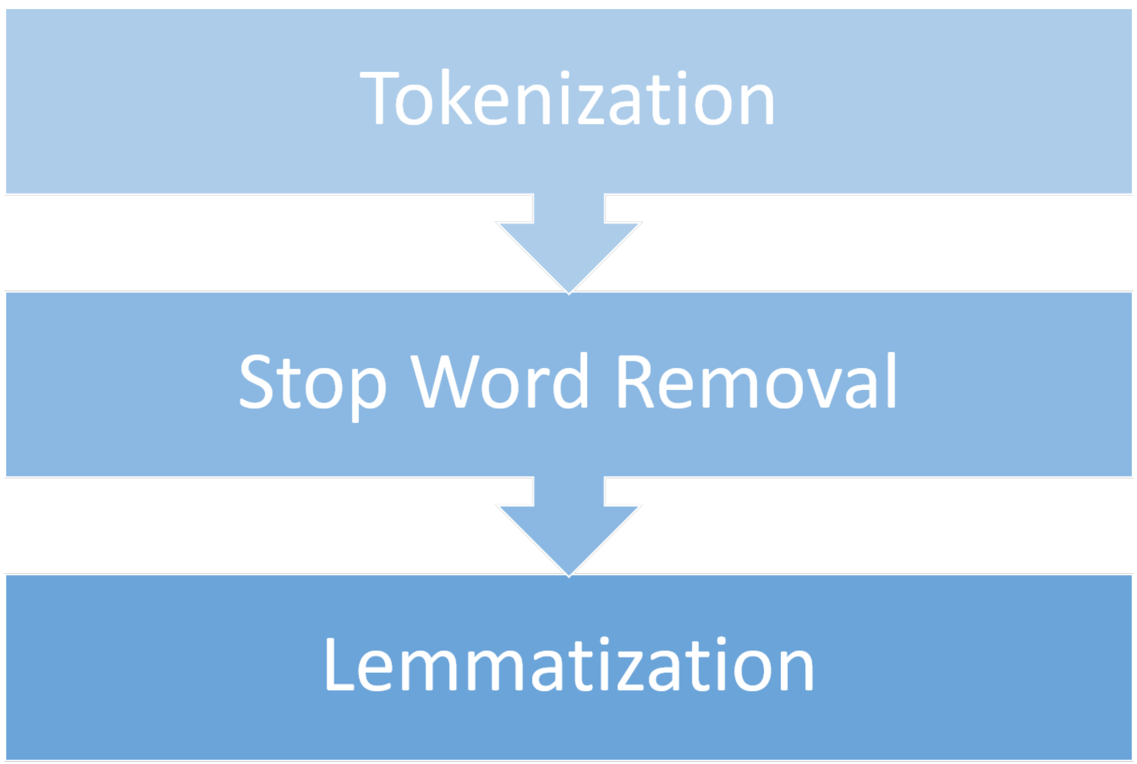


Figure 12: Steps for Corpus Preprocessing

3.5 Integrating TF-IDF Filtration

As observed in majority of the studies of topic modelling, authors converted the text into BoW vectors. During conversion either authors considered the entire corpus or removed the tokens that were present in more than 90$\%$ of the documents and tokens that only occurred once. This type of filtration often leaves the tokens that don't add much value to the document and often act as a source of noise. Therefore, to ensure that only important words are retained in the corpus we investigate integration of the TF-IDF approach in the preprocessing steps. Firstly, the BoW representation of the corpus is converted to TF-IDF representation and then a threshold score (ω) is selected. This threshold score can be treated as a hyperparameter that can be tuned along with the topic modelling algorithm to obtain high quality topics. In TF-IDF filtration, all the words in a document that have a TF-IDF score less than the threshold score are removed from the corpus and in this way only important terms for the document are retained.

3.6 Topic Modelling

The prepared corpus can then be used to conduct topic modelling to discover the latent topics.

3.6.1 Model Selection

There are several topic modelling techniques that can be applied to the unstructured data, however, the underlying principle in all of them remains the same: to break a document into topics and further represent those topics with the help of keywords. Latent Semantic Indexing (LSI) (Deerwester et al., 1990), Mixture of uni-grams model (Nigam et al., 2000), Probabilistic Latent Semantic Indexing (pLSI) (Hofmann, 1999) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003) are some of the commonly used topic modelling techniques. These techniques have been extensively studied by the machine learning community. However, as discussed in Section 2.3, LDA tends to deliver high quality topics and at the same time it is highly efficient and scalable. Hence, LDA topic modelling technique is generally preferred over other topic modelling algorithms. Implementation of LDA can be done using various packages. One of the most commonly used package for implementing topic modelling algorithms is Gensim (Řehůřek and Sojka, 2010). It provides the implementations of various topic modelling algorithms such as LDA, LSA and Random Projections. Another package that gives highly reliable results is Mallet (McCallum, 2002). Mallet was originally written in JAVA, however, Gensim provides a wrapper for implementation of Mallet LDA in Python. The major difference between Gensim LDA and Gensim Mallet LDA is that the former one implements online variational Bayes algorithm (Hoffman et al., 2010) while Mallet LDA implements Gibbs sampling algorithm for topic modelling (Yao et al., 2009). Variational Bayes sampling method is faster while Gibbs Sampling is more precise. Therefore, for analysis where more emphasis would be on the quality of the topics then Gensim LDA mallet should be preferred and if time is preferred over quality, then Gensim LDA should be preferred. The majority of the hyperparameters used to control the performance of the model are common in Gensim LDA and Gensim Mallet LDA. *num_topics* parameter is set to obtain the specified number of topics from the

LDA algorithm. In majority of the document categorization tasks, the number of labels which could be used to label the entire corpus is not known beforehand and generally involves experimentation with the *num_topics* hyperparameter to obtain satisfactory results (Řehůřek and Sojka, 2010). Hyperparameter η, governs the keyword distribution for a given topic. Higher value of η results in topic distributions containing majority of the words which makes tough to distinguish between different topics whereas lower value of η leads to keyword distributions having less words and separation between topics is high (Řehůřek and Sojka, 2010). Likewise η, α is another hyperparameter which governs the document-topic distribution. Higher value of α shows that a document is composed of most of the topics while lower value of α highlights that less number of topics are present in the documents (Řehůřek and Sojka, 2010).

3.6.2 Selection of Performance Measures

Gensim (Řehůřek and Sojka, 2010) provides the functionality to determine the quality of the topics obtained by the LDA algorithm. The Gensim package allows us to compute C_{UMass}, C_V, C_{UCI} and C_{NPMI} coherence measures respectively. For every value of *num_topics* provided to LDA algorithm, a coherence measure is calculated. The *num_topics* at which the LDA algorithm obtains the highest coherence score is generally chosen for further downstream tasks such as categorizing, information retrieval *etc.*

3.7 Visualization of Topics

The obtained topics can be visualized using LDAvis (Sievert and Shirley, 2014). The visualization further helps in understanding the derived topics.

3.8 Assessing the Results

After determination of the suitable number of topics and observing the topics visualized using LDAvis, suitable labels can be attributed to the obtained topics and these topics can be used for further downstream tasks such as information retrieval, text mining *etc.*

4 Technical Limitations

The majority of studies on domain specific topic modelling were conducted on corpora built from the abstracts of relevant research papers (Syed and Spruit, 2017). There are several reasons for that. Firstly, there is limited availability of full-text papers online. Many authors have limited access to the research portals such as Scopus, Web of Science (WoS) *etc.* which forbid them to access the full-text without subscription. Secondly, feasibility of the required analysis with limited capacity of available computational resources to handle full-text articles is another limiting factor. Lastly, very limited automation of the document extraction process is the major reason which hinders utilization of full-text articles. The authors have to go through a series of steps including web scraping, downloading the articles, converting the files from PDF format into text format which is then followed by the conventional pre-processing steps. In order to eliminate these roadblocks that come in the way of handling full-text articles, we created an interactive utility by using Python language. The utility is capable of automating the entire corpus pre-processing provided that a user has access to the full-text publications via personal or organizational subscriptions. Handling full-text articles is a complex process. To begin with, full-text link of the research paper needs to be available to automate the text extraction process. Fetching the research papers using the title of the research paper might not return the most updated revision of the document. Therefore, identification of the full-text link of the research paper is a non trivial process. Firstly, the Digital Object Identifier (DOI) (InternationalDOIFoundation, 2021) of the research paper has to be extracted using the Scopus website or the Elsevier

API. The DOI is the unique identification ID associated with a research paper and it points to the most updated version of the research paper. After identification of the full-text link, one approach was to use the requests module (Reitz, 2021) in Python to obtain the contents of the web page and then extract the text from the web page. The drawback of this approach was that the web-page of certain research articles did not contain the full-text and only contained abstract of the article as shown in Figure 13. Due to this non uniformity in the contents of the web-page of the research articles this approach was discarded.


Papers | Full Access

Methods of confining oil in unlined caverns in aquifers: Water table maintenance by well recharge

Roger Thunvik, Carol Braester

First published: February 1981 | https://doi.org/10.1029/WR017i001p00228 | Citations: 3

PDF TOOLS SHARE

Abstract

One of the methods of petroleum storage is using unlined caverns located below the water table of an aquifer. The water pressure will exceed the oil pressure in the cavern, water will penetrate the cavern, and the oil will be confined. In order to prevent excessive lowering of the water table and to maintain a prescribed constant water level it is necessary, in most cases, to inject water into the aquifer. The influence of rock properties and of flow geometry on the rates of water flow into the cavern and on the shape of the phreatic surface were studied. The equations of flow were solved by a numerical method, and the results were presented in a dimensionless form.

Citing Literature

Figure 13: Only Abstract Visible

The other approach was to download the PDF version of the research article and then apply the preprocessing steps to the extracted text. The CrossRef API (Wilkinson, 2020), was utilized to extract the full-text PDF links. The member publications are the journals that have collaborated with the Crossref API to share the full-text PDF link of the article and non-member publications are the journals which have not registered with the Crossref API. The full-text PDF links for these non member publications would not be available via the CrossRef API. Therefore, separate utility was coded to obtain full-text of member and non-member publications. The workflow of both the utilities is discussed in the following subsections.

4.1 Utility for Member Publications

The GUI will allow a researcher to prepare the corpus using full-text articles provided the researcher has access to full-text articles. In order to make use of the GUI, the user first has to obtain an API key. The API key will enable the user to query the Scopus database and fetch the results. The API key can be obtained by registering at `https://dev.elsevier.com/apikey/create`. The user is prompted to either enter the API key or upload the API key JSON file. The API key is verified (Figure 14). If the API key is incorrect an error window is displayed and the user can enter the API key once again as shown in Figure 15. If the API key is configured successfully then the user can proceed with the next step to enter the queries.

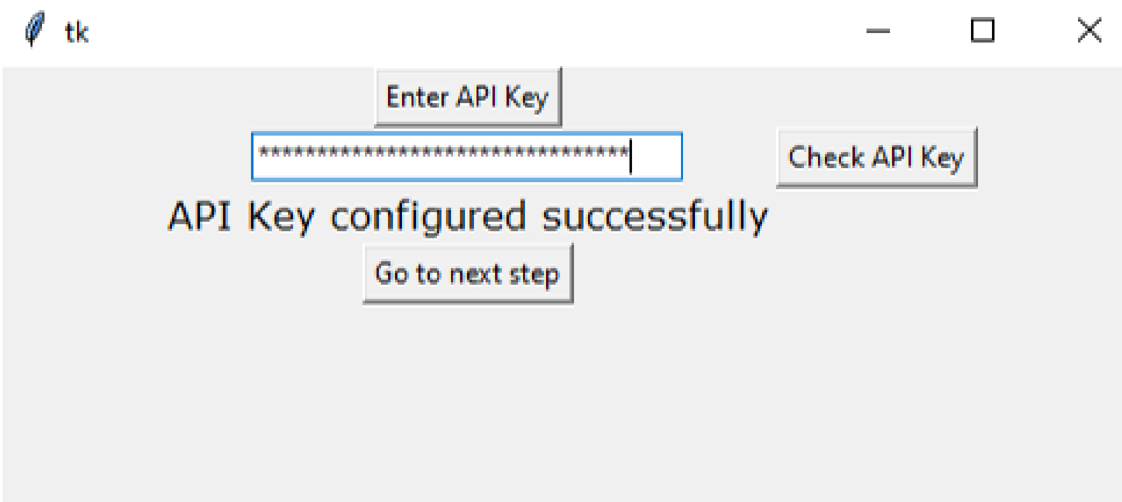


Figure 14: API Key Authentication

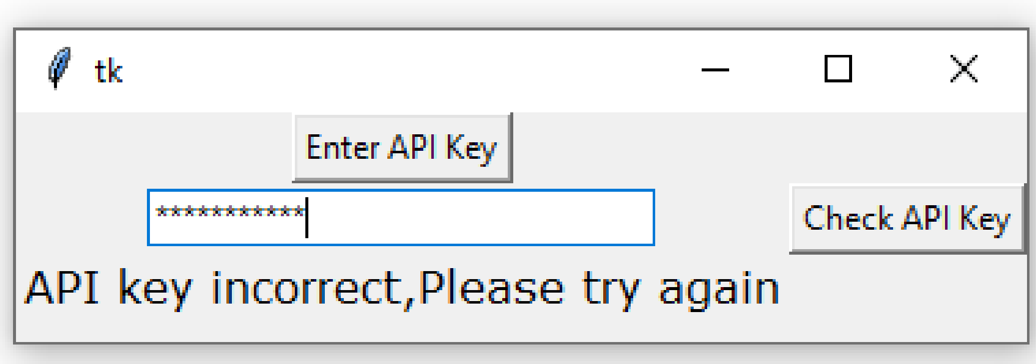


Figure 15: API Key Authentication Failed

For our study, journals displayed in the window are relevant for our queries. The query can be built using a combination of logical operators such as ‘AND’, ‘OR’ and other textual content filters such as ‘TITLE-ABS-KEY()’ *etc.* can be used. The entered query can be seen in Figure 16.

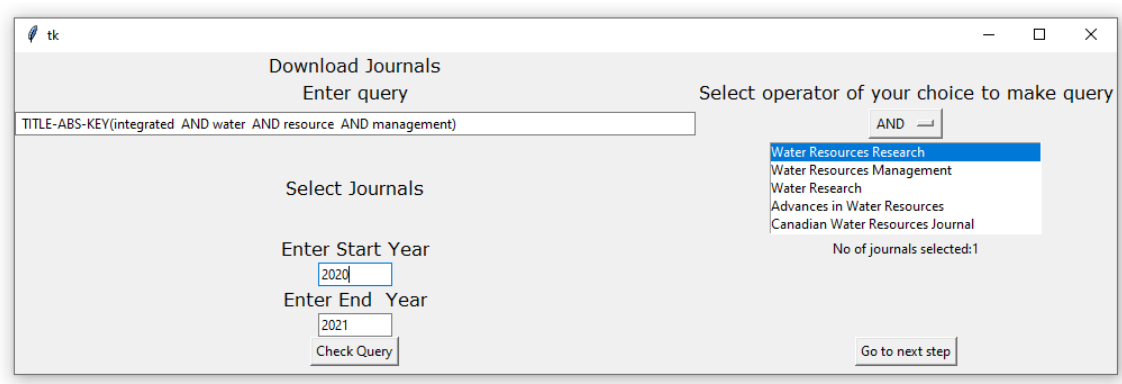


Figure 16: Query Building

The sample of the obtained results can be viewed by clicking on the ‘View Sample’ button as shown in Figure 17. The sample of the results obtained is shown in Figure 18.

