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

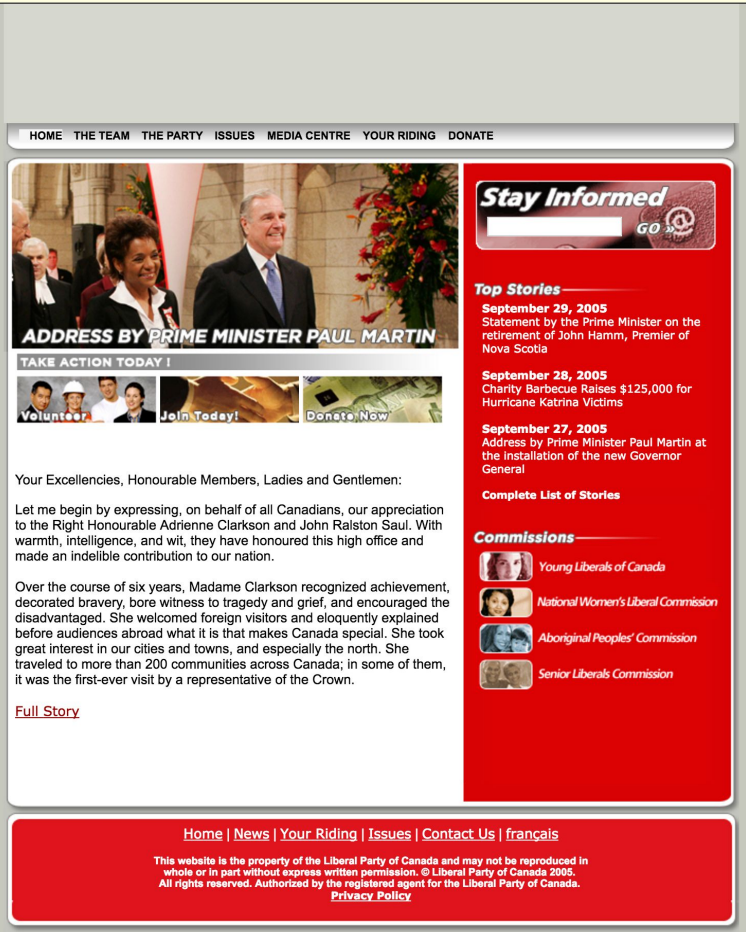
...or...

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

You are viewing an archived web page, collected at the request of University of Toronto using Archive-It. hide
This page was captured on 19:14:04 Oct 04, 2005, and is part of the Canadian Political Parties and Political Interest Groups collection. The information on this web page may be out of date. See All versions of this archived page.



Access

- **ArchiveIT** - simple search engine, some advanced options; no facets
- **wget/curl** - ~350GB

warcbase

warcbase.org | docs.warcbase.org

#elxn42

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**

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

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

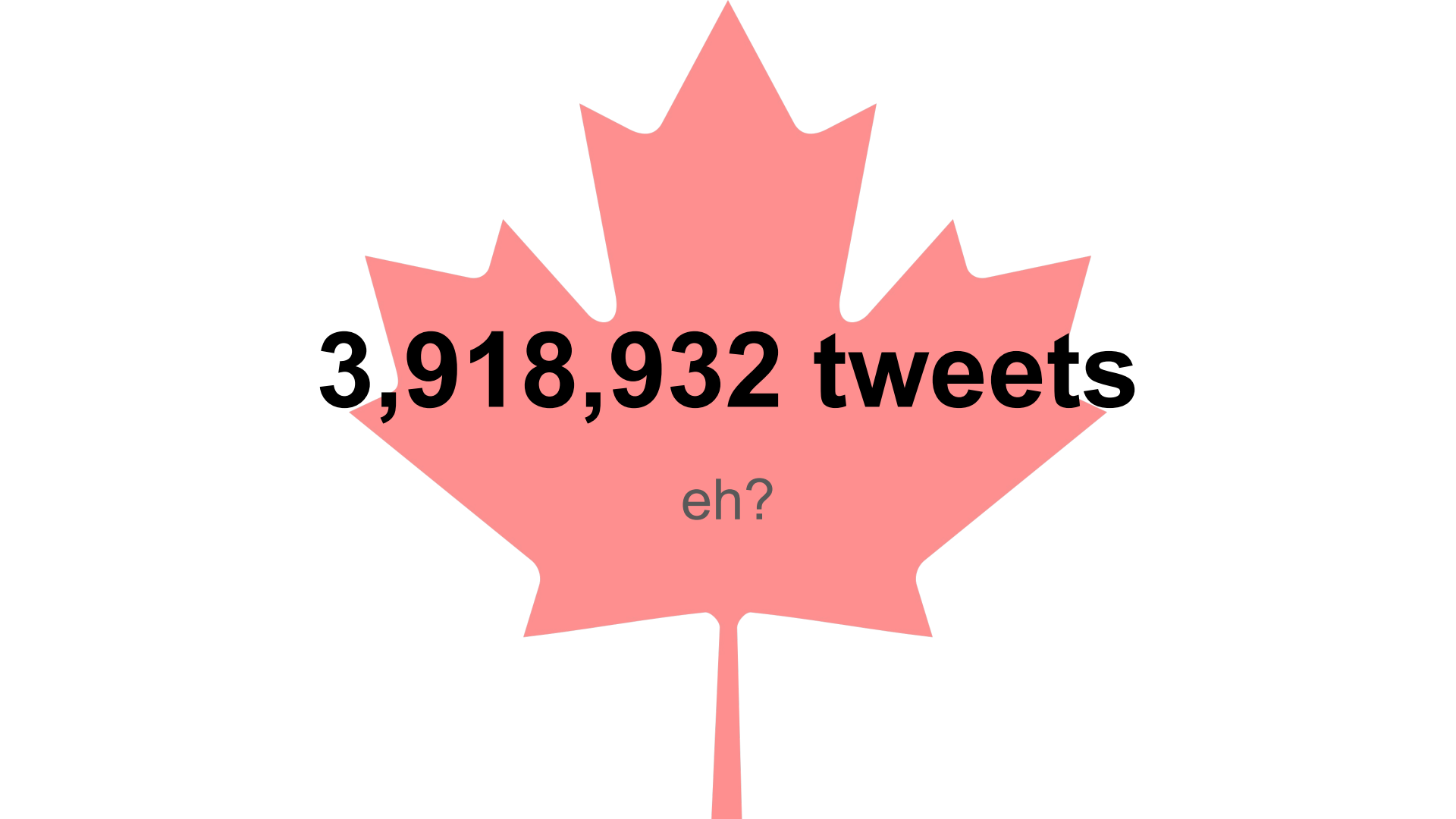
1. install Python (2 or 3)
2. pip install twarc

