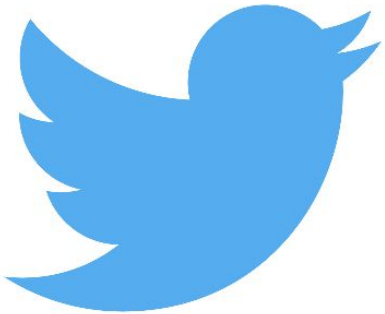
# An exploratory look at 3,039,804 #elxn42 tweets


Nick Ruest (@ruebot)
Ian Milligan (@ianmilligan1)

# Why Twitter?



- Potential to reshape multiple avenues of historical research – **318,176 unique users on the #elxn42 hashtag**
- Multiple use cases
  - Political historians
  - Military historians
  - Social and cultural historians
- Sheer scope of this data – i.e. #IdleNoMore on 11 January 2013 = 1,800 pages!


Mascot
YOU DESTROYED OUR CHILDHOODS
We Were Not Consulted
Idle NO More
Idle NO More
Harper go Suck Buffalo Balls
PROTECT OUR FUTURE GENERATIONS
NATIVE RIGHTS
ROCK UNIVERSE
CELLULAR CENTRE
First Choice
TheIndignants.org

Not a perfect source, of course...

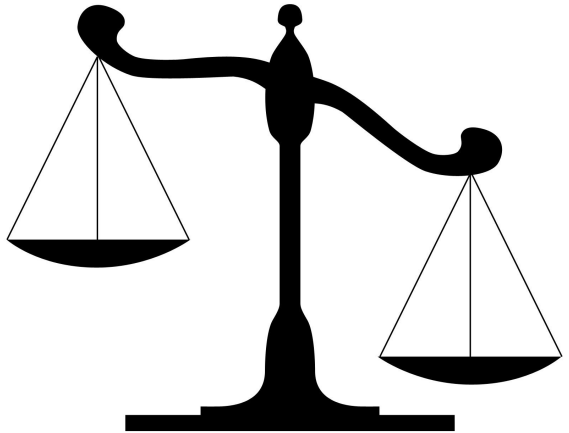
# Time is of the essence!

- 7 - 9 days: largely inaccessible after this point w/o bags of money
- But before then, you can:
  - Create your own archives using twarc;
  - Analyze tweets using twarc-report & twarc-utilities;
  - Visualize material;
  - Use this as a seed for web archiving;
  - Share datasets

# Hydration

# We can collect.. But should we? **Ethical considerations**.

- Twitter Terms of Service & Copyright
- But legal =! Ethics
  - Example of the “Black Twitter Project” at USC
  - But if we do not collect, archives will be of powerful people...



Onus on researchers…
(is that fair?)

# The Case Study

- **2015 Canadian federal election**
- Collection began on 3 August 2015, day before the formal writ dropped;
- Collection ended on 5 November 2015, the day after Justin Trudeau became 42nd Prime Minister of Canada
- #elxn42 hashtag