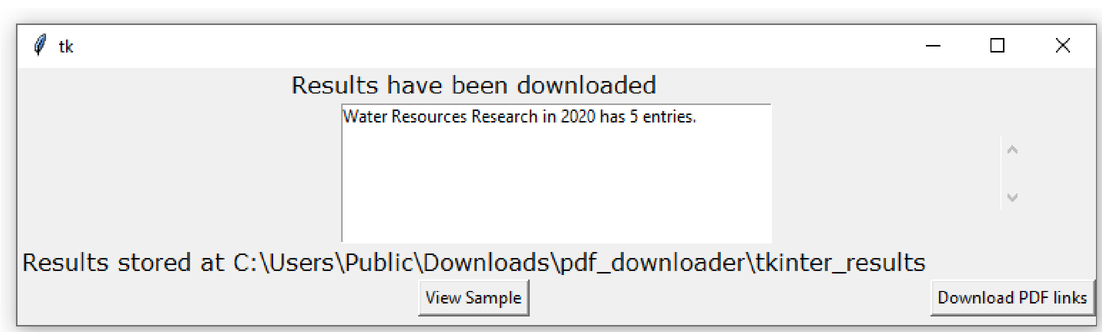


Figure 17: Query Results

To find the PDF links of the research articles is the next step. The Crossref API (Wilkinson, 2020) is utilized to extract these PDF links. The API allows users to access

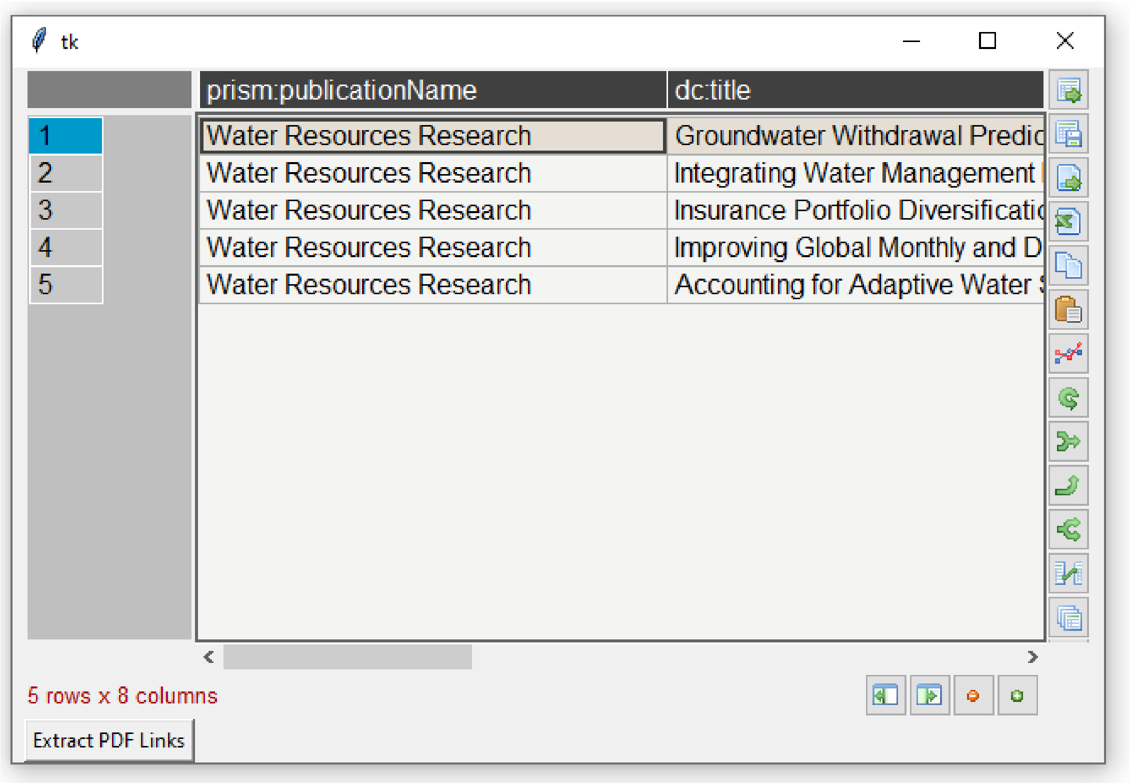


Figure 18: Sample of Results

the full-text documents from all contributing members, the content that has the type open access is delivered without any required subscription. However, if the content is not open access, then the user retrieving the results must have the necessary access to view the content. The user can then download the full-text articles by clicking on the ‘Download full-text articles’ button as shown in Figure 19.

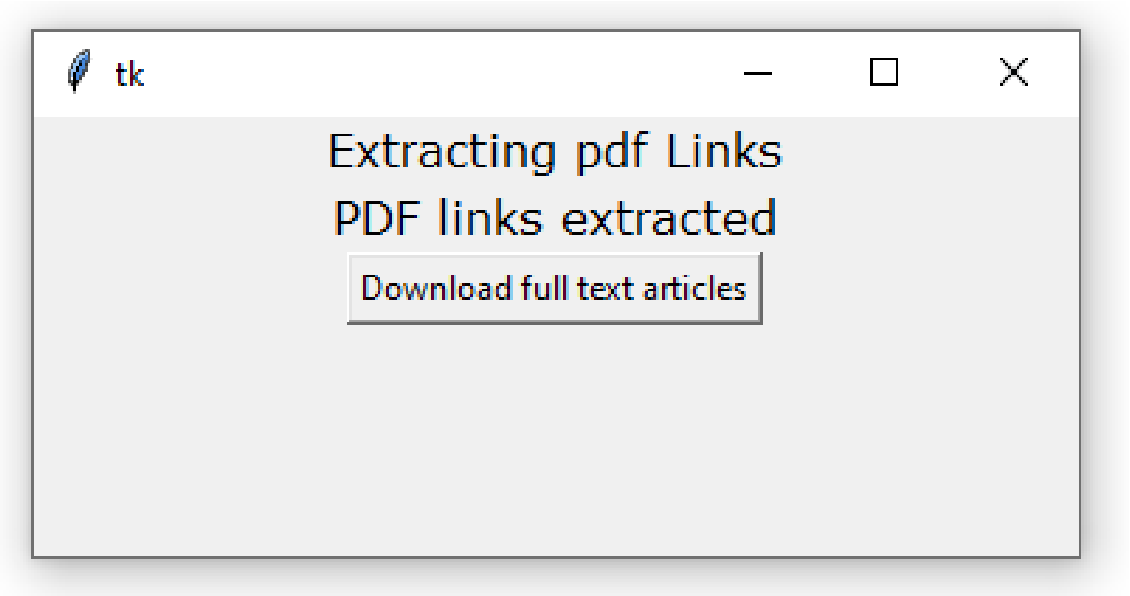


Figure 19: PDF Links Downloaded

Once the articles are downloaded they need to be preprocessed for topic modelling (Figure 20). Firstly, the articles need to be converted from PDF format to text format. This is done by utilizing the *pdfminer.six* package (Shinyaman, 2021). The transformed articles were then passed through a series of pre-processing steps described in Section 3.4. The GUI displays the path of the Excel file which contains the preprocessed corpus as shown in Figure 21. The processed corpus is shown in Figure 22.

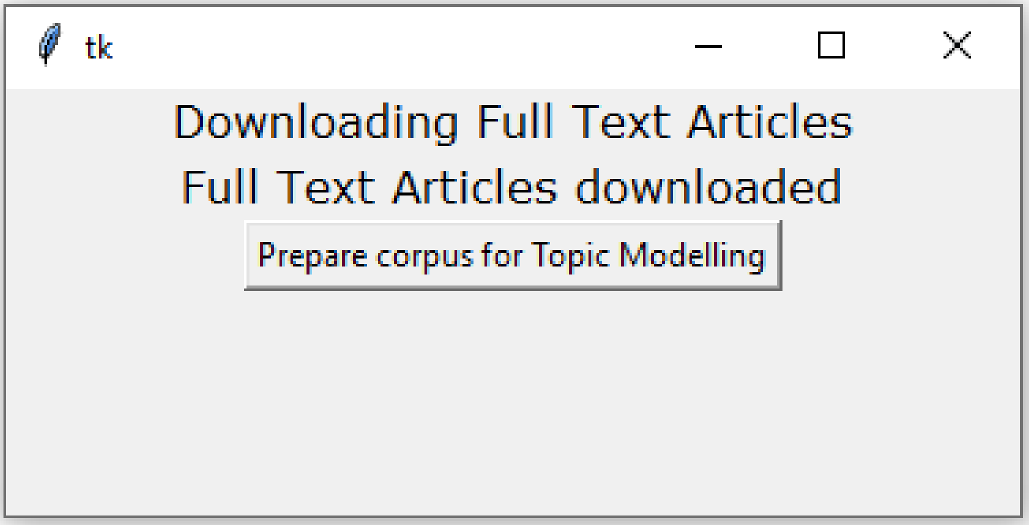


Figure 20: Full-text Articles Downloaded

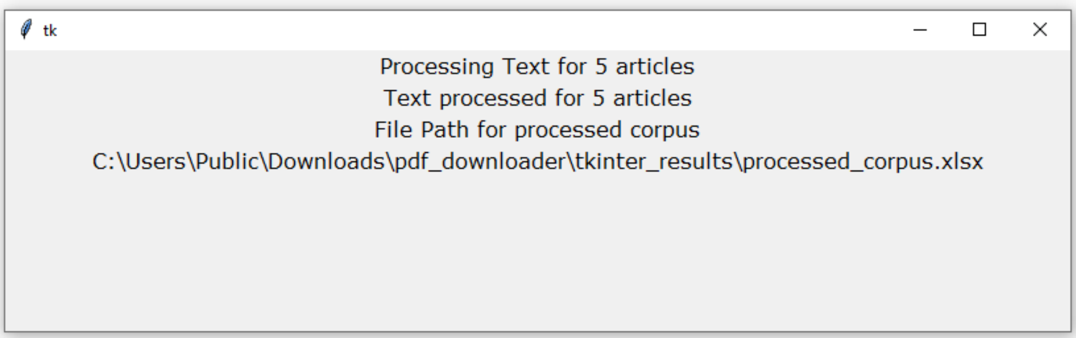
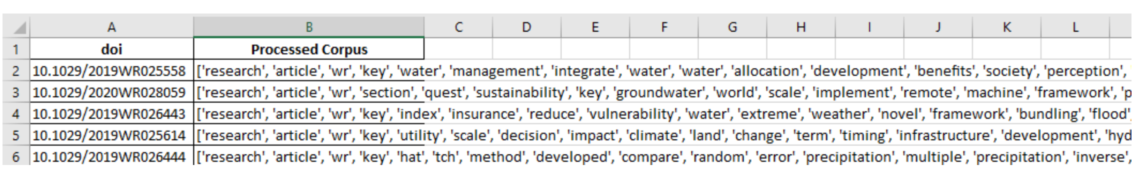


Figure 21: Pre Processed Corpus Location



doi	Processed Corpus
10.1029/2019WR025558	['research', 'article', 'wr', 'key', 'water', 'management', 'integrate', 'water', 'water', 'allocation', 'development', 'benefits', 'society', 'perception',
10.1029/2020WR028059	['research', 'article', 'wr', 'section', 'quest', 'sustainability', 'key', 'groundwater', 'world', 'scale', 'implement', 'remote', 'machine', 'framework', 'p
10.1029/2019WR026443	['research', 'article', 'wr', 'key', 'index', 'insurance', 'reduce', 'vulnerability', 'water', 'extreme', 'weather', 'novel', 'framework', 'bundling', 'flood'
10.1029/2019WR025614	['research', 'article', 'wr', 'key', 'utility', 'scale', 'decision', 'impact', 'climate', 'land', 'change', 'term', 'timing', 'infrastructure', 'development', 'hyd
10.1029/2019WR026444	['research', 'article', 'wr', 'key', 'hat', 'tch', 'method', 'developed', 'compare', 'random', 'error', 'precipitation', 'multiple', 'precipitation', 'inverse',

Figure 22: Pre Processed Corpus

The utility at present can only download the articles present in the Scopus database. However, it can be further extended to include the WoS database as well. The only requirement to use the utility is that the individual must have the subscription to access these articles. Moreover, to reduce the time spent in corpus preparation, multi-threading was used at the backend of the GUI. Therefore, the utility had been developed to download full-text papers from journals which are member-publications.

4.2 Utility for Non Member Publications

The main component of the utility is the CrossRef API which enables the retrieval of the full-text links for a particular DOI (Wilkinson, 2020). The member publication is responsible for providing the requested full-text but many times the publication is not a member and as a result the API fails to return the full-text link for the requested DOI. It is important to include the articles for the missing members otherwise it might result in an unknown bias which might hamper the quality of the results. To automate the retrieval of full-text papers of missing journals, another utility was created which primarily utilizes the Selenium package (Conservancy, 2021). The traditional workflow for manually obtaining a full-text article of a non-member publication involves three steps. Firstly, the user will find the DOI of the research article. After the DOI has been identified, the user will search for that DOI using a web browser. After the user arrives at the landing page of the article, the user can either read the content displayed on the screen or he can click on the button which allows him to download the PDF version of the article. While this workflow for a human is straightforward, for a machine the correct element of the web page which leads to PDF version of the full-text requires additional coding. Since, user interfaces of journals is different, we identified specific patterns in web page organization of the journals listed in major scientific databases. These patterns resulted in the following two scenarios for obtaining the full-text papers. In the first scenario, documents were searched using DOI, the required PDF file appeared in the web browser window and the PDF could be downloaded by locating the button and extracting the full-text link from that button as shown in Figure 23.

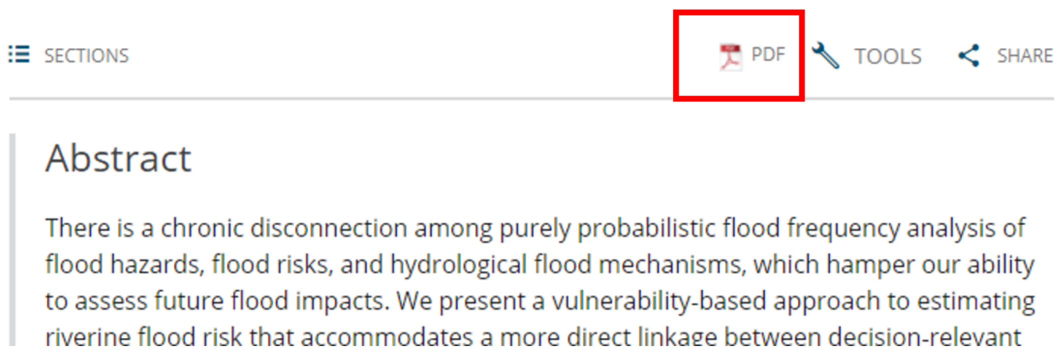


Figure 23: Download PDF Button Displayed on Landing Page

In the other scenario, the DOI was searched, however, the button containing the full-text link was not displayed at the landing web page. The web driver had to identify the web element that contained the element which had the full-text link. Once the element containing full-text link was identified, the next task was to extract the full-text link from that element (Figure 24). The drawback of this approach was that since the Selenium module does not support multi-threading, accelerating the process of automatic downloading of non member publications was not possible. This limits the application of this utility in industry grade applications. Thus, the feature to download non member publications was not integrated with the utility coded for member publications. Despite, the technical limitation, the utility still outperforms the manual process of text document collection and pre-processing. The code block for this utility can be referred in Appendix 9.1

Note: The webdriver first clicked on the button 'Download PDF' and then the dropdown appeared which had the button 'Download this article', which contained the PDF link of the full-length research paper.

