Content Selection and Curation for Web Archiving

The Gatekeepers vs the Masses

Ian Milligan (@ianmilligan1)
Nick Ruest (@ruebot)
Jimmy Lin (@lintool)
A tale of three collections
...or...
Gatekeepers vs. the Masses
Canadian Political Parties &
Political Interest Group Collection
(ARCHIVE-IT/Toronto)

● 50 Websites
  ○ All major political parties
  ○ Many minor political parties
  ○ Political interest groups
● Collected quarterly between 2005 and present

Stay Informed

Top Stories
- September 29, 2005
  Statement by the Prime Minister on the retirement of John Hamm, Premier of Nova Scotia
- September 28, 2005
  Charity Barbecue Raises $125,000 for Hurricane Katrina Victims
- September 27, 2005
  Address by Prime Minister Paul Martin at the Installation of the new Governor General

Your Excellencies, Honourable Members, Ladies and Gentlemen:

Let me begin by expressing, on behalf of all Canadians, our appreciation to the Right Honourable Adrienne Clarkson and John Ralston Saul. With warmth, intelligence, and wit, they have honoured this high office and made an indelible contribution to our nation.

Over the course of six years, Madame Clarkson recognized achievement, decorated bravery, bore witness to tragedy and grief, and encouraged the disadvantaged. She welcomed foreign visitors and eloquently explained before audiences abroad what it is that makes Canada special. She took great interest in our cities and towns, and especially the north. She traveled to more than 200 communities across Canada; in some of them, it was the first-ever visit by a representative of the Crown.

Full Story
Access

- **ArchiveIT** - simple search engine, some advanced options; no facets

- **wget/curl** - ~350GB
The Search API

The Twitter Search API is part of Twitter’s REST API. It allows queries against the indices of recent or popular Tweets and behaves similarly to, but not exactly like the Search feature available in Twitter mobile or web clients, such as Twitter.com search. The Twitter Search API searches against a sampling of recent Tweets published in the past 7 days.

Before getting involved, it’s important to know that the Search API is focused on relevance and not completeness. This means that some Tweets and users may be missing from search results. If you want to match for completeness you should consider using a Streaming API instead.

A detailed reference on this API endpoint can be found at GET search/tweets.

How to build a query

The best way to build a query and test if it’s valid and will return matched Tweets is to first try it at twitter.com/search or using the Twitter advanced search query builder. As you get a satisfactory result set, the URL loaded in the browser will contain the proper query syntax that can be reused in the API endpoint. Here’s an example:

1. We want to search for tweets referencing @twitterapi account. First, we run the search on twitter.com/search

2. Check and copy the URL loaded. In this case, we got: https://twitter.com/search?q=%40twitterapi

3. Replace “https://twitter.com/search” with “https://api.twitter.com/1.1/search/tweets.json” and you will get: https://api.twitter.com/1.1/search/tweets.json
The Streaming APIs

Overview

The Streaming APIs give developers low latency access to Twitter's global stream of Tweet data. A proper implementation of a streaming client will be pushed messages indicating Tweets and other events have occurred, without any of the overhead associated with polling a REST endpoint.

If your intention is to conduct singular searches, read user profile information, or post Tweets, consider using the REST APIs instead.

Twitter offers several streaming endpoints, each customized to certain use cases.

| Public streams | Streams of the public data flowing through Twitter. Suitable for following specific users or topics, and data mining. |
| User streams   | Single-user streams, containing roughly all of the data corresponding with a single user's view of Twitter. |
| Site streams   | The multi-user version of user streams. Site streams are intended for servers which must connect to Twitter on behalf of many users. |

Differences between Streaming and REST

Connecting to the streaming API requires keeping a persistent HTTP connection open. In many cases this involves thinking about your application differently than if you were interacting with the REST API. For example, considering a web application.
twarc

https://github.com/edsu/twarc
twarc

Twarc is a command line tool and Python library for archiving Twitter JSON data. Each tweet is represented as a JSON object that is exactly what was returned from the Twitter API. Tweets are stored as line-oriented JSON. Twarc runs in three modes: search, filter, stream and hydrate. When running in each mode twarc will stop and resume activity in order to work within the Twitter API’s rate limits.

Install

1. install Python (2 or 3)
2. pip install twarc

Twitter API Keys

Before using twarc you will need to register an application at apps.twitter.com. Once you’ve created your application, note down the consumer key, consumer secret and then click to generate an access token and access token secret. With these four variables in hand you are ready to start using twarc.