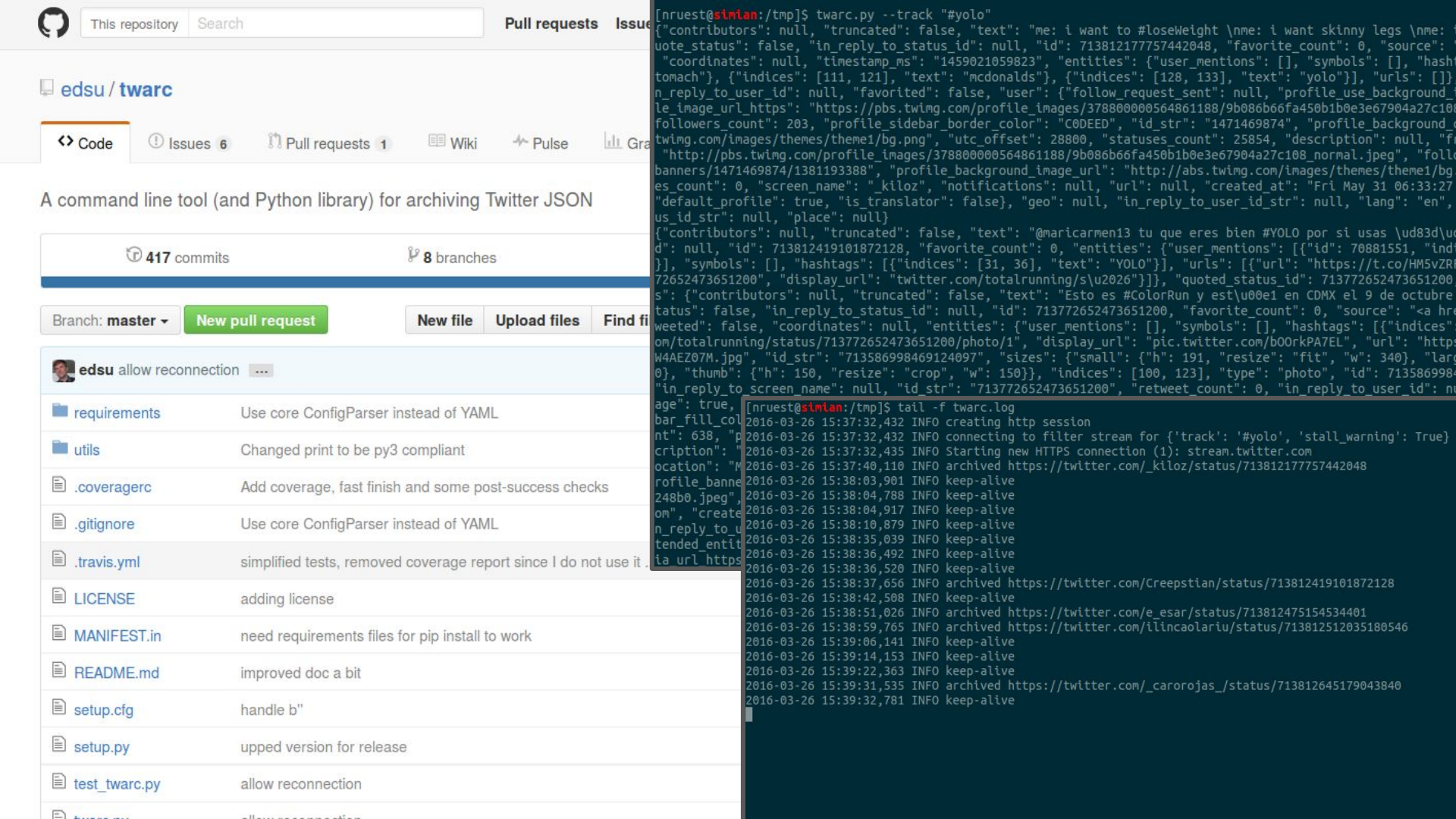

How to do this?
Over to Nick...