Figure 24: Additional Click Required to Fetch PDF Link on Landing Page

5 Problem Domain: Integrated Water Resource Management

Integrated Assessment (IA) can be defined as 'convenient frameworks for combining knowledge from a wide range of disciplines in order to conduct coordinated exploration of possible future trajectories of human and natural systems, development of insights into key questions of policy formation, and prioritization of research needs in order to enhance our ability to identify robust policy options' (Weyant et al., 1995). Policy makers while formulating policies need to take into consideration the multifariousness of societal demands, the unpredictability of relevant events, different interests of various stakeholders and how to collect and assimilate information from different sources. This calls for integrated assessment methodology (Toth and Hizsnyik, 1998). Therefore, interdisciplinary or multidisciplinary study of an environmental problem that leads to an output, supporting decision or policy making relevant to the investigated problem can be classified as an Integrated Environmental Assessment (IEA) exercise (Tol and Vellinga, 1998). To facilitate the decision making, IEA aims to harness information regarding the impact caused by human activities, current environmental state and the driving socio-economic forces. Therefore, in the context of a natural water resource, its sustainable development and management is a multi disciplinary subject which involves the integration of various disciplines such as hydrology, ecology, economics, sociology and management (Thornton et al., 2006). The Technical Committee of the Global Water Partnership (GWP) gave a formal definition to Integrated Water Resources Management: 'A process which promotes the coordinated development and management of water, land and related resources, in order to maximize the resultant economic and social welfare in an equitable manner without compromising the sustainability of vital ecosystems' (Abdel-Magid and Ahmed, 2002). Every dimension gives important information regarding the present state of the water resource and can aid in decision making for example, availability of water for various designated uses can be estimated with the help of the water quality, accompanied with the hydrological features of the region. Ecological aspects of water resources significantly depend on the cumulative effect of an impact on the environment caused by the human demands for various goods and services. The accountability of these services is based on an economic assessment of a water resource. For example, the economic valuation of a water resource should take into account numerous aquatic ecosystem services, the cost of water purification if it is necessary for various water consumption needs and the operational cost of water supply systems. However, in many cases, the cost of water is based only on the latter component. Moreover, the price of water ignores the consequences of its uses by other users and environment. For instance, imprudent abstraction results in adverse economic impact, water bodies experience pollution due to the disposal of industrial discharge which not only is harmful for the environment but also causes a menace for the water users downstream. As a result, the economic quantification of the water resource is often undervalued (Global, 2021). Thus, two major objectives of IWRM are to provide comprehensive utilization as well as the sustainability of the resource, which makes economic assessment an integral dimension for IWRM (Statistics, 2012). The water resource is vital for generating human capital as it influences the health and well-being of the human society. The sustainable management of the water resource should aim at accommodating the society's spiritual and social cultural demands in an unbiased manner. These indicators are evaluated following qualitative approaches based on surveys and interviews. Therefore, IWRM is a multidisciplinary domain where the decision can be supported by analysis of heterogeneous data of various types and scales.

This also makes management of a natural resource a political process rather than a scientific one (Balkema et al., 2002). The basis of integrated water resource management aims not only to improve the environmental quality but also to increase awareness regarding the detrimental activities of the human beings which hampers sustainable management of the resource. Sustainable integrated resource management of water resource strives to improve the current regulations with an objective to make the regulations more aligned with the multi-dimensional perspective of sustainability. The ultimate goal of integrated assessment of a water resource is to provide decision-makers with information on available options and outcomes of each option in order to weigh benefits against any negative direct and indirect consequences. The efficiency of IWRM depends on the selection of criteria for assessment (Loucks and Gladwell, 1999), selection of appropriate modelling tools (Erechtchoukova and Khaiter, 2007) and evaluation of uncertainty in obtained quantitative estimates (Erechtchoukova and Khaiter, 2009). To quantify the environmental health, environmental simulation modelling is used which derives aggregated values from the observation data recorded from various sites under consideration. The series of generated data regarding the status of the environment can help the management bodies to assess the present and future environmental states (Erechtchoukova and Khaiter, 2011). The ecological assessment of a water resource is governed by numerous aspects of water health. Water quality conditions of an aquatic ecosystem are assessed to evaluate the state of the water resource. This helps to decide which activities to undertake to mitigate the water pollution, provides better co-ordination in stream restoration projects, controls the discharge of pollutant inputs and yields better utilization of the staffing resources and funds (DepartmentofEnvironmentalProtection, 2021). Apart from ecological and environmental aspects, economic aspect of a water resource is also considered in policy making, this can be measured using indicators such as affordability (Foxon et al., 2002), user sector productivity (Liu et al., 2008), cost effectiveness of reliable access to water (van der Zaag and Gupta, 2008). To capture the contribution of the social aspect, decision makers evaluate the fraction of population that has access to clean water for drinking and sanitation purposes, while other qualitative features such as understanding of the public and their awareness with respect to the present state of the resource, social inclusion and community integration are also taken into consideration (Glasson and Wood, 2009). These indicators are evaluated using surveys and interviews. Thus, IWRM is a multifaceted problem. It involves both quantitative and qualitative data analysis. This adds complexity to IWRM because the important characteristics are not only described differently but also measured on different scales. IWRM enables us to holistically capture the interactions of different dimensions and can help us to achieve the United Nations' long-term development goals. Therefore, the primary reason of selecting IWRM as the research field was to identify the dominating themes and research trends by leveraging topic modelling algorithms. The role of the social factors is crucial in IWRM and is often left untouched because of the complexity involved to integrate the social dimension into the research methodology. Hence, another reason of analysing IWRM was to observe how much of the research space has matured with respect to the social components of IWRM.

6 Computational Experiments

The experiments were performed following the modified framework (Figure 10). The starting point of every research is to identify a suitable problem domain which can be explored to further contribute to the existing knowledge in that particular domain. This study investigated the domain of Integrated Water Resource Management (IWRM). To prepare the corpus, out of the three prominent scholarly databases *i.e.* Web of Science, Scopus and Google Scholar, Scopus was selected to download the papers of interest. The primary reason for selecting Scopus was that York University provided direct access to full-text articles and also it has been labelled as having the largest collection of peer reviewed literature ranging from journals to books to conference proceedings *etc.* (Kumar et al., 2015). In order to fetch the research papers in the domain of IWRM and Integrated Environmental Assessment, the following journals were selected after due diligence. The journals were: *Water Resources Research, Water Resources Management, Water Research, Advances in Water Resources, Canadian Water Resources Journal, Sustainable Water Resources Management, Water Environment Research, Journal of Water Resources Planning and Management, Journal of hydrology, Integrated Environmental Assessment and Management, Environmental Impact Assessment Review, Annual Review of Environment and Resources, Journal of Environmental Management, Environmental modelling and software*. All the papers from these journals were considered in the analysis. Since, it is essential to cover a substantial time period to observe the major developments in a particular field, research papers for the last 50 years were considered for the analysis. The timeframe was fixed between 1970-2020. Text extraction was performed in two steps. Firstly, the abstracts of the selected research papers were downloaded from Scopus database along with other metadata while the full-text of the selected research papers was obtained using the utility discussed in Section 4.1. Once the required research papers were downloaded, they were first tokenized and then subjected to stopwords removal. Following this, bi-grams were also integrated into the corpus. Along with the filtering of stopwords, all the parts of speech other than nouns and verbs, were also removed from the analysis. Lemmatization was preferred over stemming. LDA Mallet was selected as the topic modelling algorithm to identify the latent themes across the corpus. C_v was the chosen evaluation measure to quantitatively analyze the LDA algorithm. This section explains the results of the experiments conducted on the research articles published in the domain of Integrated Water Resource Management. The computational experiments were conducted to investigate the effectiveness of the pre-processing strategies for corpus construction on the identified topics and the quantitative impact of these strategies on the coherence score.

6.1 Corpora Development

From the journals described in Section 6, 89726 articles which were published between 1970 to 2020 were retrieved. The list of these articles including their metadata was downloaded using Scopus Database. The metadata of each article contained names of authors, unique id assigned to each author in the Scopus database, full title of the article, name of the journal in which it was published, unique digital object identifier (DOI), number of citations, abstract of the article, references, language in which the document was written, type of the document and unique article ID in Scopus. Type of the document represents article category, *e.g. 'Article', 'Review Paper' etc*. Some Journals were not indexed by the Elsevier API or by the CrossRef API and as a result the size of the final corpus containing the full-text articles was reduced to 29032 research papers. The size of the corpus is still bigger than the size of the corpus used in previous studies. Therefore, all the experiments were performed on the final corpus.

Table 2 shows availability of full-length papers in all considered journals. To perform the experiments, three different corpora were constructed. The first corpus consisted of the abstracts. In this corpus, the abstracts were converted into a BoW representation. LDA algorithm was then trained on the first corpus and the obtained coherence score served as the baseline score. The second corpus was constructed using the same abstracts but the abstracts were filtered using the TF-IDF approach discussed in Section 3.5. The LDA algorithm was trained again using the second corpus and the coherence score was compared with the baseline score to check for any improvements. The third corpus was constructed using the full-text of the research papers used in first corpus. The full-length research papers were filtered using the TF-IDF approach and LDA algorithm was trained using this corpus. The obtained coherence score of the model was then compared with the other two models to check for any significant improvement.

Table 2: List of Journals

Name of Journal	Full-text Available
Water Resources Research	Yes
Water Resources Management	Yes
Sustainable Water Resources Management	Yes
Canadian Water Resources Journal	Yes
Journal of Water Resources Planning and Management	Yes
Water Environment Research	Yes
Integrated Environmental Assessment and Management	Yes
Environmental Impact Assessment Review	No
Water Research	No
Advances in Water Resources	No
Journal of Environmental Management	No
Environmental modelling and software	No
Journal of hydrology	No
Annual Review of Environment and Resources	No

6.2 Preliminary Data Analysis

Figure 25 shows the number of publications per year. It can be observed that the research in the field picked up in the late 1990's which also coincides with the establishment of Global Water Partnership whose purpose was to foster IWRM (Smith and Jønch Clausen, 2015). Only three of the fourteen selected journals were founded after 2000, for reference see Figure 26. Journal '*Water Research*' had the highest number of publications followed by '*Water Resources Research*' (Figure 27).

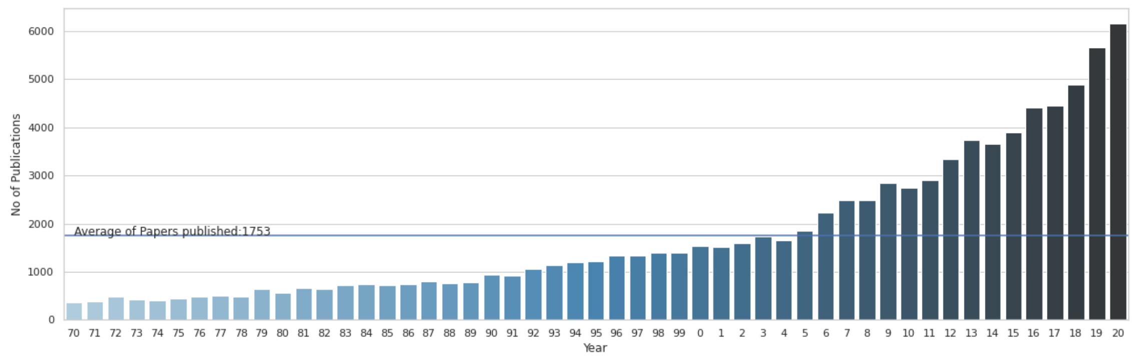


Figure 25: Publications per Year

6.3 Topic Modelling

For performing topic modelling on the prepared corpora, implementation of LDA Mallet algorithm provided by Gensim library was chosen. LDA Mallet algorithm consists of *iterations* and *optimize_interval* hyperparameters. Hyperparameter *iterations* represent the number of training iterations and *optimize_interval* governs after how many iterations the hyperparameters would be re-optimized. The LDA algorithm was trained for 1000 iterations and *optimize_interval* was set to 100. These two parameters took the same value whenever a new instance of the LDA algorithm was trained. To obtain the baseline score for further comparisons, various instances of LDA algorithm were trained with different number of topics. The corpus used in calculation of the baseline score was prepared using abstracts. The corpus was represented in the BoW form. Hyperparameter *num_topics*

	Journal Name	Year of First Publication
0	Journal of Environmental Management	1970
1	Journal of Hydrology	1970
2	Water Research	1970
3	Water Resources Research	1970
4	Canadian Water Resources Journal	1976
5	Advances in Water Resources	1977
6	Environmental Impact Assessment Review	1980
7	Journal of Water Resources Planning and Manage...	1983
8	Water Resources Management	1987
9	Water Environment Research	1992
10	Environmental Modelling and Software	1997
11	Annual Review of Environment and Resources	2003
12	Integrated Environmental Assessment and Manage...	2006
13	Sustainable Water Resources Management	2015

Figure 26: Year of First Publication for Each Journal

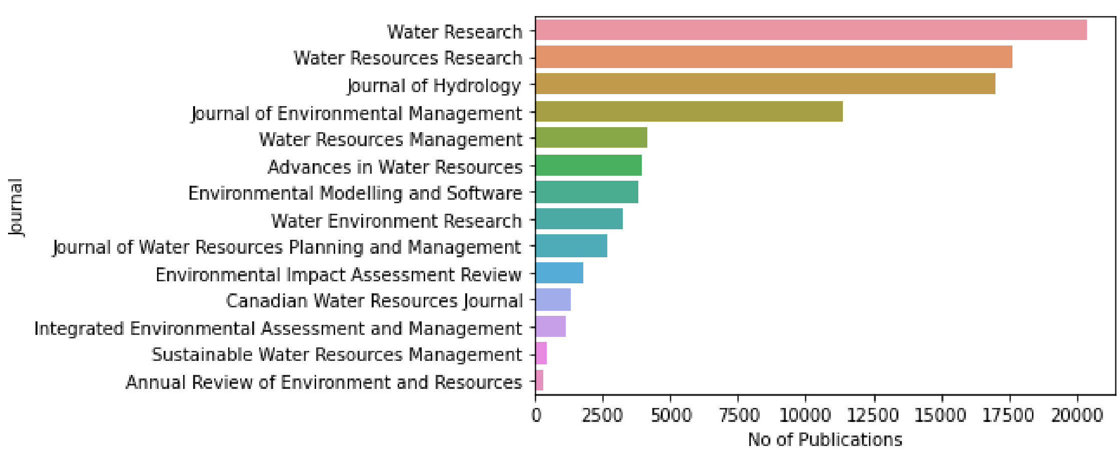
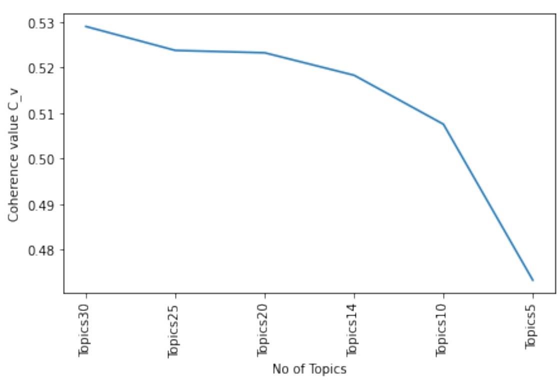


Figure 27: Distribution of Publications across Journals

was varied from 5, 10, 14, 20, 25 and 30. Coherence metric C_v was calculated for each LDA algorithm which was obtained as *num_topics* was varied from 5 to 30. The coherence score C_v was plotted against *num_topics* (Figure 28). As seen in Figure 28, LDA algorithm trained for 30 number of topics gave the best results, however, qualitative analysis of derived topics showed that interpretation of individual topics was difficult. It was observed that multiple topics indicated to the same theme. Thus, it made sense to investigate *num_topics* which were less than 30. Similar inference was drawn when *num_topics* was equal to 25. LDA algorithm trained for 20 topics delivered interpretable topics and sufficiently high coherence score. Therefore, the coherence score C_v of LDA algorithm trained for *num_topics* set to 20, was considered as the baseline coherence score for further comparisons. The topics obtained from LDA algorithm trained on corpus prepared using only abstracts are shown in Figure 29.



