Acquiring Social Media Data

Laura Wrubel, Dan Kerchner

September 20, 2018

slides:

The hands-on part of the workshop requires a Twitter account.
Go to twitter.com to create one if you don't have one.
You can delete it later.

Agenda

- Overview of social media APIs and data formats
- Twitter's API in depth
- Using existing datasets
  - Hands-on: TweetSets
- Collecting new datasets
  - Hands-on: Social Feed Manager
- Ethics of social media collecting

APIs, social media APIs, and their data

What's an API?

- Short for “Application Programming Interface”
- Allows code to request or send data to a website
- API calls consist of:
  - **requests**: *http://an.api.com/request?foo=15*
  - **response:** structured data, e.g., XML or JSON

Why use an API for working with social media?

- You don't want to scrape it from the web page!
  - It's hard, will break, and is incomplete.
- An API gives you:
  - Data similar to what the platform stores.
  - Slices of data you can't get by scraping.
  - Sata in structured format, which makes it easy to analyze as data, with analysis tools.

JSON: JavaScript Object Notation

- `{ key: value, key: value… }`
- keys are strings
- a value may be:
  - string - in quotes: `"GW"`
  - number
  - boolean - `true` or `false`
  - another JSON object
  - array (denoted by square brackets `[ ]`) of JSON objects
  - `null`

JSON example


```
{
   "full_text": "Yesterday, #GWU students, faculty,
staff...https://t.co/8Tz29odc11",
   "favorite_count": 56,
   "truncated": false,
   "entities": {
      "user_mentions": [],
      "hashtags": {
         "indices": [11, 15],
         "text": "GWU"
      }
   }
}
```

Tweets are JSON too

- Example: go.gwu.edu/emse4197sampletweet
- Twitter's guide to the structure of a tweet: developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object

The Twitter API

Example Twitter API methods

- Get a tweet: GET users/lookup
- Post a tweet: POST statuses/update
- Search tweets: GET search/tweets
- Follow a user: POST friendships/create
- Get user info: GET users/lookup
- Get trends near a location: GET trends/place

More: developer.twitter.com/en/docs

Most useful API methods for collecting tweets

- User timeline: GET statuses/user_timeline
- Search: GET search/tweets
- Filter stream: POST statuses/filter

User timeline: GET statuses/user_timeline

- Gets most recent tweets posted by a user.
- Limited to last 3,200 tweets.
- Returns 200 at a time, so must page.
- Rate limit: 900 tweets per 15 minutes
- `https://api.twitter.com/1.1/statuses/user_timeline.json?screen_name=gelmanlibrary&max_id=829886156334571 5`

Search: GET search/tweets

- Search recent tweets.
  - Sampling of tweets from last 7 days.
  - Query by keyword, phrases, hashtags, author, date, more.
- Returns up to 100 at a time, so must page.
- Not the same as search on Twitter website.
- Rate limit: 180 tweets per 15 minutes
- `https://api.twitter.com/1.1/search/tweets.json?q=%23onlyatgw`

Filter Stream: POST statuses/filter

- Realtime filtering of all public tweets.
  - Filter by keyword, user, or location.
- Continue to receive additional tweets over a single call to API. (No paging.)
- Limits:
  - When high volume, will not receive all tweets.
  - One stream at a time per set of credentials.
- `https://stream.twitter.com/1.1/statuses/filter.json?track=gwu`

Geotagging

- When posting a tweet:
  - Geotagging is opt-in. Only ~2% geotagged.
  - Lat, long or place name (e.g., DC or Middle Earth)
- API support:
  - Search API: Limit to a specified distance of a lat, long.
  - Filter Stream: Limit to a bounding box.

More:
gwu-libraries.github.io/sfm-ui/posts/2017-04-12-geographic-collecting

Acquiring Twitter data sets

Options for acquiring a Twitter dataset

- Use an existing dataset.
- Collect a new dataset.
- Other options:
  - Purchase it from Twitter.
  - Access / purchase from a Twitter service provider.