The first time you run twarc it will prompt you for these keys and store them in a .twarc file in your home directory. Sometimes it can be handy to store multiple authorization keys for different Twitter accounts in your config file. So if you can have multiple profiles to your .twarc file, for example:

```
[main]
consumer_key=lksdfljklksdjf
consumer_secret=lkjsdflkjSDLKFj
```
3,918,932 tweets

eh?
318,176 unique URLs

(1,988,693 URLs total)
unshrtntn

https://github.com/edsu/unshrtntn
Intersection Analysis
## Top Domains

<table>
<thead>
<tr>
<th>Twitter</th>
<th>CPP (Aug - Nov 2015)</th>
</tr>
</thead>
<tbody>
<tr>
<td>twitter.com</td>
<td>liberal.ca</td>
</tr>
<tr>
<td>cbc.ca</td>
<td>55,536</td>
</tr>
<tr>
<td>youtube.com</td>
<td>greenparty.ca</td>
</tr>
<tr>
<td>huffingtonpost.ca</td>
<td>policyalternatives.ca</td>
</tr>
<tr>
<td>theglobeandmail.com</td>
<td>socialist.ca</td>
</tr>
<tr>
<td>thestar.com</td>
<td>davidsuzuki.org</td>
</tr>
<tr>
<td>ctvnews.ca</td>
<td>canadians.org</td>
</tr>
<tr>
<td>globalnews.ca</td>
<td>ccrweb.ca</td>
</tr>
<tr>
<td>twimg.com</td>
<td>afn.ca</td>
</tr>
<tr>
<td>macleans.ca</td>
<td>blocquebecois.org</td>
</tr>
<tr>
<td></td>
<td>egale.ca</td>
</tr>
<tr>
<td></td>
<td>CPP</td>
</tr>
<tr>
<td>-------</td>
<td>---------</td>
</tr>
<tr>
<td>CPP</td>
<td></td>
</tr>
<tr>
<td>Twitter</td>
<td>0.269%</td>
</tr>
<tr>
<td>Wayback</td>
<td>N/A</td>
</tr>
<tr>
<td>Included</td>
<td>Included</td>
</tr>
<tr>
<td>-------------------</td>
<td>----------</td>
</tr>
<tr>
<td>cbc.ca</td>
<td>3,035</td>
</tr>
<tr>
<td>youtube.com</td>
<td>2,639</td>
</tr>
<tr>
<td>thestar.com</td>
<td>1,665</td>
</tr>
<tr>
<td>theglobeandmail.com</td>
<td>1,644</td>
</tr>
<tr>
<td>huffingtonpost.ca</td>
<td>1,561</td>
</tr>
<tr>
<td>twitter.com</td>
<td>1,550</td>
</tr>
<tr>
<td>ctvnews.ca</td>
<td>1,423</td>
</tr>
<tr>
<td>nationalpost.com</td>
<td>1,262</td>
</tr>
<tr>
<td>globalnews.ca</td>
<td>1,062</td>
</tr>
<tr>
<td>ottawacitizen.com</td>
<td>836</td>
</tr>
</tbody>
</table>
Conclusions
#elxn42 web crawl

**Description**
Consists of a web crawl of unique URLs tweeted with the #elxn42 hashtag. #elxn42 collection took place from August 1, 2015 – November 5, 2015. Unique URLs were extracted from the dataset, and harvested with Heritrix on January 23, 2016 – February 6, 2016.

**Download**
- web: elxn42 web crawl.gz
- cdr: elxn42 web crawl.cdx
- web: elxn42 web crawl interruptions.gz
- Seed list: elxn42 web crawl.cdx
- Heritrix configuration: crawler-beans.xml

**In collections**
- #elxn42

---

#panamapapers crawl; May 1-7, 2016

**Description**
Consists of a web crawl of unique URLs tweeted with the #panamapapers hashtag. Panamapapers collection took place from April 29, 2016. Unique URLs were extracted from the dataset, and harvested with Heritrix on May 1-7, 2016.

**Download**
- web: panamapapers crawl-May 1-7, 2016.gz
- cdr: panamapapers crawl-May 1-7, 2016.cdx
- web: panamapapers crawl-May 1-7, 2016 interruptions.gz
- Seed list: panamapapers crawl-May 1-7, 2016.cdx
- Heritrix configuration: crawler-beans.xml

**In collections**
- #panamapapers

---

**Details**
- **Title:** #panamapapers crawl; May 1-7, 2016
- **Creator(s):** Nick Rusen
- **Note:** Tweet id: http://t.co/109h411590
- **Identifier (local):** WEB:201605020020310201300010-1049-Y.o; library.jorka.ca-ca;7717-PAAW4994F92
- **Identifier (metadata):** m4n3VX09G95jV59628z17sV6796256t5a
- **Type:** Website
- **Subject(s):** #panamapapers, Panama Papers, Mossack Fonseca, International Consortium of Investigative Journalists, tax havens, web archiving, twitter
- **Date captured:** 2016-05-01
- **Size:** 9.96M
- **File size:** 1.97GB
- **Format:** x-amz-cdxh

**Funding:** This research was supported by a research grant – R-255-2015-0111 – issued by Social Sciences and Humanities Research Council.

**Rights:** Use of this resource is governed by the terms and conditions of the Creative Commons 'Attribution' License ([http://creativecommons.org/licenses/by/2.0/](http://creativecommons.org/licenses/by/2.0/))
Questions

ruestn@yorku.ca

@ruebot