(a) Coherence Plot

	Topics	Coherence Type	Coherence Value
0	Topics30	c_v	0.529041
1	Topics25	c_v	0.523786
2	Topics20	c_v	0.523263
3	Topics14	c_v	0.518328
4	Topics10	c_v	0.507565
5	Topics5	c_v	0.473292

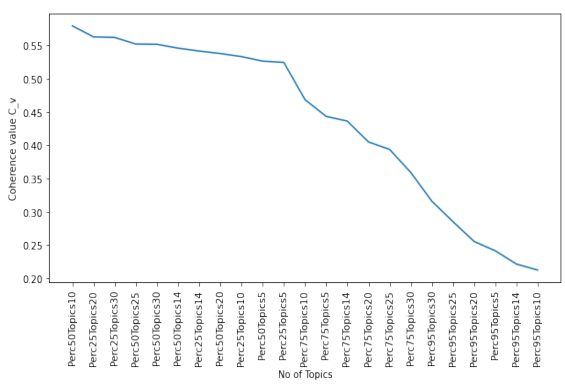
(b) Coherence Score

Figure 28: Analysis of Topics Obtained by LDA Algorithm Trained on Abstracts BoW

Topic No.	Words forming the Topic	Topic Label
1	soil water flow fracture surface infiltration saturation permeability depth layer pore drainage pressure gas property	***Infiltration and Saturation***
2	method base problem approach result technique solution apply propose present procedure obtain develop application require	***Methodological Approaches***
3	measure site sample test field datum measurement study result determine range difference location obtain find	***Monitoring***
4	estimate datum uncertainty parameter error base estimation method observation information approach reserve measurement union_right provide	***Data Processing Techniques***
5	catchment stream runoff dynamic control reserve water response storage union_right pattern soil discharge area vegetation	***Watershed Hydrology***
6	system process approach provide analysis include development research information paper develop framework management application tool	***Methodological Approaches***
7	distribution function scale parameter show time derive give relationship property analysis find structure correlation characteristic	***Statistical Analysis***
8	water demand irrigation level system management supply crop region source area year surface production efficiency	***Irrigation***
9	concentration risk assessment level chemical effect specie exposure base contamination pollution factor quality study source	***Risk Assessment of Pollution***
10	management policy project cost benefit plan program decision resource planning development service implementation community risk	***Water Related Project Management***
11	flow transport equation concentration solution field solute velocity time tracer dispersion case result medium condition	***Mesocosm for Model Calibration***
12	system reservoir network design optimization operation cost constraint reliability performance control pipe pressure uncertainty develop	***Reservoir Operation Management***
13	rainfall precipitation flood year drought period region event climate index runoff storm station record trend	***Prediction of Droughts***
14	temperature water surface lake flux evaporation snow heat energy depth measurement condition soil period summer	***Lake Water Balance***
15	model simulation base simulate parameter prediction result develop performance datum predict forecast input calibration apply	***Simulation Modelling***
16	concentration treatment removal process wastewater system study plant sludge reactor adsorption remove bacteria investigate degradation	***Wastewater Treatment***
17	rate condition time result process effect phase experiment study show observe reserve behaviour find determine	***Experimental Hydrology***
18	flow river discharge sediment channel stream bed surface velocity reach transport depth particle wave slope	***Hydrodynamics***
19	increase change effect impact result decrease study condition reduce reduction scenario show area level affect	***Land Use Change Impact***
20	groundwater aquifer flow zone recharge system salinity area head pump surface salt depth drawdown response	***Ground waters***

Figure 29: Topics obtained from Abstracts BoW

After obtaining the baseline score, TF-IDF filtration was applied to the corpus consisting of only abstracts. Let ω be the threshold parameter, words having TF-IDF score less than ω in a document would be removed for a document, ensuring only relevant words for the document are retained. LDA algorithm was trained for each combination of *num_topics* and ω, *num_topics* were varied from 5 to 30 and the value of ω ranged from 25 percentile, 50 percentile, 75 percentile and 95 percentile of the TF-IDF scores. Coherence metric C_v was calculated for each LDA algorithm and the variation in coherence score C_v by changing *num_topics* and ω is shown in Figure 30. By analyzing the results obtained in Figure 30, LDA algorithm trained for 10 topics with value of ω set to 50 percentile of the TF-IDF score yielded the highest coherence score, however, the qualitative analysis of those ten topics revealed that the topics were too generic. LDA algorithm trained with *num_topics* set to 20 and ω set to 25 percentile of the TF-IDF scores, yielded the second highest coherence score.



(a) Coherence Plot

	Combination	Coherence Type	Coherence Value
0	Perc50Topics10	c_v	0.579076
1	Perc25Topics20	c_v	0.562532
2	Perc25Topics30	c_v	0.561724
3	Perc50Topics25	c_v	0.551755
4	Perc50Topics30	c_v	0.551409

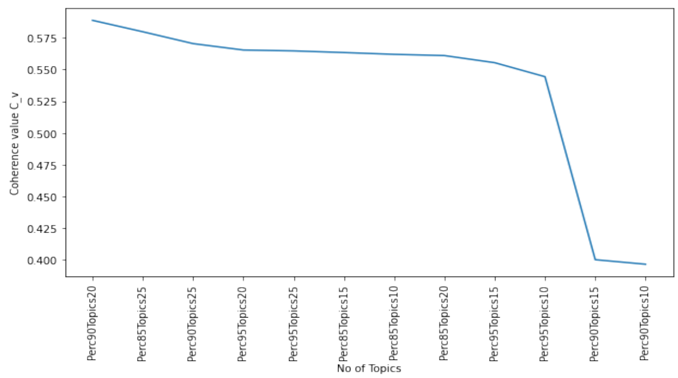
(b) Coherence Score of top 5 Topic models

Figure 30: Analysis of LDA Algorithm Trained on Abstract TF-IDF Representation

Analysis of the obtained topics (Figure 31) revealed that the topics were informative and meaningful. Integration of the TF-IDF filtration to abstracts led to an improvement of the coherence score by **7.5%** as coherence score jumped from 0.5232 to 0.5625. Lastly, the effect of integration of the full-length research papers on the coherence score C_v was investigated. To perform this experiment the values of ω were set to 25 percentile, 50 percentile, 75 percentile, 85 percentile, 90 percentile and 95 percentile of the TF-IDF scores. The remaining hyperparameters were set in accordance to the previous experiments so that the effect of full-text corpus could be analyzed in isolation. LDA algorithm was trained for each combination of *num_topics* and ω. Coherence metric C_v was calculated for each LDA algorithm and the variation in coherence score C_v by changing *num_topics* and ω is shown in Figure 32.

Topic No.	Words forming the Topic	Topic Label
1	water irrigation demand crop cost supply price allocation policy household user transfer farmer production conservation	***Irrigation***
2	removal treatment process wastewater concentration sludge plant reactor remove adsorption bacteria degradation day rate solid	***Wastewater Treatment***
3	groundwater aquifer zone recharge salinity pump storage head level salt age drawdown response area depth	***Groundwaters***
4	system reservoir design network optimization operation cost demand performance reliability constraint pipe control pressure energy	***Optimization of Reservoir Operations***
5	community exposure program research population review assessment specie include risk fish chemical development survey issue	***Community Risk Assessment of Water Toxicity***
6	distribution time function show find measure derive determine range obtain rate coefficient equation case give	***Mathematical Modelling***
7	flow river discharge stream channel bed velocity sediment reach rate depth transport surface wave exchange	***Stream Hydrological Modelling***
8	snow vegetation measurement area site observation sensor map elevation product tree ground forest plant measure	***Forest Hydrology***
9	transport concentration solute tracer particle dispersion rate mix plume velocity reaction diffusion experiment sorption column	***Aquatic Mesocosms***
10	management scenario risk project development impact plan strategy level flood integrate decision planning tool index	***Flood Risk Mitigation***
11	soil surface flux depth infiltration evaporation profile layer drainage field condition moisture root measurement conductivity	***Infiltration***
12	fracture permeability field pressure saturation conductivity hydraulic_conductivity medium phase formation heterogeneity interface rock pore density	***Water Saturation***
13	model simulation prediction forecast parameter calibration simulate performance input predict calibrate develop output forecasting improve	***Predictive Simulation Modelling***
14	increase change effect temperature decrease condition impact trend climate period scenario lake reduction response affect	***Climate Change Impact Assessment***
15	process scale pattern structure dynamic variability network control behaviour influence response interaction relationship characteristic link	***Scale Issues in Hydrology***
16	stream concentration source area sediment site load lake discharge control wetland sample pollution nutrient watershed	***Lake Organic Pollution***
17	method estimate datum parameter error sample estimation test measurement site field procedure observation sampling data	***Monitoring***
18	approach base solution problem equation propose formulation present solve develop scheme term algorithm technique case	***Mathematical Techniques***
19	rainfall catchment precipitation runoff year event flood drought storm storage region period basin record station	***Watershed Hydrology***
20	uncertainty analysis study result provide application information paper make technique include state exist evaluate present	***Methodological Approaches***

Figure 31: Topics Obtained from Abstract TF-IDF Corpus



(a) Coherence Plot

	Combination	Coherence Type	Coherence Value
0	Perc90Topics20	c_v	0.588880
1	Perc85Topics25	c_v	0.579867
2	Perc90Topics25	c_v	0.570515
3	Perc95Topics20	c_v	0.565450
4	Perc95Topics25	c_v	0.564766

(b) Coherence Score of top 5 Topic Models

Figure 32: Analysis of LDA Algorithms Trained on Full-Length Research Papers

By analyzing the results displayed in Figure 32, LDA algorithm trained for 20 topics and ω equal to 90 percentile of the TF-IDF scores yielded the highest coherence score. The qualitative analysis of the topics (Figure 33) also revealed that the topics were able to capture the contribution of multiple disciplines involved in IWRM. Thus, integration of full-length research papers led to an improvement of the coherence score by **12.54%** over the baseline score. It also showed an improvement of **4.68%** over the coherence score obtained by TF-IDF filtration of the abstracts.

Topic No.	Words forming the Topic	Topic Label
1	flow equation solution field function velocity variance head transport case cid coefficient dispersion problem element	***Hydrodynamics***
2	fracture particle transport flow permeability test experiment velocity pressure injection simulation porosity image point rock	***Sediments***
3	flood distribution level probability variable flood cluster variance event correlation series frequency estimator estimation duration	***Flood Predictions***
4	concentration rate transport source reaction stream adsorption mix column load solute sediment sorption nitrogen equilibrium	***Hydrochemistry***
5	cost optimization design solution system constraint problem scenario management uncertainty fig pump plan objective objective_function reliability	***Optimisation of Water Distribution***
6	concentration exposure specie chemical risk sediment assessment toxicity fish test metal pesticide substance organism contaminant	***Aquatoxicology***
7	reservoir storage period operation release inflow volume stage dam demand rule energy fig capacity month	***Hydropower Management***
8	surface temperature figure snow lake soil_moisture evaporation depth vegetation heat flux canopy layer day wind	***Snow Hydrology***
9	groundwater area recharge aquifer zone table depth head age pump fig factor wetland unit layer	***Groundwater***
10	site sample measurement pattern location sampling group probe monitoring monitor plot survey sensor sampler frequency	***Water Monitoring***
11	network pipe node demand pressure control fig method tank leakage failure link pump sensor valve	***Water Distribution Systems***
12	water demand cost price supply household consumption allocation transfer user benefit market sector utility policy	***Water Allocation***
13	management project policy criterion state risk indicator plan quality program water_quality community decision development resource	***Water Quality Management***
14	soil depth saturation phase infiltration pressure air layer gas experiment drainage profile water_content root column	***Water Infiltration***
15	treatment concentration removal wastewater sludge reactor plant cod bacteria membrane day solid cell substrate biomass	***Wastewater Treatment***
16	precipitation year climate region station drought basin change trend period month temperature rainfall index streamflow	***Precipitations***
17	model parameter estimate error uncertainty calibration prediction input datum observation simulation performance forecast estimation output	***Simulation Modelling***
18	irrigation scenario crop basin impact land farmer yield salinity land_use farm production reduction canal fig	***Irrigation***
19	discharge stream river flow sediment flow bed velocity channel reach depth transport stress slope run	***Water Constituents Transport***
20	rainfall catchment runoff event storm scale response soil slope watershed basin hillslope peak forest rain	***Watershed Modelling***

Figure 33: Topics obtained from Full-Text TF-IDF Corpus

6.3.1 Visualization of Topics

PyLDAvis is a python package which has the capability to visualize the topics obtained via the Gensim LDA model. The package fetches the information from the trained LDA model to display an interactive web-based visualization (Mabey, 2021). pyLDAvis is the python extension of the original LDAvis (Sievert and Shirley, 2014) implemented in R programming language. The topics generated by the LDA algorithm trained on corpus prepared using abstracts (Corpus A), corpus prepared using abstracts and filtered using TF-IDF (Corpus B) and corpus prepared using full-length research papers and filtered using TF-IDF (Corpus C) were visualized using pyLDAvis. The topics obtained from corpus A are visualized in Figure 34, topics obtained from corpus B are visualized in Figure 35 and topics obtained from corpus C are visualized in Figure 36. Consider Figure 34, each bubble in the visualization corresponds to a topic. The size of the bubble determines the percentage contribution of the topic to the corpus. Large size of the bubble highlights that more number of documents in the corpus have that topic as their central theme. The blue bars that are displayed in the visualization corresponds to the count of words in the corpus and the overlaying red bars that appear when a topic is selected indicates how many times that word appeared under the selected topic. If size of red bars is large, it indicates that the words that formed the topic had a high occurrence just for that topic only. The distance between the bubbles helps to distinguish between the topics and also captures the semantic relationship between the topics. Generally, if the corpus consist of non overlapping themes then high separation between the bubbles is an indication of high quality topics, however, if the corpus consists of themes which are highly inter-correlated then overlapping bubbles can be obtained and is an indication that the topics were able to capture the subtle differences between the two themes. The pyLDAvis obtained for each corpus was able to capture the semantic relationships existing between the topics. For instance, consider the topics 'Topic 2 (Methodological Approaches)', 'Topic 4 (Data Processing Techniques)' and 'Topic 15 (Simulation Modelling)' in Figure 34. Intuitively, these three labels are expected to have high semantic relationship and this can be observed in Figure 34. Likewise consider topics 'Topic 19 (Watershed Hydrology)', 'Topic 14 (Climate Change Impact Assessment)' and 'Topic 8 (Forest Hydrology)' in Figure 35. These three research areas are highly correlated and this was rightly captured by the visualization (Figure 35). Lastly, consider 'Topic 11 (Water Distribution Systems)', 'Topic 7 (Hydropower Management)' and 'Topic 5 (Optimization of Water Distribution)' in Figure 36. The high semantic relationship between these three labels can be again observed in Figure 36. This highlights that the obtained topics from *Corpus A*, *Corpus B* and *Corpus C* captured the semantic relationships between the topics that existed in IWRM domain.