# Twitter APIs

# Search API

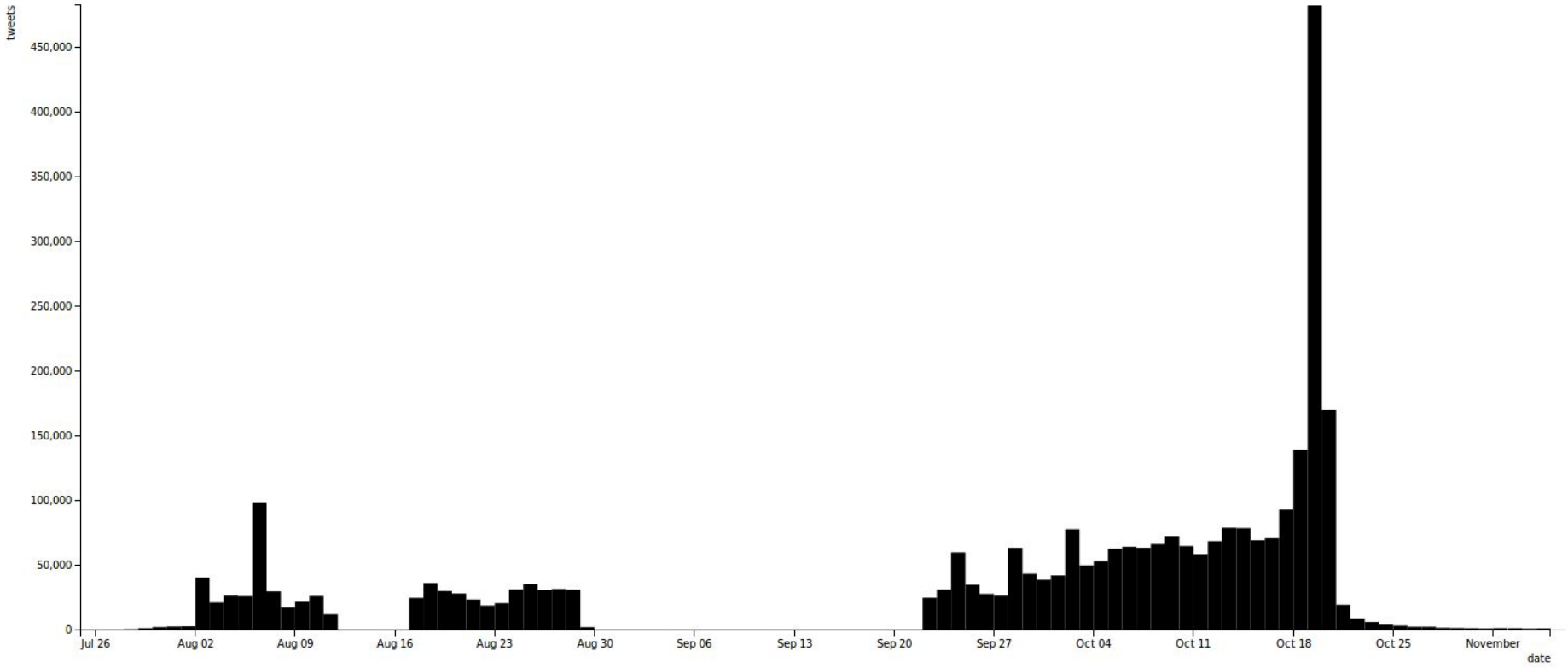
# Streaming API

# twarc


This repository
Search
Pull requests
Issue
edsu / twarc
Code
Issues 6
Pull requests 1
Wiki
Pulse
A command line tool (and Python library) for archiving Twitter JSON
417 commits
8 branches
Branch: master
New pull request
New file
Upload files
edsu allow reconnection
requirements
Use core ConfigParser instead of YAML
utils
Changed print to be py3 compliant
.coveragerc
Add coverage, fast finish and some post-success checks
.gitignore
Use core ConfigParser instead of YAML
.travis.yml
simplified tests, removed coverage report since I do not use it
LICENSE
adding license
MANIFEST.in
need requirements files for pip install to work
README.md
improved doc a bit
setup.cfg
handle b''
setup.py
upped version for release
test_twarc.py
allow reconnection
[nruest@simian:/tmp]$ twarc.py --track "#yolo"
[nruest@simian:/tmp]$ tail -f twarc.log
2016-03-26 15:37:32,432 INFO creating http session
2016-03-26 15:37:32,432 INFO connecting to filter stream for {'track': '#yolo', 'stall_warning': True}
2016-03-26 15:37:32,435 INFO Starting new HTTPS connection (1): stream.twitter.com
2016-03-26 15:37:40,110 INFO archived https://twitter.com/_kiloz/status/713812177757442048
2016-03-26 15:38:03,901 INFO keep-alive
2016-03-26 15:38:04,788 INFO keep-alive
2016-03-26 15:38:04,917 INFO keep-alive
2016-03-26 15:38:10,879 INFO keep-alive
2016-03-26 15:38:35,039 INFO keep-alive
2016-03-26 15:38:36,492 INFO keep-alive
2016-03-26 15:38:36,520 INFO keep-alive
2016-03-26 15:38:37,656 INFO archived https://twitter.com/Creepstian/status/713812419101872128
2016-03-26 15:38:42,508 INFO keep-alive
2016-03-26 15:38:51,026 INFO archived https://twitter.com/e_esar/status/713812475154534401
2016-03-26 15:38:59,765 INFO archived https://twitter.com/ilincaolariu/status/713812512035180546
2016-03-26 15:39:06,141 INFO keep-alive
2016-03-26 15:39:14,153 INFO keep-alive
2016-03-26 15:39:22,363 INFO keep-alive
2016-03-26 15:39:31,535 INFO archived https://twitter.com/_carorojas_/status/713812645179043840
2016-03-26 15:39:32,781 INFO keep-alive