More:
gwu-libraries.github.io/sfm-ui/posts/2017-09-14-twitter-data

Using existing Twitter data

Using an existing dataset

- DocNow Catalog: www.docnow.io/catalog/
  - Tweet ids only. Will need to hydrate.
- TweetSets: tweetsets.library.gwu.edu/
  - Filter existing datasets collected by GW Libraries.
  - Full tweets as JSON or CSV (when on campus network).
- Other:
  - Data repositories, e.g., Dataverse: dataverse.harvard.edu
  - Kaggle: www.kaggle.com

Datasets collected by GW Libraries

- 2016 U.S. election (280 million tweets)
- 2018 U.S. midterm election
- Congress (all senators and representatives)
- Federal govt (3000 U.S. government accounts)
- News outlets (4500 media organization accounts)
- Hurricane Florence/Harvey / Irma
- Trump Admin officials
- Make America Great Again
- Tax reform
- Immigration & travel ban
- Charlottesville
- Climate change

More ...

Hands-on: TweetSets

Steps we'll perform:

1. Select a source dataset.
2. Filter the source dataset.
3. Create a new dataset.
4. Generate and download dataset derivatives.

Go to tweetsets.library.gwu.edu/

Collecting new Twitter data

Collecting a new dataset

- Command line:
  - Twarc: github.com/docnow/twarc
  - Twurl: github.com/twitter/twurl
- Libraries:
  - Python
    - twarc github.com/DocNow/twarc
    - tweepy: www.tweepy.org/
  - R - rtweet: github.com/mkearney/rtweet

Collecting a new dataset (continued)

- Web application:
  - Social Feed Manager: go.gwu.edu/sfmgw
- Other tools:
  - TAGS (Twitter Archiving Google Sheet) - tags.hawksey.info/

Twurl

- Command line access to Twitter APIs. Tutorial
- Requires a developer account and app credentials
- To search, use GET search/tweets: developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets

```
twurl authorize --consumer-key Etwe247ksBgflP5nUalEfhaeo
--consumer-secret
ZtUpmfsT8ResEmaqiy52Ddihu9FPAiLebuMOmqN0jeQtXe
twurl /1.1/search/tweets.json?q=gwu | jq
```

More: github.com/twitter/twurl

Social Feed Manager software

- Open source software by GW Libraries.
- User interface for collecting, managing, and exporting social media data.
- Collect from Twitter, Tumblr, Flickr, Sina Weibo.
- Libraries run this for their users as a service. (Not typically a local install on your laptop.)

More: go.gwu.edu/sfm

Hands-on: Social Feed Manager

Steps we'll perform:

1. Sign up
2. Request credentials (API keys)
3. Create a collection
4. Perform a harvest
5. Export data

Go to gwsfm-sandbox.wrlc.org

Exporting datasets

- Formats: Excel, CSV, JSON
- Limit by date ranges
- Splits into separate files

Exploring and analyzing Twitter data

Before analysis

- Clean and validate your data
- Examples of why this is necessary:
  - Our 2016 U.S. election collection includes tweets from the Indian election.
  - Our U.S. government collection includes accounts that were deleted, claimed by other users, and tweeting in Russian.

Working with datasets

Jupyter notebooks:

Example w/pandas: bit.ly/2uhN252 (also see here)

R

jq (jq recipes for Twitter data: bit.ly/2t9cStF)

Excel

Ethical considerations

Social media data comes from people

- Consider impact of your work on the creator of the social media.
- Do not have creator's permission for research.
- Impact on creator is balanced against public good of your research.
- Requires judgment call.

More: go.gwu.edu/sfmethics

Data collecting

Be thoughtful collecting social media of:

- Vulnerable individuals (e.g., minors, social activists)
- Sensitive or harmful topics (e.g., questionable behavior, mental illness)
- Geography-based collecting

Data sharing

- Get familiar with platform terms of use.
  - Don’t republish full datasets
  - Share in accordance with terms (e.g., tweet ids only)
  - Consider copyright
- Sharing summary statistics is usually OK.

Publishing

- When possible, get permission from creator for quotes.
- Do not rely on anonymizing posts.

Questions?

Make a consultation appointment:
calendly.com/social-media-consulting-gw

- sfm@gwu.edu
- @liblaura lwrubel@gwu.edu
- @DanKerchner kerchner@gwu.edu