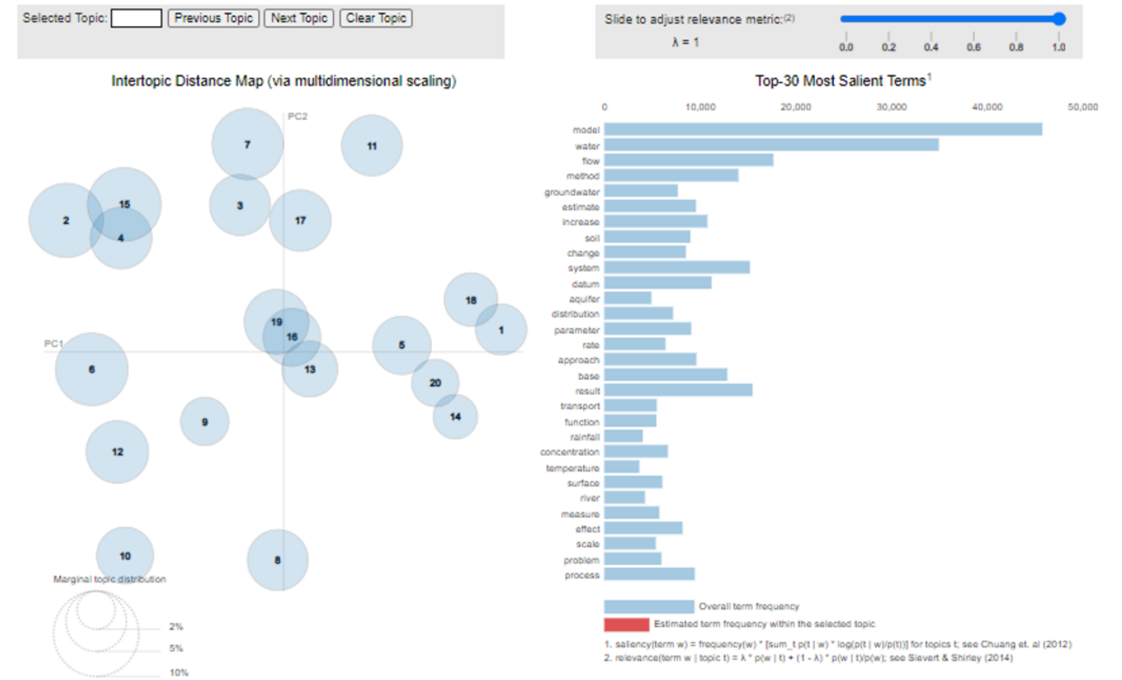


Figure 34: pyLDAvis for Corpus A

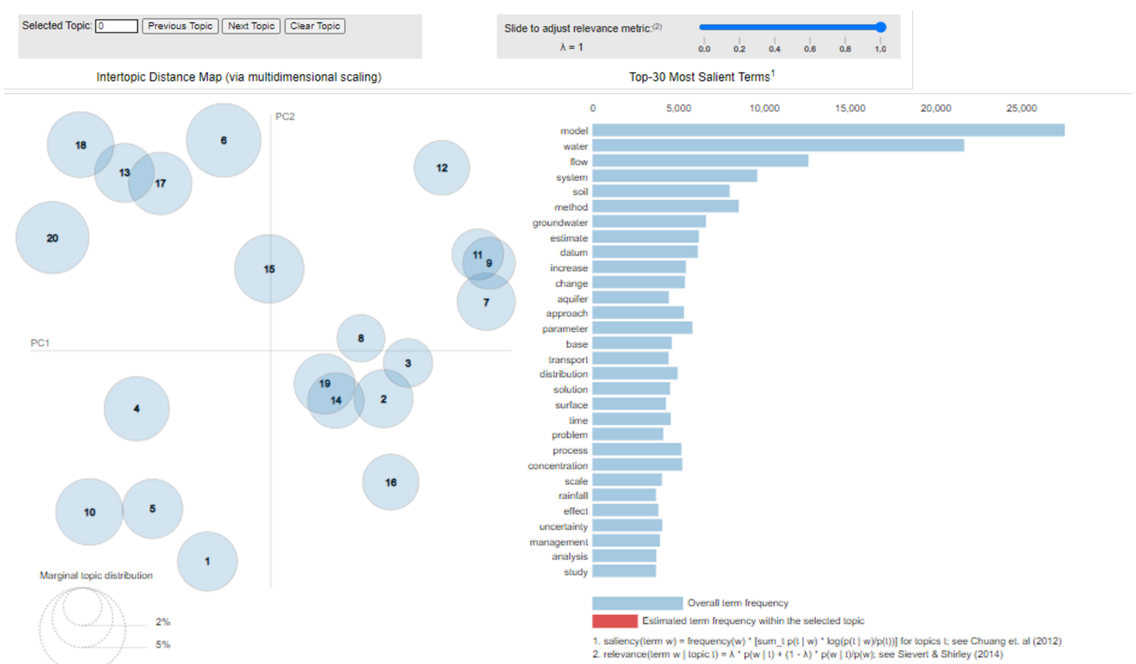


Figure 35: pyLDAvis for Corpus B

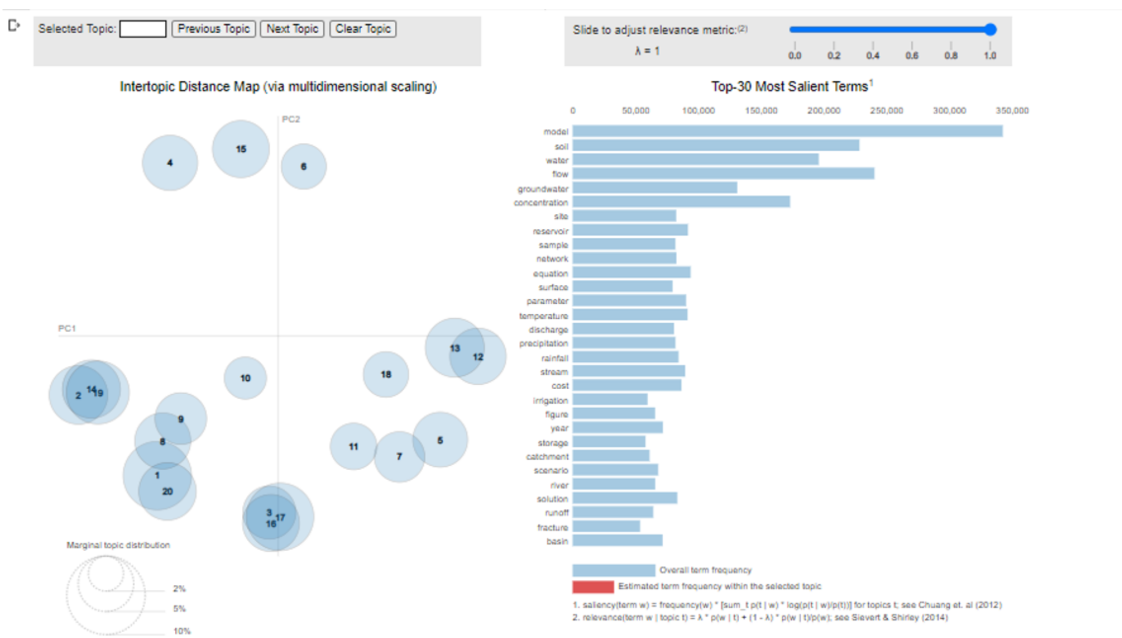


Figure 36: pyLDAvis for Corpus C

6.4 Qualitative Analysis of Results

6.4.1 Qualitative Analysis of TF-IDF Filtration

A document was randomly sampled from the corpus and the words removed by the TF-IDF filtration were observed. The title of the sampled document was "*How Dynamic Boundary Conditions Induce Solute Trapping and Quasi-stagnant Zones in Laboratory Experiments Comprising Unsaturated Heterogeneous Porous Media* ". Table 3 contains the words that were eliminated for the document along with the overall frequency of the words in the corpus.

The eliminated tokens had high frequency in the corpus and low frequency in the document. For the LDA algorithms trained on corpus consisting of abstracts and represented in the BoW format, these words were considered for this document. These tokens can disturb the topic proportion that would be allocated by the topic modelling algorithm and thus, act as a source of noise for this particular document. The token 'water' is important in general for the whole corpus but inclusion of it in this document would likely disturb the composition of the topics. This does not happen in training of LDA algorithms, where the abstracts were filtered using the TF-IDF model. This makes the LDA algorithm less susceptible to noise and as a result the improvement in the coherence score is observed.

Table 3: Tokens eliminated by TF-IDF filtration

Token Removed	Token Frequency in Corpus	Token Frequency in Document
water	33563	1
result	14138	1
study	12056	2
process	8823	1
parameter	8398	1
effect	7631	1
condition	7574	1
consider	4758	1
observe	4554	1
control	4260	1
present	4058	1
application	4010	1
structure	2911	1
analyze	2634	1
form	2460	1

6.4.2 Qualitative Analysis of Topics obtained from corpus prepared using Abstracts

Consider Figure 37 and Figure 38. Topic 1 was selected in both the figures and it was observed that the order of the relevant terms differed in both. The reason behind this difference is the relevance metric also denoted by λ. λ determines the order of the relevant words for a topic. Sievert and Shirley (2014) viewed relevance of a word for a topic from two perspectives. One perspective is that, a word can be highly relevant for a topic if the frequency of occurrence of that word in the topic was high and the other perspective is that, a word can be highly relevant for a topic if the ratio of probability of occurrence of word in the selected topic and the marginal probability of the word in the corpus is high. This ratio is referred as ‘Lift’ and from Figure 37, it can be calculated as the ratio of the length of the red bar and the length of the blue bar. When λ is set to one the words appear in the descending order of their frequency and when λ is set to zero they appear in the descending order of their lift values. Sievert and Shirley (2014) found out that setting λ to 0.6 yielded the best results. Therefore, to obtain highly relevant terms for each topic λ was set to 0.6.

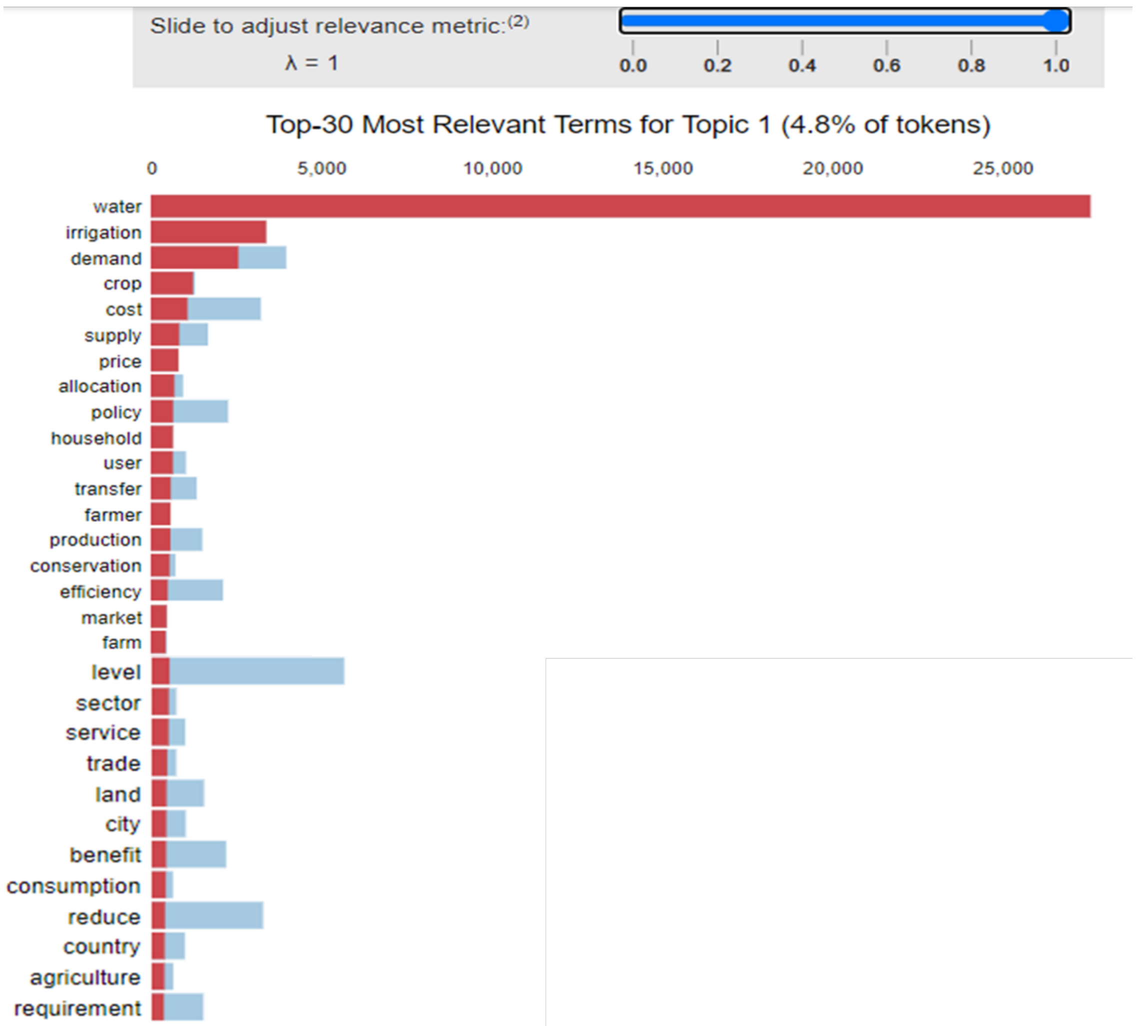


Figure 37: Top 30 Most Relevant Terms for Topic 1 ($\lambda = 1$)