# Data Collection

2015-07-25 17:56:45 EDT to 2015-11-05 06:46:45 EST

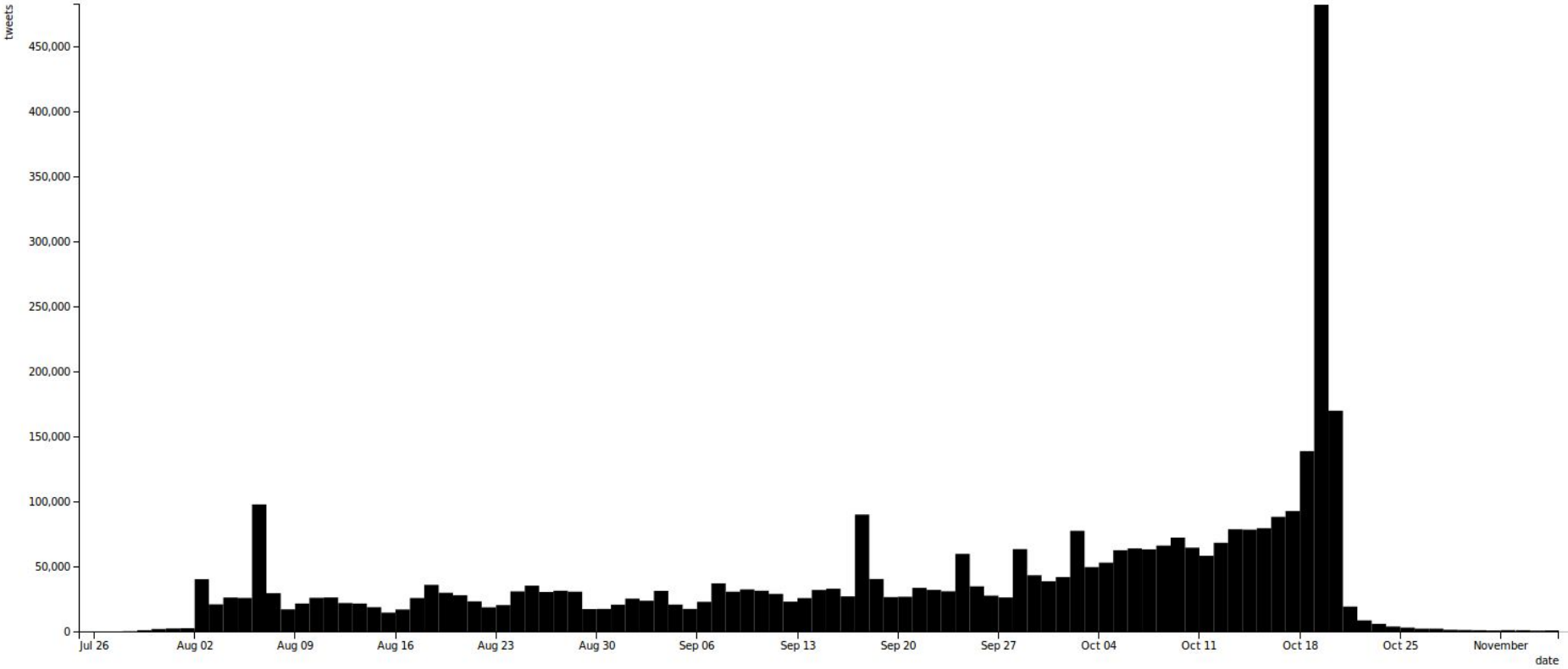


Well actually…

**3,918,932 tweets**

Thanks Library and Archives Canada!

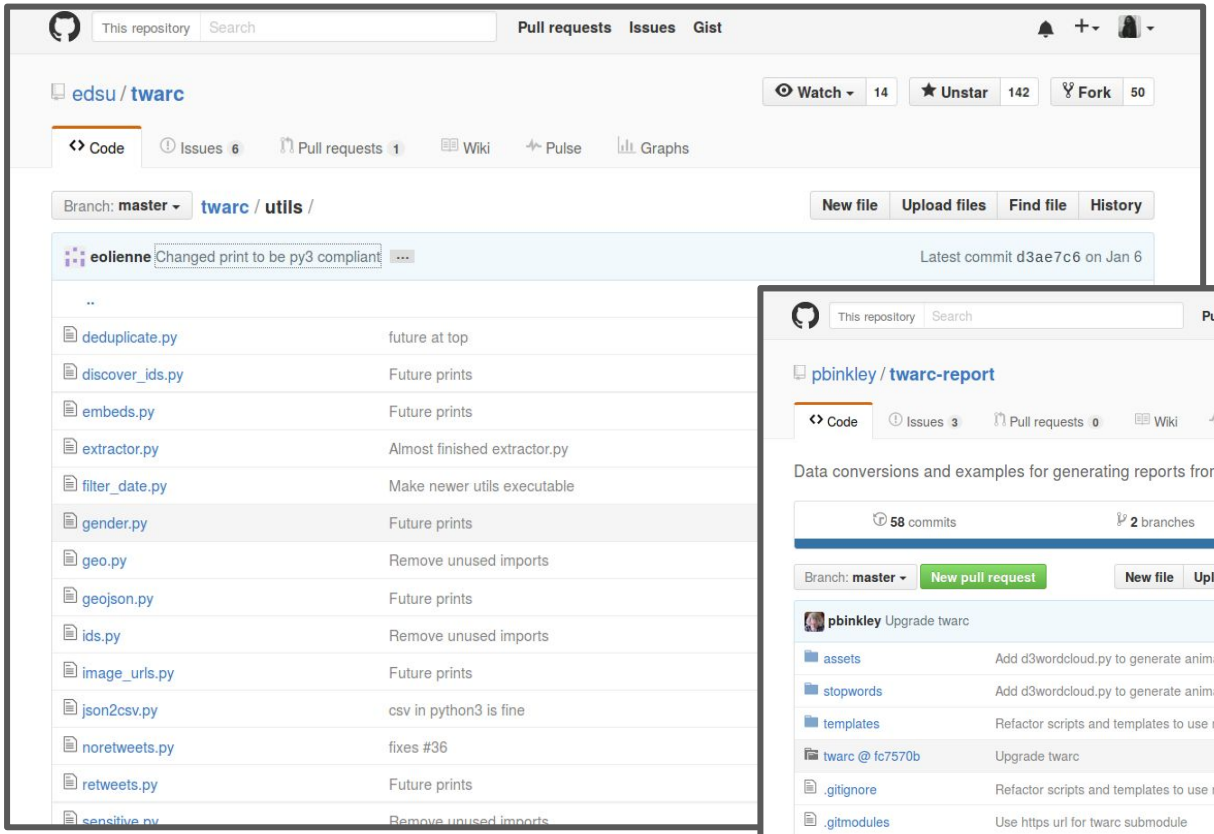
2015-07-25 17:56:45 EDT to 2015-11-05 06:46:45 EST