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)

unshrtn

https://github.com/edsu/unshrtn


INTERNET ARCHIVE

Intersection Analysis

Top Domains

Twitter		**CPP (Aug - Nov 2015)**	
twitter.com	**615,421**	liberal.ca	
cbc.ca	**143,941**	**55,536**	
youtube.com		greenparty.ca	**45,788**
66,886		policyalternatives.ca	**37,810**
huffingtonpost.ca	**66,758**	socialist.ca	**26,856**
theglobeandmail.com	**63,401**	davidsuzuki.org	**25,487**
thestar.com	**53,051**	canadians.org	**24,424**
ctvnews.ca	**49,295**	ccrweb.ca	**19,521**
globalnews.ca	**46,488**	afn.ca	**15,879**
twimg.com	**39,989**	blocquebecois.org	**10,899**
macleans.ca	**35,280**	egale.ca	**7,837**

	CPP	Twitter	Wayback
CPP		0.341%	74.30%
Twitter	0.269%		10.06%
Wayback	N/A	N/A	

Included Domains (Internet Archive)

Included		**Excluded**	
cbc.ca	**3,035**	twitter.com	**173,931**
youtube.com	**2,639**	linkis.com	**11,071**
thestar.com	**1,665**	youtube.com	**6,026**
theglobeandmail.com	**1,644**	instagram.com	**5,302**
huffingtonpost.ca	**1,561**	globalnews.ca	**4,709**
twitter.com	**1,550**	cbc.ca	**4,529**
ctvnews.ca	**1,423**	facebook.com	**4,282**
nationalpost.com	**1,262**	rabble.ca	**3,859**
globalnews.ca	**1,062**	huffingtonpost.ca	**3,762**
ottawacitizen.com	**836**	fw.to	**3,284**

Conclusions

Advanced search
Username *
Password *
Log in

#elxn42 web crawl

Description

Consists of a web crawl of unique URLs tweeted with the #elxn42 hashtag. #elxn42 collection took place from August 3, 2015 - November 5, 2015. Unique URLs were extracted from the dataset, and harvested with Heritrix on January 29, 2016 - February 8, 2016.

Download

- warc: #elxn42 web crawl.gz
- cdx: #elxn42 web crawl.cdx
- wat: #elxn42 web crawl.wat.gz
- Seed list: #elxn42 web crawl.txt
- Heritrix configuration: crawler-beans.cxml

In collections

- #elxn42

Details

Title:	#elxn42 web crawl
Creator(s):	Nick Ruest
Note:	Tweet ids: http://hdl.handle.net/10864/11311
Identifier (local):	WEB-20160208134917869-00013-3991~rho.library.yorku.ca~9191-ELXN42
Identifier (md5):	63d707352a6fb45a62889c448154610f
Type:	Website
Subject(s):	#elxn42 Canadian federal election, 2015 Canadian politics Federal politics Canada
Date captured:	2016-01-29
Size:	13GB
File size:	12972947583
PUID:	x-fmt/266
Funding	This research was supported by a research grant -- 435-2015-0011 -- 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/)

Advanced search
Username *
Password *
Log in

#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 4-29, 2016. Unique URLs were extracted from the dataset, and harvested with Heritrix on May 1-7, 2016.

Download

- warc: #panamapapers crawl; May 1-7, 2016.gz
- cdx: #panamapapers crawl; May 1-7, 2016.cdx
- wat: #panamapapers crawl; May 1-7, 2016.wat.gz
- Seed list: #panamapapers crawl; May 1-7, 2016.txt
- Heritrix configuration: crawler-beans.cxml

In collections

- #panamapapers

Details

Title:	#panamapapers crawl; May 1-7, 2016
Creator(s):	Nick Ruest
Note:	Tweet ids: http://hdl.handle.net/10864/11592
Identifier (local):	WEB-20160505182658813-00010-9949~rho.library.yorku.ca~9191-PANAMAPAPERS
Identifier (md5):	e43e8940bf56e340a85ab38975676dfa
Type:	Website
Subject(s):	#panamapapers politics Panama Papers Mossack Fonseca International Consortium of Investigative Journalists tax havens web archiving twitter
Date captured:	2016-05-01
Size:	9.9GB
File size:	10547578814
PUID:	x-fmt/266
Funding	This research was supported by a research grant -- 435-2015-0011 -- 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/)

Questions

ruestn@yorku.ca
@ruebot