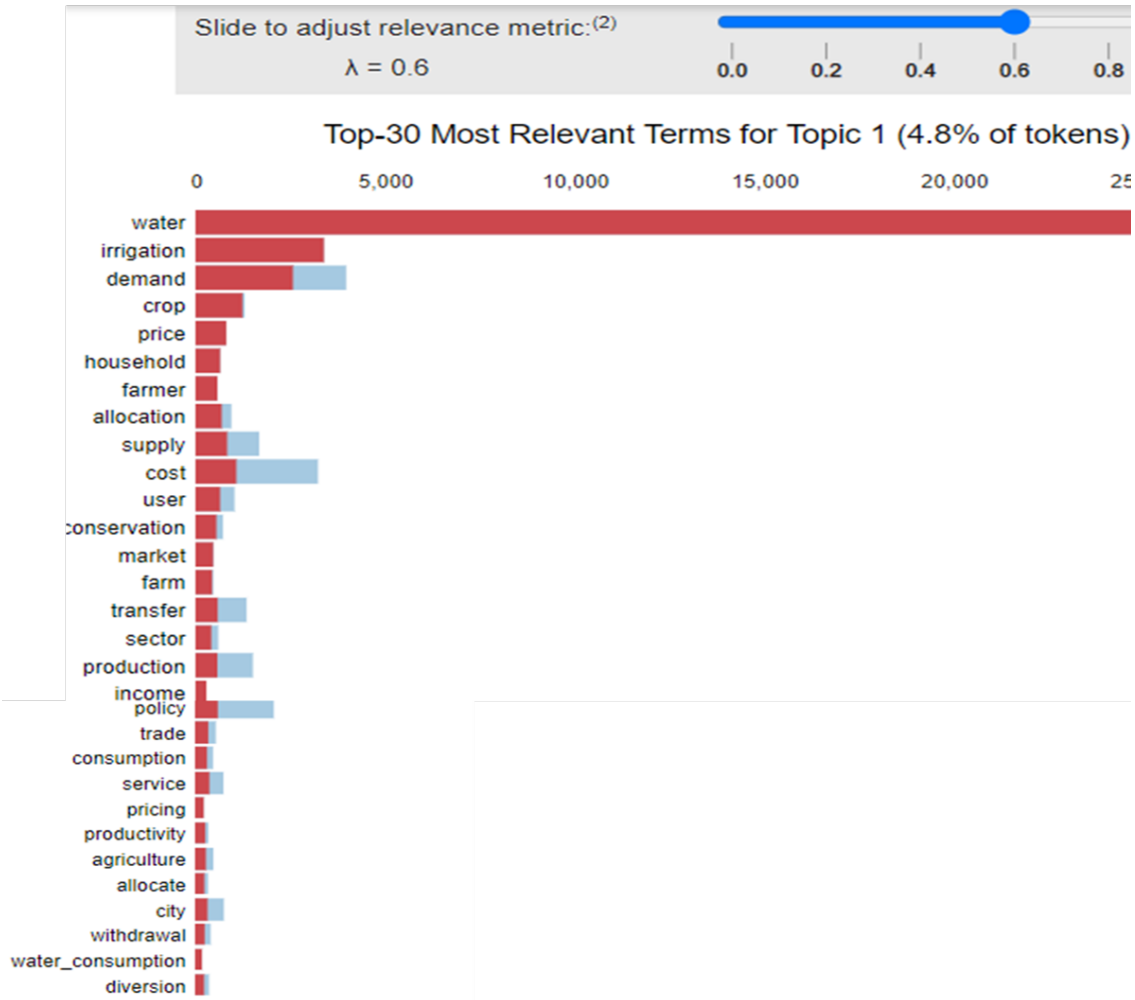


Figure 38: Top 30 Most Relevant Terms for Topic 1 ($\lambda = 0.6$)

By analysing the words for each topic shown in Figure 31 and the relevant words from the pyLDAvis visualization, the following qualitative analysis was performed for each corpus. For the topics obtained by corpus prepared using abstracts and filtered using TF-IDF approach, Topic 1 is represented by words 'water', 'irrigation', 'demand', 'crop', 'farmer', 'allocation', 'supply', 'water_consumption' *etc.* which suggests that the theme is '***Irrigation***'. Topic 2 incorporates 'removal', 'treatment', 'wastewater', 'sludge', 'reactor', 'adsorption', 'degradation', 'bacteria' highlights wastewater treatment. Thus, topic 2 can be labelled as '***Wastewater Treatment***'. Topic 3 corresponds to '***Groundwaters***' due to the dominating terms, 'groundwater', 'aquifer', 'recharge', 'salinity', 'leakage', 'piezometer', 'pumping_test'. Topic 4 is devoted to optimization techniques of water distribution network. The terms 'system', 'reservoir', 'pipe', 'optimization', 'reservoir_operation', 'network', 'pressure', 'energy' indicate towards optimizing the multi-objective functions of water distribution or reservoir systems such as pipe cost, pipe pressure *etc*. Topic 4 discusses an important part of IWRM which is about making the water distribution systems reliable and efficient. Hence, Topic 4 can be labelled as '***Optimization of Reservoir Operations***'

Topic 5 describes the coordination of science, regulation and management needed to more effectively achieve a common goal of managing chemicals on our planet. The terms 'community', 'chemical', 'toxicity', 'exposure', 'population', 'fish', 'research', 'development', 'risk' indicates the integration of chemicals into the food chain. Therefore, Topic 5 can be labelled as '***Aquatic Toxicology***'. Topic 6 is about modelling the flow of water. The terms 'distribution', 'time', 'function', 'derive', 'variance', 'coefficient', 'calculate', 'relationship' are indicating towards estimating the flow of water with the help of statistical analysis. Therefore, Topic 6 can be labelled as '***Mathematical Modelling***'. Topic 7 is about modelling sediment deposition and formation. It also discusses the flow structure and morphodynamics. This conclusion can be derived by analyzing the terms 'flow', 'velocity', 'stress', 'sediment', 'rate', 'transport', 'slope', 'bed'. Topic 7 is a crucial part of IWRM as sediment deposition and the flow structure of the channels affect the economic and social dimensions of IWRM. Topic 7 can be labelled as '***Stream Hydrological Modelling***'. Topic 8 discusses the measurement of snow cover as well as vegetation cover on the land. The terms 'image', 'measurement', 'sensor', 'site', 'snow' indicate the theme of capturing descriptive properties of the snow via sensors preferable aerial. Meanwhile the terms 'difference', 'vegetation', 'forest' discusses about the land cover and the difference in the properties of the land. Topic 8 can be labelled as '***Forest Hydrology***'. The terms 'transport', 'concentration', 'solute', 'tracer', 'particle', 'dispersion', 'rate', 'mix', 'plume', 'velocity', 'sorption', 'column', 'spread' indicate that the articles were discussing topics pertaining to the diffusion of a fluid, its properties and how it changes *etc.* Topic 9 discusses the physical phenomenon such as diffusion and adsorption observed in natural processes and hence, Topic 9 can be labelled as '***Aquatic Mesocosms***'. The terms such as 'management', 'scenario', 'risk', 'project', 'development', 'impact', 'planning', 'assessment', 'decision', 'policy' highlights the concept of the risk management. The terms like 'tool', 'index', 'indicator', 'level' further indicates quantification of risk. Topic 10 is about identifying and mitigating the risks involved in complex projects. This also captures an important component of Integrated Environment Assessment. Thus, Topic 10 can be labelled as '***Prediction of extreme hydrological***

events'. The terms in Topic 11 such as 'soil', 'surface', 'flux', 'layer', 'root', 'measurement', 'evaporation', 'infiltration', 'conductivity', 'soil_texture', 'heat', 'temperature', 'drainage', 'depth' indicate towards a theme about hydraulic properties of the soil. Topic 11 can be labelled as '***Infiltration***'. Terms such as 'fracture', 'permeability', 'field', 'pressure', 'interface', 'rock', 'pore', 'porosity', 'experiment', 'density' highlight that the underlying theme would have been analysing the permeability of soil and performing different experiments to record the value of porosity for different mediums. Thus, topic 12 can be labelled as '***Water Saturation***'. Topic 13 contain terms such as 'model', 'simulation', 'prediction', 'forecast', 'input', 'predict', 'forecasting', 'develop', 'parameter', 'performance' which collectively conveys the theme of forecasting/time-series models. Thus, topic 13 can be labelled as '***Predictive Simulation Modelling***'. Topic 14 contain terms such as 'lake', 'climate', 'change', 'trend', 'impact', 'temperature', 'response', 'reduce', 'level', 'affect', 'response'. These terms in conjunction tries to convey the idea of what parameters affect the level of water in water-bodies and how sensitive the water-bodies are to climate change parameters. Therefore, topic 14 can be labelled as '***Climate Change Impact Assessment***'. The terms 'process', 'scale', 'pattern', 'structure', 'dynamic', 'variability', 'network', 'control', 'behaviour' and 'influence' indicate that topic 15 can be labelled as '***Scale Issues in Hydrology***'. Topic 16 contain terms such as 'lake', 'watershed', 'stream', 'site', 'discharge', 'pollution', 'sediment', 'reduction', 'nitrogen', 'quality', 'carbon'. These terms indicate towards the pollution of the waterbodies. Topic 16 can be attributed with water pollution and identifying the sources of pollution. Topic 16 can be labelled as '***Lake Organic Pollution***'. Topic 17 contains terms like 'method', 'estimate', 'parameter', ‘error’, ‘estimation’, ‘test’, ‘procedure’, ‘observation’, ‘sampling’, ‘accuracy’, ‘technique’, ‘carbon’. Topic 17 can be labelled as ‘***Monitoring***’. Topic 18 contains terms like ‘solution’, ‘problem’, ‘equation’, ‘formulation’, ‘solve’, ‘algorithm’, ‘technique’, ‘simulation’. Topic 18 can be labelled as ‘***Mathematical Techniques***’. Topic 19 contains terms such as ‘rainfall’, ‘runoff’, ‘catchment’, ‘precipitation’, ‘flood’, ‘drought’, ‘storm’, ‘record’, ‘station’, ‘climate’, ‘variability’. Therefore, topic 19 can be labelled as ‘***Watershed Hydrology***’. Topic 20 contain terms like ‘uncertainty’, ‘analysis’, ‘application’, ‘information’, ‘technique’, ‘include’, ‘exist’, ‘framework’, ‘design’. Thus, Topic 20 can be labelled as ‘***Methodological Approaches***’. The topics are shown in Figure 39. Similarly, the significant words for the topics obtained by the LDA algorithm trained only on abstracts BoW representation are shown in Figure 40 .

Topic No.	Significant Words Explaining Topic Label	Topic Label
1	water irrigation demand crop farmer allocation water_consumption supply	***Irrigation***
2	removal treatment wastewater sludge reactor bacteria adsorption degradation	***Wastewater Treatment***
3	groundwater aquifer recharge salinity leakage piezometer pumping_test	***Groundwaters***
4	system reservoir pipe optimization reservoir_operation network pressure energy	***Optimization of Reservoir Operations***
5	community chemical toxicity exposure population fish research development risk	***Community Risk Assessment of Water Toxicity***
6	distribution time function derive variance coefficient calculate relationship	***Mathematical Modelling***
7	flow velocity stress sediment rate transport slope bed	***Stream Hydrological Modelling***
8	image measurement sensor site snow vegetation difference forest	***Forest Hydrology***
9	transport concentration solute tracer particle dispersion rate mix plume velocity sorption column spread	***Aquatic Mesocosms***
10	management scenario risk project development impact planning assessment decision policy tool index indicator level	***Prediction of extreme hydrological events***
11	soil surface flux layer root measurement evaporation infiltration conductivity soil_texture heat temperature drainage depth	***Infiltration***
12	fracture permeability field pressure interface rock pore porosity experiment density	***Water Saturation***
13	model simulation prediction forecast input predict forecasting develop parameter performance	***Predictive Simulation Modelling***
14	climate lake change trend impact temperature response reduce level affect response	***Climate Change Impact Assessment***
15	process scale pattern structure dynamic variability network control influence behaviour	***Scale Issues in Hydrology***
16	lake watershed stream site discharge pollution sediment reduction nitrogen quality carbon	***Lake Organic Pollution***
17	method estimate parameter error estimation test procedure observation sampling accuracy technique carbon	***Monitoring***
18	solution problem equation formulation solve algorithm technique simulation	***Mathematical Techniques***
19	rainfall runoff catchment precipitation flood drought storm record station climate variability	***Watershed Hydrology***
20	uncertainty analysis application information technique include exist framework design	***Methodological Approaches***

Figure 39: Significant Words for each Topic Label for Abstracts TF-IDF

Topic No.	Significant Words Explaining Topic Label	Topic Label
1	soil fracture infiltration saturation permeability pore drainage retention	***Infiltration and Saturation***
2	method base problem approach technique solution procedure formulation result solve	***Methodological Approaches***
3	measure site sample test field datum detection conduct sensor monitoring sampling	***Monitoring***
4	estimate datum uncertainty estimation error observation data bias estimator method reserve measurement union_right provide	***Data Processing Techniques***
5	catchment stream runoff dynamic vegetation hillslope landscape topography variability watershed pattern	***Watershed Hydrology***
6	system process approach provide analysis include development research information knowledge framework	***Methodological Approaches***
7	distribution function scale parameter show derive relationship property correlation variability	***Statistical Analysis***
8	water demand irrigation level system management supply crop agriculture production	***Irrigation***
9	risk exposure contamination concentration chemical risk_assessment fish toxicity pollutant pesticide	***Risk Assessment of Pollution***
10	policy project benefit management conflict community infrastructure regulation price	***Water Related Project Management***
11	transport flow equation solute concentration hydraulic_conductivity coefficient plume medium maximize inflow	***Mesocosm for Model Calibration***
12	system reservoir network design optimization operation cost constraint reliability performance control pipe pressure uncertainty develop	***Reservoir Operation Management***
13	rainfall precipitation flood year drought period region event climate index runoff storm station record trend	***Prediction of Droughts***
14	temperature lake evaporation snow flux transpiration heat radiation energy atmosphere depth	***Lake Water Balance***
15	model simulation base simulate parameter prediction result develop performance datum predict forecast input calibration apply	***Simulation Modelling***
16	concentration treatment removal process wastewater system sludge reactor adsorption remove bacteria denitrification	***Wastewater Treatment***
17	rate condition time result process effect phase experiment show observe colloid dissolution equilibrium determine	***Experimental Hydrology***
18	flow river sediment channel discharge bed stream velocity reach turbulence transport wave slope	***Hydrodynamics***
19	increase change effect impact result decrease study condition reduce reduction scenario land_use	***Land Use Change Impact***
20	groundwater aquifer flow zone recharge system salinity salt head flow confined_aquifer	***Ground waters***

Figure 40: Significant Words for each Topic Label for Abstracts BoW

The topics are shown in Figure 39. Similarly, the significant words for the topics obtained by the LDA algorithm trained only on abstracts BoW representation are shown in Figure 40

6.4.3 Qualitative Analysis of Topics obtained from corpus prepared using Full-Length Research Papers

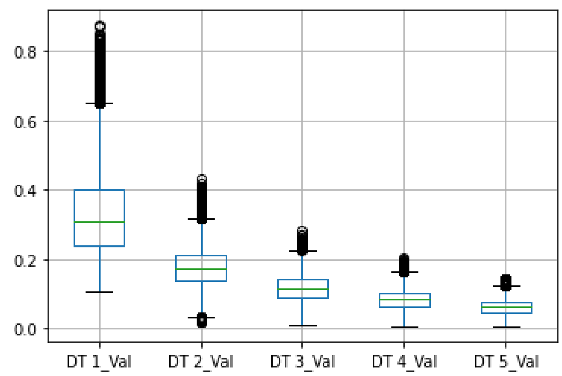
The significant words for topics obtained from full-length research papers were determined using the same methodology followed in Section 6.4.2 and are shown in Figure 41. Consider ‘Topic 10 (Prediction of extreme hydrological events)’ in Figure 39 and ‘Topic 13 (Prediction of extreme hydrological events)’ in Figure 40. Both these topics indicated that the research papers had content related to extreme hydrological events, however, it is difficult to obtain any other extra information only by reading these labels. Similarly, consider ‘Topic 3 (Flood Predictions)’, ‘Topic 16 (Precipitations)’ and ‘Topic 17 (Simulation Modelling)’ in Figure 41. These three topics not only indicate that the research papers had a central theme related to hydrological modelling but also further subdivides the area into separate themes such as flood prediction, role of precipitation in hydrological cycle and simulation modelling of hydrological processes. These concepts are associated with hydrological events, thus providing extra information to the researcher. Similarly, consider ‘Topic 12 (Water Allocation)’ in Figure 41. The concept of ‘Water Allocation’ is an integral component of IWRM domain, however, this theme was not observed in any of the topics obtained from abstracts (Abstracts BoW and Abstracts TF-IDF). This highlights that integration of full-length research articles helps to identify the breadth as well as depth of the domain under consideration.