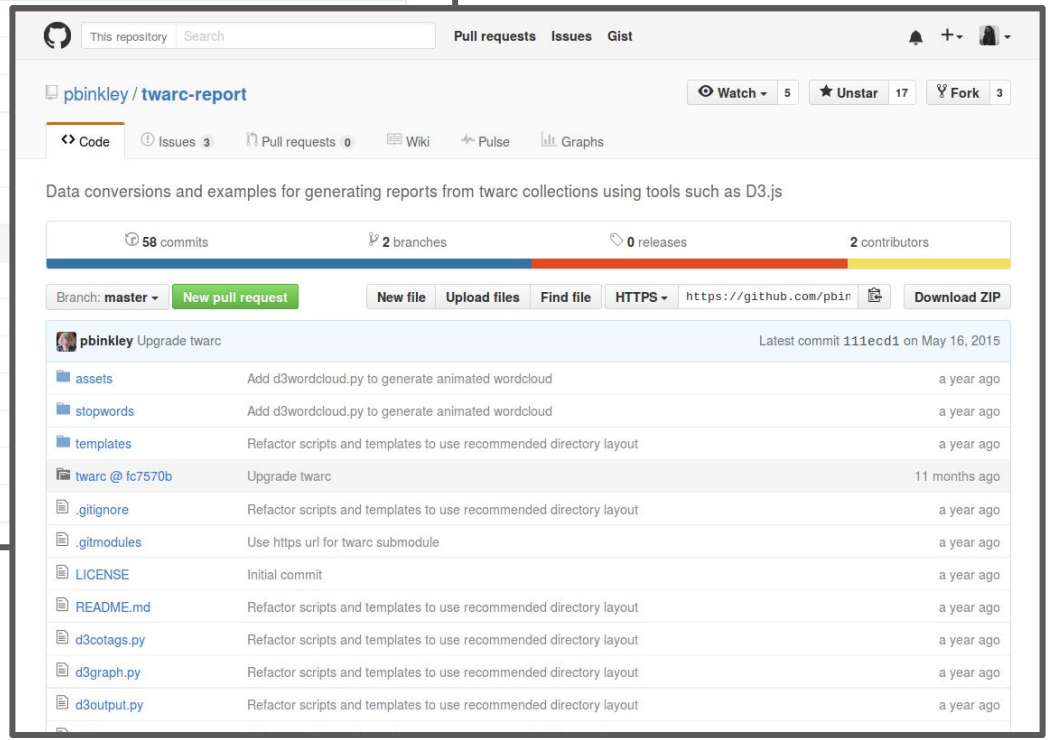


```
$ twarc.py --search "#elxn42"
> elxn42-search.json

$ twarc.py --track "#elxn42" >
elxn42-stream.json
```

# Analysis


This repository Search
Pull requests Issues Gist
edsu / twarc
Watch 14 Unstar 142 Fork 50
Code Issues 6 Pull requests 1 Wiki Pulse Graphs
Branch: master twarc / utils /
New file Upload files Find file History
eolienne Changed print to be py3 compliant ...
Latest commit d3ae7c6 on Jan 6
..
deduplicate.py future at top
discover_ids.py Future prints
embeds.py Future prints
extractor.py Almost finished extractor.py
filter_date.py Make newer utils executable
gender.py Future prints
geo.py Remove unused imports
geojson.py Future prints
ids.py Remove unused imports
image_urls.py Future prints
json2csv.py csv in python3 is fine
noretweets.py fixes #36
retweets.py Future prints


