Collecting Social Media Data with Social Feed Manager

February 14, 2017

Dan Kerchner, Justin Littman, Laura Wrubel

The hands-on part of the workshop requires a Twitter account.
Go to http://twitter.com to create one. (You can delete it later.)

Agenda

- Overview of social media APIs and data formats
- Twitter's API in depth
- Datasets available from GW Libraries
- Ethics of social media collecting
- Social Feed Manager walkthrough: hands-on

What's an API?

- Application Programming Interface
- A system's API provides a way for you (or your code) to interact with the system
- Consists of a set of:
  - requests you can make
    - might look like `http://some.api.com/somerequest?thingnumber=15`
  - each with a well-defined response structure
    - can be anything, but data is often returned as XML or **JSON**

Why use an API for working with social media?

- You *really* don't want to scrape it from the web page!
  - It's hard
  - It will break
  - It's incomplete

- But using the API
  - generally gives you exactly what the platform stores
  - can give you useful slices of data you can't get by any amount of scraping
  - gives you social media data in JSON format, which makes it easy to analyze as data

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`

Tweets are JSON too -- Data structure of a Tweet

- Example: http://go.gwu.edu/emse4197sampletweet
- Twitter's "Field Guide" to its API: https://dev.twitter.com/overview/api/tweets

Twitter's APIs

REST APIs (https://dev.twitter.com/rest/public)

- Creating, getting, and updating users, tweets, followers, lists, and more.
- Getting user timeline.
- Lookup user by screen name or user id.
- Getting tweets by tweet ids.

Search API (https://dev.twitter.com/rest/public/search)

Streaming API (https://dev.twitter.com/streaming/public)

- Filter
- Sample

User timeline

Docs: https://dev.twitter.com/rest/reference/get/statuses/user_timeline

Gets most recent tweets for a user.

- User is specified by a screen name or user id.
- Limited to last 3,200 tweets.
- Returns 200 at a time, so must make multiple calls (“paging”) to get full timeline.

Example: GET https://api.twitter.com/1.1/statuses/user_timeline.json?screen_name=socialfeedmgr&max_id=829886156334571524

Limits: 900 tweets per 15 minutes

Search

Docs: https://dev.twitter.com/rest/reference/get/search/tweets

Search recent tweets.

- Sampling of tweets from last 7 days.
- Query by keyword, phrases, hashtags, author, date, sentiment and more.
- Returns up to 100 at a time, so must use paging to get full results.
- Not the same as search on Twitter website.

Example: GET https://api.twitter.com/1.1/search/tweets.json?q=%23onlyatgw

Limits: 180 tweets per 15 minutes

Sample Stream

Docs: https://dev.twitter.com/streaming/reference/get/statuses/sample

Sample of all (public) tweets.

- Estimated to be 0.5-1% of tweets.
- Continue to receive additional tweets over a single call to API. (No paging.)
- Around 3GB per day (compressed).

Example: GET https://stream.twitter.com/1.1/statuses/sample.json
(Notice there are no parameters.)

Limits: Only one stream at a time.

Filter Stream

Docs: https://dev.twitter.com/streaming/reference/post/statuses/filter

Filtering of all (public) tweets.

- Filter by keyword, user, or location.
- Continue to receive additional tweets over a single call to API. (No paging.)

Example: POST https://stream.twitter.com/1.1/statuses/filter.json?track=gwu

Limits:

- Only one stream at a time.
- Number of tweets at a time.

API Summary

Get tweets from the past:

- By user: User timeline
- By query: Search

Get tweets from the present:

- A sample: Sample stream
- By query: Filter stream

APIs can be really confusing/complicated.

- Read the API docs.

What do you want to collect?

What are your research questions?

What data do you need?

- Real-time, recent past, far past
- Do you need to proactively collect or is the data already available?
- Metadata
- Images and other media, websites referenced
- Comments, responses, conversation

What do you plan to do with the data?

Datasets available to GW community

- U.S. Presidential Election
  - 280 million tweets
  - Collected between July 13, 2016 and November 10, 2016
  - Candidates, parties, conventions, debates, hashtags
- Women's March
  - 7.2 million tweets
  - #WomensMarch, #WMW, #WhyIMarch, more
- News outlets
- 115th U.S. Congress

“Social media data is public: it’s fair game.”

Social media data comes from people.

What is the impact on the person
who created the social media,
if I use the data in the way I am planning?

Sharing social media data

Get familiar with platform terms of use. See links within SFM.

- Don’t republish full datasets
- Share in accordance with terms (tweet identifiers only)
- Consider copyright

Think through the ethical aspects of collecting and using social media data. Read more:
https://gwu-libraries.github.io/sfm-ui/resources/ethics

Options for accessing Twitter API

Web:

- API Console (https://dev.twitter.com/rest/tools/console)

Commandline:

- Twarc (https://github.com/docnow/twarc)

Library:

- Python - Tweepy (http://www.tweepy.org/)
- R - rtweet (https://github.com/mkearney/rtweet)

User interface:

- Social Feed Manager

Social Feed Manager

Open source software developed by GW Libraries.

Provides a user interface for collecting, managing, and exporting social media data from Twitter, Tumblr, Flickr, and Sina Weibo.

An instance of SFM is hosted by GW Libraries and available to GW community.

More info:

- SFM info for GW: http://go.gwu.edu/sfmgw
- Project site: http://go.gwu.edu/sfm
- Twitter: @SocialFeedMgr

SFM walkthrough

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 http://go.gwu.edu/sfmsandbox

- This is a sandbox. You can't break anything.
- To access SFM, you must be connected to GW network (on-campus or VPN).

Questions?

Schedule a consultation with us to get started!

libdata@gwu.edu