Topic No.	Significant Words Explaining Topic Label	Topic Label
1	flow equation solution field function velocity variance head approximation coefficient boundary_condition dispersion	***Hydrodynamics***
2	fracture particle permeability transport flow test velocity pressure injection simulation porosity borehole point rock	***Sediments***
3	flood distribution level probability variable flood cluster variance event correlation series frequency estimator estimation generate	***Flood Predictions***
4	concentration rate transport source reaction stream adsorption mix column load solute ion chloride nitrogen equilibrium	***Hydrochemistry***
5	cost optimization design solution system constraint problem scenario management uncertainty fig pump plan objective objective_function reliability	***Optimisation of Water Distribution***
6	exposure specie chemical toxicity fish pesticide substance concentration plastic expose population	***Aquatoxicology***
7	reservoir storage period operation release inflow volume stage dam demand rule energy fig capacity month	***Hydropower Management***
8	surface temperature snow lake soil_moisture evaporation depth vegetation heat flux canopy layer day wind	***Snow Hydrology***
9	groundwater area recharge aquifer zone table depth head age pump seepage piezometer factor wetland unit	***Groundwater***
10	site sample measurement pattern location sampling group probe monitoring monitor plot survey sensor sampler frequency	***Water Monitoring***
11	network pipe node demand pressure control valve distribution_network method tank leakage failure link pump	***Water Distribution Systems***
12	water demand cost price supply household consumption allocation transfer user benefit market sector utility policy	***Water Allocation***
13	management project policy criterion state risk indicator plan quality program water_quality community decision development resource	***Water Quality Management***
14	soil depth saturation phase infiltration pressure air layer gas experiment drainage profile water_content root capillary_pressure	***Water Infiltration***
15	treatment concentration removal wastewater sludge reactor plant cod bacteria membrane day solid cell substrate biomass	***Wastewater Treatment***
16	precipitation year climate region station drought basin change trend period month temperature rainfall index streamflow	***Precipitations***
17	model parameter estimate error uncertainty calibration prediction input datum observation simulation performance forecast estimation output	***Simulation Modelling***
18	irrigation scenario crop basin impact land farmer yield salinity land_use farm production reduction canal fig	***Irrigation***
19	discharge stream river flow sediment flow bed velocity channel reach depth transport stress slope run	***Water Constituents Transport***
20	rainfall catchment runoff event storm scale response soil slope watershed basin hillslope peak forest rain	***Watershed Modelling***

Figure 41: Significant Words for each Topic Label for Full-Text TF-IDF Corpus

6.4.4 Qualitative Comparison of Topics obtained using Abstracts and Full-Length Research Papers

To further determine the distribution of topics in the two corpus *i.e.* corpus prepared using only abstracts and filtered using TF-IDF approach and the corpus prepared using full-length research papers filtered using TF-IDF approach, the following analysis was conducted. For each document the contribution of topics were recorded. If the contributions are nearly identical then it becomes difficult to determine which topic label correctly describes the document. On the contrary, if there is a clear separation between topic proportions then the document could be correctly labelled. Boxplot of topic distribution in corpus prepared using only abstracts and filtered using TF-IDF approach (Figure 42) revealed that for 50% of the research papers, the most dominating topic had less than 30% contribution and the second most dominating topic had a contribution of 17%. Identical distribution of topic contributions makes it difficult to identify the label to which the document would belong to. This shows that abstracts showed over representation of different topics. Boxplot of topic distribution in corpus prepared using full-length research papers filtered using TF-IDF approach (Figure 43) revealed that for 50% of the research papers, the most dominating topic had a contribution of over 55% and the second most dominating topic had a contribution of 20% which shows there is high topic separation.

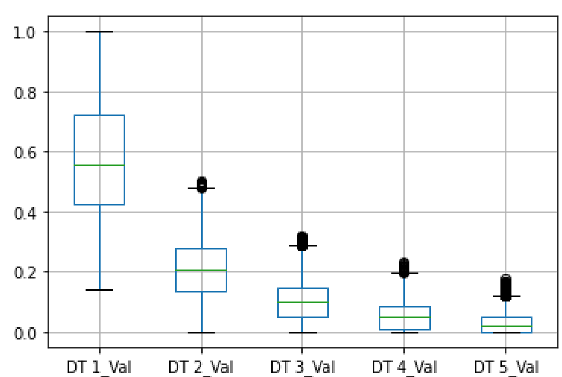


(a) Topic Distribution in Abstracts Filtered using TF-IDF

	DT 1_Val	DT 2_Val	DT 3_Val	DT 4_Val	DT 5_Val
mean	0.332792	0.177999	0.117784	0.084678	0.062141
std	0.130368	0.053283	0.038202	0.029007	0.023426
min	0.108055	0.018287	0.008786	0.007028	0.005872
25%	0.238543	0.139969	0.087915	0.064944	0.045753
50%	0.309840	0.173308	0.117420	0.083001	0.061458
75%	0.403384	0.211032	0.143891	0.104563	0.077796
max	0.875222	0.431782	0.281282	0.202264	0.143593

(b) Descriptive Statistics for Top 5 Topic Proportions

Figure 42: Topic Distribution in Corpus Prepared using Abstracts and TF-IDF Filtration



(a) Topic Distribution in Full-Length Research Papers

	DT 1_Val	DT 2_Val	DT 3_Val	DT 4_Val	DT 5_Val
mean	0.582503	0.207974	0.100575	0.054046	0.028319
std	0.196201	0.099976	0.063662	0.045374	0.032065
min	0.139884	0.000014	0.000003	0.000003	0.000003
25%	0.425769	0.138484	0.051871	0.010328	0.000150
50%	0.557708	0.206894	0.098810	0.049156	0.017966
75%	0.724683	0.276077	0.145842	0.085945	0.049097
max	0.999837	0.499045	0.317091	0.235139	0.175641

(b) Descriptive Statistics for Top 5 Topic Proportions

Figure 43: Topic Distribution in Corpus Prepared using Full-Length Research Papers and TF-IDF Filtration

To further demonstrate the above findings, consider the article '*Food-processing wastes*'(Frenkel et al., 2020). The abstract for this article is 'Literature published in 2018 and literature published in 2019 related to food-processing wastes treatment for industrial applications are reviewed. This review is a subsection of the Treatment Systems section of the annual Water Environment Federation literature review and covers the following food-processing industries and applications: general, meat and poultry, fruits and vegetables, dairy and beverage, and miscellaneous treatment of food wastes. ' (Frenkel et al., 2020). The topic distribution for this research paper only on the basis of the abstract is shown in Table 4.

Table 4: Distribution of Top 3 Topics (Abstract Topics (TF-IDF))

Dominant Topic	Topic Label	Topic Contribution
5	***Aquatic Toxicology***	0.3886932064852141
2	***Wastewater Treatment***	0.3856861497552013
20	***Methodological Approaches***	0.10721089269965177

The abstract of the research paper indicates that the article must be about treating the waste produced by the food processing industry and it is expected that the abstract topic model would return high proportion for Topic 2 *i.e.* 'Wastewater Treatment'. However, Topic 2 only had a representation of 0.38 which shows that the model was not able to correctly represent the latent theme observed in the article. The topic distribution for the same article obtained by the LDA algorithm trained on full-length research papers is shown in Table 5. Topic 15 which is about 'Wastewater Treatment' was returned as the most dominant topic with a contribution of over 0.99. This highlights that the full-length research papers helped the LDA algorithm to return better and relevant topics for documents where the abstract was too broad. This highlights that integration of full-length research papers can help in better labelling the documents for further downstream tasks such as classification and information retrieval.

Table 5: Distribution of Top 3 Topics (Full-Length Research Papers)

Dominant Topic	Topic Label	Topic Contribution
15	***Wastewater Treatment***	0.9991686693840754
1	***Hydrodynamics***	$7.047840662612521 \times 10^{-5}$
3	***Flood Predictions***	$5.4557013527051996 \times 10^{-5}$

7 Conclusion

The suggested topic modelling framework can be applied to any research domain in general and should be used specifically, for a multidisciplinary domain like IWRM. Moreover, previous studies adopted a keyword search approach to selecting relevant research papers. This often led to under-representation of an entire domain. To provide full coverage of the domain, the study has adopted a journal-based approach which eliminated biases towards certain topics representation in the corpus. Journals which scope covers IWRM were identified. The initial exploration of the scientific journals revealed 14 relevant outlets. However, due to the unavailability of full-length papers from a few journals, the analysis was conducted on 7 journals.

The selected domain of IWRM is a complex multidisciplinary research area and to ensure complete representation of publications related to various disciplines, all research papers from the selected journals were considered. Since, generation of topics by the LDA algorithm is a probabilistic process, different topics can be obtained from the same corpus even if all algorithm's hyperparameters have the same values. Therefore, to prevent occurrence of different topics every time the experiment is run, the random state hyperparameter was set at the beginning. This ensured that the set of topics for a specific corpus generated by the LDA algorithm remains the same.

Previous studies did not consider full-length research papers. Inclusion of full-length papers in the corpus makes the current analysis to be the first study on an in-depth exploration of IWRM domain.

The labels that were assigned to the topics identified in the corpus are subjective. Inputs from the domain expert helped to enhance label accuracy. However, further investigation might be required to determine whether the deduced labels correctly capture the underlying theme.

In summary, the study investigated approaches to building corpora for topic modelling analysis in a multidisciplinary domain. The analysis was conducted using text representation based on word frequency and word distribution approaches. The conventional topic modelling framework was modified to improve the quality of generated topics which was reflected in a higher coherence score. The results of the study led to the following conclusions.

TF-IDF filtration increased the coherence score of the obtained topics. Inclusion of the full-length research papers in a corpus not only improved the coherence score but also revealed new themes which could not have been obtained using corpus of abstracts of the same papers.

Topics generated from full-length paper corpus can support better labelling of the documents aiding information retrieval process. The conducted analysis required automation of the document extraction and preprocessing to eliminate the time and labour-consuming manual paper downloading and transformation process. To automate corpus preparation, a utility was developed in Python. Over 29,000 full-length research papers were downloaded and used for topic modelling.

The topic modelling analysis was conducted for IWRM domain which has uncontested priority in achieving UN Sustainable Development Goals. The enhanced topic modelling framework helped to identify dominating topics on IWRM in leading scientific peer-reviewed journals for the last fifty years. The analysis demonstrated that the number of publications in this domain increased steadily reflecting growing interest to the IWRM problems among researchers and practitioners. However, the main research themes of IWRM remain within discipline areas with very few publications addressing integration of knowledge from natural sciences with economic and social analysis. The majority of publications fit the area of natural sciences and engineering, with fewer studies on management and economic issues. The analysis clearly demonstrated that the gap in integration of social components and water related policy assessment into water resource management at all levels of governance remain wide and calls for intensification of research in this area.

8 References

Abdel-Magid, I. and Ahmed, S. (2002). Integrated water resources management and global water partnership.

Aggarwal, C. C. (2018). *Machine Learning for Text*. Springer Publishing Company, Incorporated, 1st edition.

Aletras, N. and Stevenson, M. (2013). Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*, pages 13–22.

Balkema, A. J., Preisig, H. A., Otterpohl, R., and Lambert, F. J. (2002). Indicators for the sustainability assessment of wastewater treatment systems. *Urban water*, 4(2):153–161.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Blei, D. M., Jordan, M. I., et al. (2006). Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143.

Blei, D. M. and McAuliffe, J. D. (2010). Supervised topic models. *arXiv preprint arXiv:1003.0783*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Bonilla, T. and Grimmer, J. (2013). Elevated threat levels and decreased expectations: How democracy handles terrorist threats. *Poetics*, 41:650–669.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40.

Bryan, K., Cunningham, P., and Bolshakova, N. (2006). Application of simulated annealing to the biclustering of gene expression data. *IEEE transactions on information technology in biomedicine*, 10(3):519–525.

Buckley, C., Allan, J., and Salton, G. (1994). Automatic routing and ad-hoc retrieval using smart: Trec 2. *NIST SPECIAL PUBLICATION SP*, pages 45–45.

Cai, D., He, X., Wu, X., and Han, J. (2008). Non-negative matrix factorization on manifold. In *2008 Eighth IEEE International Conference on Data Mining*, pages 63–72. IEEE.

Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., and Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Neural information processing systems*, volume 22, pages 288–296. Citeseer.

Cheng, X., Shuai, C., Liu, J., Wang, J., Liu, Y., Li, W., and Shuai, J. (2018). Topic modelling of ecology, environment and poverty nexus: An integrated framework. *Agriculture, Ecosystems & Environment*, 267:1–14.

Conservancy, S. F. (2021). selenium. https://pypi.org/project/selenium, accessed 14. Apr. 2021.

Daume, S., Albert, M., and von Gadow, K. (2014). Assessing citizen science opportunities in forest monitoring using probabilistic topic modelling. *Forest Ecosystems*, 1(1):1–12.

De Saussure, F. (2011). *Course in general linguistics*. Columbia University Press.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

DepartmentofEnvironmentalProtection (2021). Ecological assessments. https://dep.wv.gov/WWE/watershed/wqmonitoring/Pages/EcologicalAssessments.aspx; accessed 1. Feb. 2021.

DiMaggio, P., Nag, M., and Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of u.s. government arts funding. *Poetics*, 41:570–606.

Erechtchoukova, M. and Khaiter, P. (2007). Data-model issues in environmental impact assessment. *The International Journal of Environmental, Cultural, Economic, and Social Sustainability: Annual Review*, 3:149–156.

Erechtchoukova, M. and Khaiter, P. (2009). Quantifying model uncertainty in environmental impact assessment. *The International Journal of Environmental, Cultural, Economic, and Social Sustainability: Annual Review*, 5:65–72.

Erechtchoukova, M. G. and Khaiter, P. (2011). A model-driven approach to uncertainty reduction in environmental data. In *ITEE*.

Foxon, T., Mcilkenny, G., Gilmour, D., Oltean-Dumbrava, C., Souter, N., Ashley, R., Butler, D., Pearson, P., Jowitt, P., and Moir, J. (2002). Sustainability criteria for decision support in the uk water industry. *Journal of Environmental Planning and Management*, 45:285–301.

Frenkel, V. S., Cummings, G. A., Maillacheruvu, K. Y., and Tang, W. Z. (2020). Food-processing wastes. *Water Environment Research*, 92(10):1726–1740.

Ghosh, D. and Guha, R. (2013). What are we 'tweeting' about obesity? mapping tweets with topic modeling and geographic information system. *Cartography and geographic information science*, 40:90–102.

Glasson, J. and Wood, G. (2009). Urban regeneration and impact assessment for social sustainability. *Impact Assessment and Project Appraisal - Impact Assess Proj Apprais*, 27:283–290.

Global, S. (2021). Valuing water to drive more effective decisions | Trucost. https://www.trucost.com/trucost-news/valuing-water-drive-effective-decisions; accessed 1. Feb. 2021.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

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. Citeseer.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, page 50–57, New York, NY, USA. Association for Computing Machinery.

Hofmann, T. (2013). Probabilistic latent semantic analysis. *arXiv preprint arXiv:1301.6705*.