This repository Search
Pull requests Issues Gist
pbinkley / twarc-report
Watch 5 Unstar 17 Fork 3
Code Issues 3 Pull requests 0 Wiki Pulse Graphs
Data conversions and examples for generating reports from twarc collections using tools such as D3.js
58 commits 2 branches 0 releases 2 contributors
Branch: master New pull request
New file Upload files Find file HTTPS https://github.com/pbin Download ZIP
pbinkley Upgrade twarc
Latest commit 111ecd1 on May 16, 2015
assets Add d3wordcloud.py to generate animated wordcloud a year ago
stopwords Add d3wordcloud.py to generate animated wordcloud a year ago
templates Refactor scripts and templates to use recommended directory layout a year ago
twarc @ fc7570b Upgrade twarc 11 months ago
.gitignore Refactor scripts and templates to use recommended directory layout a year ago
.gitmodules Use https url for twarc submodule a year ago
LICENSE Initial commit a year ago
README.md Refactor scripts and templates to use recommended directory layout a year ago
d3cotags.py Refactor scripts and templates to use recommended directory layout a year ago
d3graph.py Refactor scripts and templates to use recommended directory layout a year ago
d3output.py Refactor scripts and templates to use recommended directory layout a year ago


jq is a lightweight and flexible command-line JSON processor.

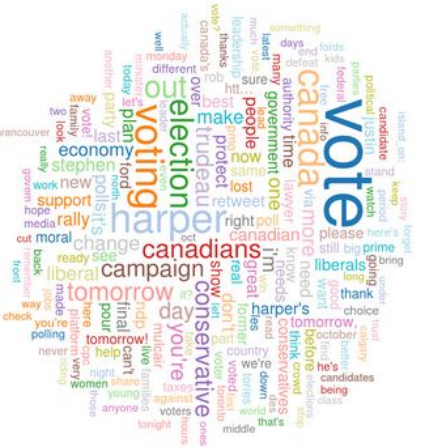
Download jq 1.5 ▾ | Try online at jqplay.org!

## CLI Utilities

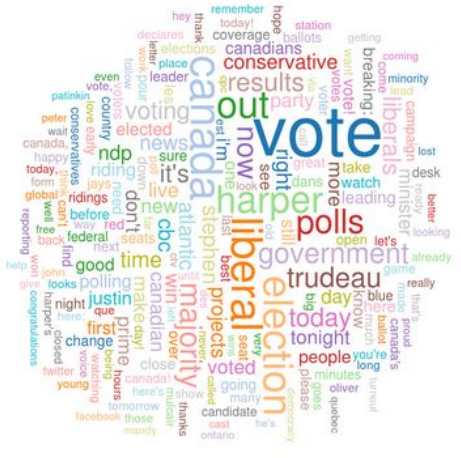
`cat, awk, sed, grep, uniq, sort, wc, etc...`