InternationalDOIFoundation (2021). Digital Object Identifier System. https://www.doi.org; accessed 17. May 2021.

Jiang, H., Qiang, M., and Lin, P. (2016). A topic modeling based bibliometric exploration of hydropower research. *Renewable and Sustainable Energy Reviews*, 57:226–237.

Jivani, A. G. et al. (2011). A comparative study of stemming algorithms. *Int. J. Comp. Tech. Appl*, 2(6):1930–1938.

Jockers, M. and Mimno, D. (2013). Significant themes in 19th-century literature. *Poetics*, 41:750–769.

Jones, K. S. (1988). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 60:493–502.

Joshi, P. (2020). Topic Modelling In Python Using Latent Semantic Analysis. https://www.analyticsvidhya.com/blog/2018/10/stepwise-guide-topic-modeling-latent-semantic-analysis; accessed 6. Apr. 2021.

Kherwa, P. and Bansal, P. (2017). Latent semantic analysis: An approach to understand semantic of text. In *2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC)*, pages 870–874. IEEE.

Kherwa, P. and Bansal, P. (2020). Topic modeling: a comprehensive review. *EAI Endorsed transactions on scalable information systems*, 7(24).

Kontoghiorghes, E. J. (2005). *Handbook of parallel computing and statistics*. CRC Press.

Kreutz-Delgado, K., Murray, J. F., Rao, B. D., Engan, K., Lee, T.-W., and Sejnowski, T. J. (2003). Dictionary learning algorithms for sparse representation. *Neural computation*, 15(2):349–396.

Kumar, A., Choukimath, P. A., et al. (2015). Popular scientometric analysis, mapping and visualisation softwares: An overview.

Kwon, H., Kim, J., and Park, Y. (2017). Applying lsa text mining technique in envisioning social impacts of emerging technologies: The case of drone technology. *Technovation*, 60:15–28.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.

Lensu, A. (2002). Computationally intelligent methods for qualitative data analysis.

Li, W. and Zhao, Y. (2015). Bibliometric analysis of global environmental assessment research in a 20-year period. *Environmental Impact Assessment Review*, 50:158–166.

Liu, Y., Guo, H., Yu, Y., Dai, Y., and Zhou, F. (2008). Ecological–economic modeling as a tool for watershed management: A case study of lake qionghai watershed, china. *Limnologica - Ecology and Management of Inland Waters*, 38:89–104.

Loucks, D. P. and Gladwell, J. S. (1999). *Sustainability criteria for water resource systems*. Cambridge University Press.

Mabey, B. (2021). pyLDAvis. https://github.com/bmabey/pyLDAvis; accessed 28. Apr. 2021.

Marvuglia, A., Havinga, L., Heidrich, O., Fonseca, J., Gaitani, N., and Reckien, D. (2020). Advances and challenges in assessing urban sustainability: An advanced bibliometric review. *Renewable and Sustainable Energy Reviews*, 124:109788.

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Minx, J. C., Lamb, W. F., Callaghan, M. W., Bornmann, L., and Fuss, S. (2017). Fast growing research on negative emissions. *Environmental Research Letters*, 12(3):035007.

Mohr, J. W. and Bogdanov, P. (2013). Introduction—topic models: What they are and why they matter.

Muita, S. et al. (2020). *A Model for processing public participation feedback using topic Modeling*. PhD thesis, University of Nairobi.

Muñoz, E. (2021). Getting started with NLP: Tokenization, Document-Term Matrix, TF-IDF. *Medium*.

Narke, S. P. (2017). *A Study of Different Pre-Processing Approaches of Text Categorization*. PhD thesis, Dublin, National College of Ireland.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pages 100–108.

Newman, D., Smyth, P., Welling, M., and Asuncion, A. (2007). Distributed inference for latent dirichlet allocation. *Advances in neural information processing systems*, 20:1081–1088.

Nigam, K., Mccallum, A. K., Thrun, S., and Mitchell, T. (2000). Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, 39(2):103–134.

Paatero, P. (1997). Least squares formulation of robust non-negative factor analysis. *Chemometrics and intelligent laboratory systems*, 37(1):23–35.

Paatero, P. and Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126.

Ramasubramanian, C. and Ramya, R. (2013). Effective pre-processing activities in text mining using improved porter's stemming algorithm.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Reitz, K. (2021). Requests module. https://pypi.org/project/requests; accessed 17. May 2021.

Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408.

Sathi, V. R. and Ramanujapura, J. S. (2016). A quality criteria based evaluation of topic models.

Shinyaman, Y. (2021). pdfminer.six. https://github.com/pdfminer/pdfminer.six; accessed 11. Apr. 2021.

Sievert, C. and Shirley, K. (2014). Ldavis: A method for visualizing and interpreting topic. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 63–70, Baltimore, Maryland, USA. Association for Computational Linguistics. https://www.aclweb.org/anthology/W14-3110", doi = "10.3115/v1/W14-3110.

Singh, J. and Gupta, V. (2016). Text stemming: Approaches, applications, and challenges. *ACM Comput. Surv.*, 49(3). https://doi.org/10.1145/2975608.

Smith, M. and Jønch Clausen, T. (2015). Integrated water resource management: A new way forward. *World Water Council, Marseille*.

Statistics, D. U. (2012). *System of Environmental-Economic Accounting for Water*. UN,. https://digitallibrary.un.org/record/728076?ln=en.

Syed, S. and Spruit, M. (2017). Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 165–174.

Thornton, K. W., Laurin, C., Shortle, J., Fisher, A., Sobrinho, J., and Stewart, M. (2006). A framework and guidelines for moving toward sustainable water resources management. *Proceedings of the Water Environment Federation*, 2006(10):2762–2777.

Tol, R. S. and Vellinga, P. (1998). The european forum on integrated environmental assessment. *Environmental Modeling & Assessment*, 3(3):181–191.

Toth, F. L. and Hizsnyik, E. (1998). Integrated environmental assessment methods: Evolution and applications. *Environmental Modeling & Assessment*, 3(3):193–207.

van der Zaag, P. and Gupta, J. (2008). Scale issues in the governance of water storage projects. *Water Resources Research, 44, 2008 ; doi:10.1029/2007WR006364*, 44.

Wallach, H. M. (2006). Topic modeling: Beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 977–984, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/1143844.1143967.

Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning*, pages 1105–1112.

Wang, M.-H., Yu, T.-C., and Ho, Y.-S. (2010). A bibliometric analysis of the performance of water research. *Scientometrics*, 84(3):813–820.

Weyant, J., Davidson, O., Dowlatabadi, H., Edmonds, J., Grubb, M., Parson, E., Richels, R., Rotmans, J., Shukla, P., Tol, R. S., et al. (1995). Integrated assessment of climate change: an overview and comparison of approaches and results. *Climate change*, 3.

Wilkinson, L. J. (2020). Text and data mining for researchers - Crossref. https://www.crossref.org/education/retrieve-metadata/rest-api/text-and-data-mining-for-researchers; accessed 13. Apr. 2021.

Yao, L., Mimno, D., and McCallum, A. (2009). Efficient methods for topic model inference on streaming document collections. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 937–946.

Zin, H. M., Mustapha, N., Murad, M. A. A., and Sharef, N. M. (2017). The effects of pre-processing strategies in sentiment analysis of online movie reviews. In *AIP conference proceedings*, volume 1891, page 020089. AIP Publishing LLC.

9 Appendix

9.1 Appendix A: Code Block for Downloading Non member Publications

The following code block drives the downloading of the full-text PDF links of the articles retrieved from the Scopus database

```
import time
import requests
import time
from selenium import webdriver
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from selenium.webdriver.common.by import By
from selenium.common.exceptions import TimeoutException
from selenium.webdriver import ActionChains
from selenium.webdriver.chrome.options import Options
from pyvirtualdisplay import Display
import os
import pandas as pd
import test_logic_sample
import processing_doc
```

```
opts = Options()

opts.add_argument('--profile-directory=Profile 2')
opts.add_argument('--window-size=1280,800')
prefs =
{"profile.default_content_settings.popups": 0,

"download.default_directory": m,
"directory_upgrade": True}
opts.add_experimental_option("prefs", prefs)

opts.add_argument("--window-size=1920,1080");

def find_link(doi):
    driver = webdriver.Chrome(
    executable_path=
    r'C:\Users\aksha\Downloads\chromedriver_win32 (2)
    \chromedriver.exe',
    options=opts)
```

```
doi = doi
path = 'http://dx.doi.org/{}'.format(doi)
try:

    driver.get(path)

    WebDriverWait(driver, 20).until(
    EC.element_to_be_clickable(
    (By.XPATH,
    "//span[@class=
    'pdf-download-label u-show-inline-from-lg']"
    )))
    .click()



    WebDriverWait(driver, 20).until(
    EC.visibility_of_element_located(
    (By.XPATH,
    "//a[@class=
    'link-button u-margin-s-bottom link-button-primary']")))
    button = driver.find_element_by_xpath(
```

```
            "//a[@class=
            'link-button u-margin-s-bottom link-button-primary']"
            )

            link = button.get_attribute('href')
            print("Link")
            print(link)
            driver.close()
            return link

        except Exception as e:

            print(e)
            return "Link not found"

            driver.close()
    t1 = time.perf_counter()
```

```
if __name__=='__main__':

    non_member_publications=[
        "Environmental Impact Assessment Review"
        ,"Journal of Environmental Management"
        ,"Water Research"
        ,"Advances in Water Resources"
        ,"Environmental modelling and software"
        ,"Journal of hydrology"
        ,"Annual Review of Environment and Resources"

    ]

    pth=
    r'C:\Users\Public\Downloads\pdf_downloader\Selected_Journals_Try'
    for (root, dirs, files) in os.walk(pth,topdown=True):
        for dir in dirs:

            journal_name=os.path.join(root,dir).split("\\")[-1]
            if journal_name in non_member_publications:
                print(journal_name)
                for (root1, dirs1, files1)
```

```
in os.walk(os.path.join(root,dir)
, topdown=True):
    for file in files1[:1]:

        print(os.path.join(root1,file))
        df=pd.read_csv(
        os.path.join(root1,file),
        usecols=['Authors','Author(s) ID','Title','Year',
        'Source title','Cited by','DOI','Link','Abstract']
        )
        print(df.shape)
        df.dropna(subset=['DOI'],inplace=True)
        df.reset_index(drop=True,inplace=True)
        print(df.shape)



        df['Full link']=df['DOI'].apply(find_link)
        df['Full link'] = list(zip(
        df['Source title'],df['DOI'],
        df['Full link'],[str(x) for x in list(df.index)]))
        df['PDF Location']=df['Full link'].apply(
        test_logic_sample.test_logic_sample)
```

```
                        df['Full link'] = list(zip(
                        df['DOI'],
                        df['PDF Location'],
                        df['Source title']))
                        df['Text Location'] = df['Full link'].apply(
                            processing_doc.processing_doc)



                        df.to_csv(os.path.join(root1,file))




t2 = time.perf_counter()
print(f'Finished in {t2-t1} seconds')
```

The code block written in Section 9.1 utilized two helper functions. The first function *i.e. test_logic_sample* is used for downloading the PDF of the full-length research papers. The code is written below

```
import shutil
import pickle
from selenium import webdriver
```

```
import threading
import time
from selenium.webdriver.chrome.options import Options
from selenium import webdriver
import os
from _collections import defaultdict
import concurrent.futures
import time
file_location_test=defaultdict(str)


def test_logic_sample(link):
    print(link[3])
    try:

        m =
        r"C:\Users\Public\Downloads\pdf_downloader\Download\{}\Sample{}".
        format(link[0],link[3])
        print(m)

        if os.path.exists(m):
            shutil.rmtree(m)
```

```
    print("Directory has been deleted")

if not os.path.exists(m):
    os.makedirs(m)

opts = Options()
opts.add_argument('--profile-directory=Profile 2')

opts.add_argument('--headless')
opts.add_argument('--no-sandbox')
prefs = {"profile.default_content_settings.popups": 0,
"download.default_directory": m,
"directory_upgrade": True}
opts.add_experimental_option("prefs", prefs)
driver = webdriver.Chrome(
executable_path=
r'C:\Users\aksha\Downloads\chromedriver_win32 (2)\chromedriver.exe',
options=opts)
driver.get(link[2])
time.sleep(1)
file_location_test[link[1]] = m + "\\" + sorted(
os.listdir(m),
```

```
    key=lambda x: os.path.getmtime(m + "\\" + x))[-1]
    print(link[2])



    fileends = "crdownload"
    while "crdownload" == fileends:
        time.sleep(2)
        files = sorted(os.listdir(m),
        key=lambda x: os.path.getmtime(m + "\\" + x))[-1]
        filename = m + "\\" + files
        print(files)
        file_location_test[link[1]] = m + "\\" + files

        if "crdownload" in files:
            print("Still Downloading")
            fileends = "crdownload"
        else:

            fileends = "none"
    driver.quit()
    return (file_location_test[link[1]])
except Exception as e:
```

```
        print(e)
        file_location_test[link[1]] = "full-text not found"
        print("full-text not found")

        return (file_location_test[link[1]])
```

The other helping function utilized is *processing_doc*. This function converts the PDF into a text file which can then be easily used for topic modelling. The code is written below.

```
import io
import os

from pdfminer.pdfinterp import PDFResourceManager, PDFPageInterpreter
from pdfminer.converter import TextConverter
from pdfminer.layout import LAParams
from pdfminer.pdfpage import PDFPage
from pdfminer.high_level import extract_text
from _collections import defaultdict
import shutil
import time
import math

file_location_text=defaultdict(str)
```

```python
import signal
import functools


def processing_doc(args):
    print(args)
    try:
        rsrcmgr = PDFResourceManager()
        codec = 'utf-8'
        laparams = LAParams()

        outstream = io.StringIO()
        laparams = LAParams()
        rsrcmgr = PDFResourceManager(caching=True)
        device = TextConverter(rsrcmgr, outstream, laparams=laparams,
                               imagewriter=None)
        interpreter = PDFPageInterpreter(rsrcmgr, device)
        password = ""
        maxpages = 0
        caching = True
        pagenos = set()
        path = args[1]
```

```
if args[1] == "full-text not found":
    file_location_text[args[0]] = "full-text not found"
    return 0
folder_name = args[1].split("\\")[-2]
file_name = args[1].split("\\")[-1].split(".pdf")[0] + ".txt"

m =
r"C:\Users\Public\Downloads\pdf_downloader\SampleTextFiles\{}"
.format(args[2])
print(os.path.join(m, folder_name))
if os.path.exists(os.path.join(m, folder_name)):
    shutil.rmtree(os.path.join(m, folder_name))
    print("Directory has been deleted")

if not os.path.exists(os.path.join(m, folder_name)):
    os.makedirs(os.path.join(m, folder_name))

write_file = m + "\\" + folder_name + "\\" + file_name
print(write_file)

try:
```

```
        with open(path, 'rb') as fp:
            text = extract_text(fp)
            print("{}:{}".format(folder_name,text[:5] ))
            with open(write_file, 'w+', encoding="utf-8") as wp:
                wp.write(text)
            file_location_text[args[0]] = write_file
        if os.path.exists("\\".join(path.split("\\")[:-1])):
            shutil.rmtree("\\".join(path.split("\\")[:-1]))
            print("Directory has been deleted")
            return write_file
    except Exception as e:
        print(e)
        return e
except Exception as e:
    print(e)
    return e
```