October 17, 2015


October 18, 2015




**October 19, 2015**




October 20, 2015

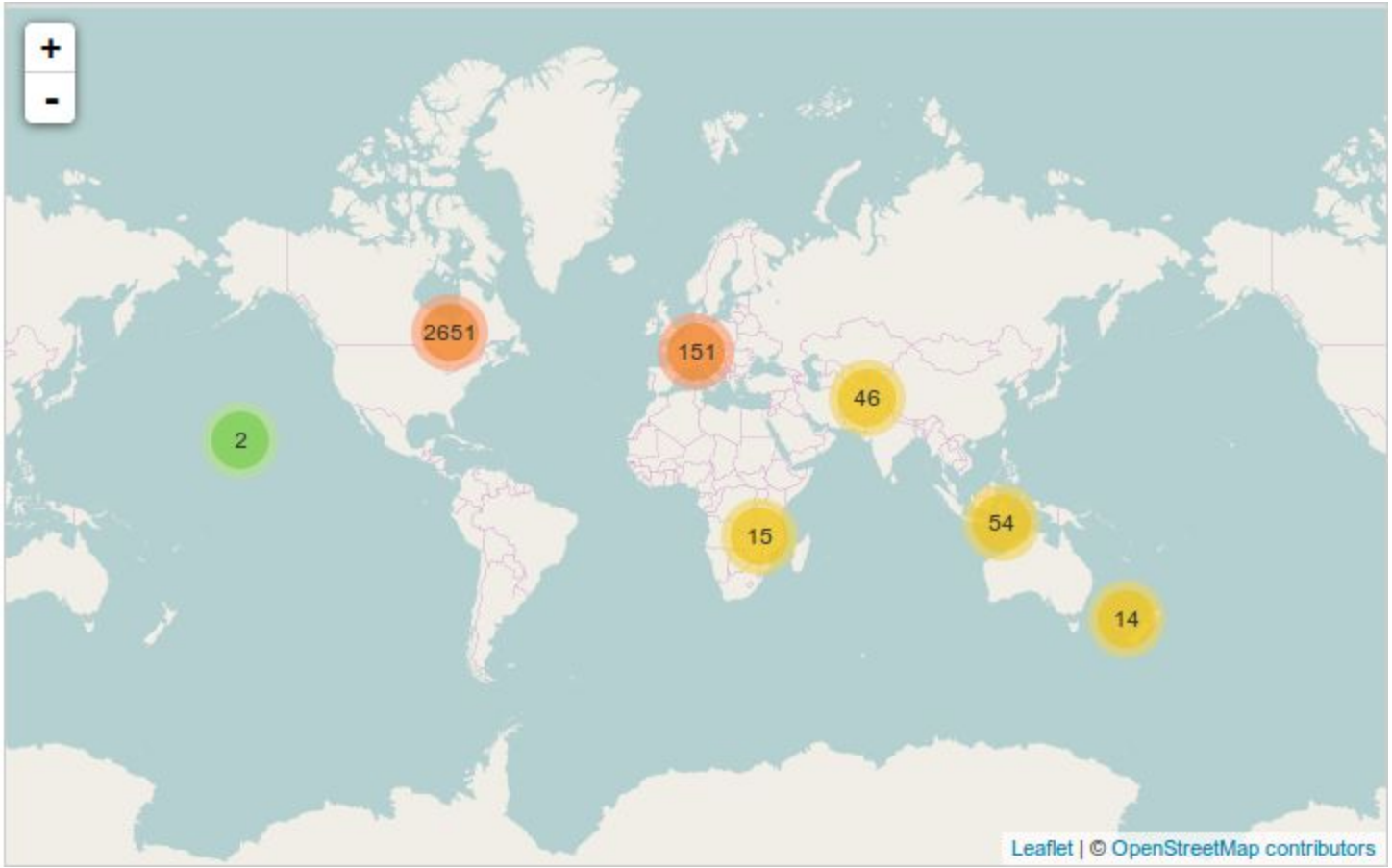
**Justin Trudeau**
@JustinTrudeau

Follow

# Ready. #elxn42


RETWEETS 5,319 LIKES 9,469

1:33 AM - 20 Oct 2015


+
-
2651
151
46
2
15
54
14
Leaflet | © OpenStreetMap contributors

Home Notifications Messages Search Twitter Tweet



TWEETS 176K | FOLLOWING 2,224 | FOLLOWERS 1,542 | LIKES 15.3K

Follow

**David Morrison**
@DavidMorrison17

My FB timeline is open to the public if you'd like to view some interesting documents and read my opinions concerning an illegal CSIS operation against me.

Whitby, ON
facebook.com/david.morrison…
Joined August 2011

Tweets | Tweets & replies | Photos & videos

Pinned Tweet

**David Morrison** @DavidMorrison17 · 15 Oct 2015

Synopsis of #CSIS' illegal operation against me. Please share: tl.gd/n_1s2qglv #cdnpoli #elxn42


Who to follow · Refresh · View all


**Dan Broadway Project** @D…
Follow


**La Batteria** @LaBatteria
Follow


**Herbcraft** @theherbcraft
Follow

Find friends

# 21,423 tweets

# hashtags

1. 3,685,885
2. 1,390,783
3. 164,339
4. 139,070
5. 129,082
6. 89,303
7. 68,387
8. 64,718
9. 62,282
10. 61,700

1. #elxn42
2. #cdnpoli
3. #ndp
4. #cpc
5. #lpc
6. #elxn2015
7. #polcan
8. #realchange
9. #polqc
10. #globedebate

www.cbc.ca/includes/federalelection/dashboard/index.html

Search

# Canada Votes 2015

Election Day October 19, 2015 FR

## Liberal majority government

ALERTS 02:19 a.m. CBC News projects Kent Hehr (LIB) elected in Calgary Centre

Show all alerts

## National Results

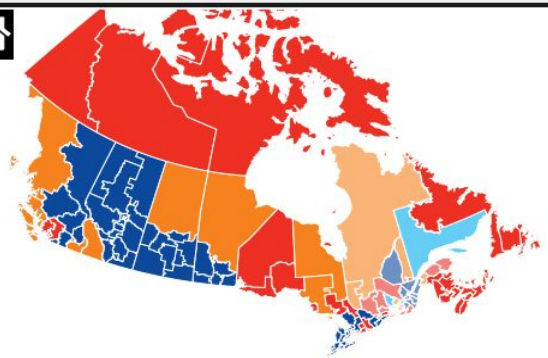
RIDING TOTALS | VOTE SHARE

| | MAJORITY ELECTED | LEADING | TOTAL |
|---|---|---|---|
| L | 184 | 0 | **184** |
| C | 99 | 0 | **99** |
| NDP | 44 | 0 | **44** |
| BQ | 10 | 0 | **10** |
| G | 1 | 0 | **1** |

ELECTED LEADING TO COME: 0 RIDINGS

Show all parties

## Riding Results



Search by postal code or riding name

MY FAVOURITES | RIDINGS TO WATCH | CANDIDATES TO WATCH

**Top Ridings to Watch**

Show why these ridings matter

| | |
|---|---|
| Ajax | LIB |
| Beauport-Côte-de-Beaupré-Île d'Orléans-Charlevoix | CON |
| Berthier-Maskinongé | NDP |
| Brampton Centre | LIB |
| Brome-Missisquoi | LIB |
| Burnaby North-Seymour | LIB |
| Calgary Centre | LIB |

Waiting for discussion.scribblelive.com...

**11,956 tweets**

# domains

1. 615,421
2. 143,941
3. 66,886
4. 66,758
5. 63,401
6. 53,051
7. 49,295
8. 46,488
9. 39,989
10. 35,280

1. twitter.com
2. cbc.ca
3. youtube.com
4. huffingtonpost.ca
5. theglobeandmail.com
6. thestar.com
7. ctvnews.ca
8. globalnews.ca
9. twimg.com
10. macleans.ca


CON FLIP:




Libéral

# 1,203,867 #elxn42 images

Dataset is available here. | Original 32G png available here. | Tiled with deepzoom.py.



Nick Ruest, Web Archives for Historical Research

*Submitted by Nick Ruest on Thu, 03/31/2016 - 17:23*

# 1,203,867 #ELXN42 IMAGES

## BACKGROUND

Last August, I began capturing the #elxn42 hashtag as an experiment, and potential research project with Ian Milligan. Once Justin Trudeau was sworn in as the 23rd Prime Minister of Canada, we stopped collection, and began analysing the dataset. We wrote that analysis up for the Code4Lib Journal, which will be published in the next couple weeks. In the interim, you can check out our pre-print here. Included in that dataset is a line-deliminted list of a url to every embedded image tweeted in the dataset; 1,203,867 images. So, I downloaded them. It took a couple days.

getTweetImages

```
IMAGES=/path/to/elxn42-image-urls.txt
cd /path/to/elxn42/images

cat $IMAGES | while read line; do
  wget "$line"
done
```

Now we can start doing image analysis.

## 1,203,867 IMAGES, NOW WHAT?

I really wanted to take a macroscopic look at all the images, and looking around the best tool for the job looked like `montage`, an ImageMagick command for creating composite images. But, it wasn't that so simple. 1,203,867 images is a lot of images, and starts getting you thinking about what big data is. Is this big data? I don't know. Maybe?

### ATTEMPT #1

I can just point `montage` at a directory and say go to town, right? NOPE.

```
$ montage /path/to/1203867/elxn42/images/* elxn42.png
```

Too many arguments! After glancing through the man page, I find that I can pass it a line-delimited text file with the paths to each file.

```
file paths find `pwd` -type f -exec cat {} > images.txt
```

Now that I have that, I can pass `montage` that file, and I should be golden, right? NOPE.

Over to Ian for some concluding thoughts..

# Deleted Tweets

Twitter Development Agreement & Policy

Spam tweets lost… an invaluable part of the Twitter experience?

Ultimately: **archives are always full of large gaps and omissions; at least here we know that people could make decisions to be removed.**

# Twitter & Web Archiving

Similar to IIPC Twittervane: could we use the tweets of users as a seed list? **Could it help us find absences?** How does it compare to **Global Wayback versus the Archive-It CPP Collection**?

| | CPP | Twitter | Wayback |
|---|---|---|---|
| CPP | - | 0.341% | 74.3% |
| Twitter | 0.269% | - | 10.06% |
| Wayback | N/A | N/A | - |

# Final Thoughts


To do twitter event archiving, you don’t need to be Scrooge McDuck..

You can DIY, making just-in-time, responsive collections – with an open-source analytics solution!

# Thanks and Acknowledgements


Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada

compute canada | calcul canada

Ontario

MINISTRY OF RESEARCH AND INNOVATION
MINISTÈRE DE LA RECHERCHE ET DE L'INNOVATION

journal.code4lib.org

“*An Open-Source Strategy for Documenting Events: The Case Study of the 42nd Canadian Federal Election on Twitter*”

# Contact

Nick Ruest: @ruebot

ruestn@yorku.ca

Ian Milligan: @ianmilligan1

i2milligan@waterloo.ca