Acquiring Social Media Data

Justin Littman, Dan Kerchner

February 5, 2018

slides: bit.ly/2FCB3nn

The hands-on part of the workshop requires a Twitter account.
Go to http://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

Agenda

- Bonus: Facebook Graph API
  - Demo: Graph API Explorer
  - Demo: f(b)arc

APIs, Social Media APIs, and their data

What's an API?

- Short for “Application Programming Interface”
- Allows software to interact with a website
  - Compared to a web interface, which allows people to interact with a website.
- API calls consist of:
  - **requests**: *`http://an.api.com/somerequest?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.
- 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 structured 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`

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: http://go.gwu.edu/emse4197sampletweet
- Twitter's guide to the structure of a tweet: https://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: https://developer.twitter.com/en/docs

Twurl

- Like Curl, but for Twitter.
- To search, use GET search/tweets: https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets

```
twurl authorize --consumer-key EHdoTe7ksBgflP5nUalEfhaeo
--consumer-secret
ZtUpemtBkf2cEmaqiy52Ddihu9FPAiLebuMOmqN0jeQtXe
twurl /1.1/search/tweets.json?q=gwu | jq
```

More: https://github.com/twitter/twurl

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

Using existing Twitter data

Using an existing dataset

- DocNow Catalog: http://www.docnow.io/catalog/
  - Tweet ids only. Will need to hydrate.
- TweetSets: https://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: http://dataverse.harvard.edu
  - Kaggle: https://www.kaggle.com

Datasets collected by GW Libraries

- 2016 U.S. election (280 million tweets)
- Congress (all senators and representatives)
- Federal govt (3000 U.S. government accounts)
- News outlets (4500 media organization accounts)
- Hurricane Harvey / Irma
- Healthcare
- Trump Admin officials
- Make America Great Again
- Tax reform
- Immigration & travel ban
- Charlottesville
- Solar Eclipse
- 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 https://tweetsets.library.gwu.edu/

Collecting new Twitter data

Collecting a new dataset

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

Collecting a new dataset (continued)

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

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, and Sina Weibo.
- Intended for organizations to run for their users.

More: http://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 http://gwsfm-sandbox.wrlc.org

Exporting datasets via the UI

- Formats: Excel, CSV, JSON
- Limit by date ranges
- Splits into separate files
- But files must be downloaded and export process is serial

Exporting datasets via the command line

- Formats: JSON
- Can be piped through other tools for filtering / transformation (e.g., jq)
- Within a Docker container; requires shell access to server.
- Can be parallelized for faster export.

Exploring and analyzing Twitter data

Working with datasets

- Jupyter notebooks:
  - Example w/Pandas: http://bit.ly/2uhN252 (also see here)
- ElasticSearch / Logstash / Kibana (ELK stack):
  - For simple exploration, visualization, and analytics
  - Docs: http://sfm.readthedocs.io/en/latest/exploring.html
- Other tools:
  - jq (jq recipes for Twitter data: http://bit.ly/2t9cStF)
  - parallel

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 judgement call.

More: http://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

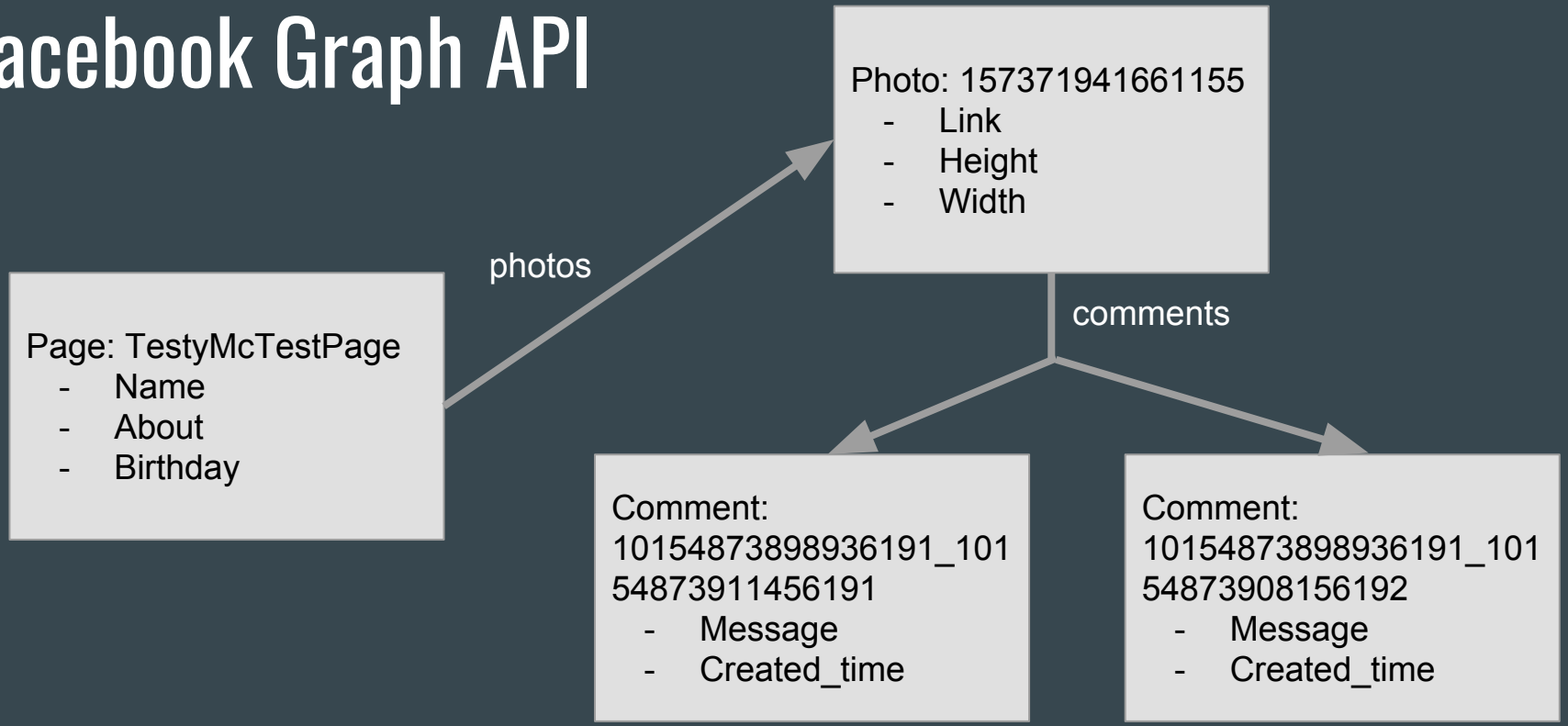
Data sharing

- 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.

Facebook Graph API



Facebook Graph API

`https://graph.facebook.com/v2.11/TestyMcTestpage?fields=about,birthday,photos.limit(100)`

node

fields

connections

- API optimized for retrieving only needed data.
- API is really flaky.

More:

https://gwu-libraries.github.io/sfm-ui/posts/2018-01-02-facebook

Facebook is JSON too

```
{
   about: "This is a page for testing the
Facebook API.",
   "birthday": "01/01/2000",
   "id": "157371584994524"
}
```

More:

https://gist.github.com/justinlittman/05dc05532a1e624adba76892f286ba09

Demo: Graph API Explorer

More: https://developers.facebook.com/tools/explorer

Demo: f(b)arc

- Command line tool and python library.
- Fields and connections are configurable.
- Can recursively collect connected nodes.
- Includes a viewer web application.
- To retrieve a page:

```
python fbarc.py graph page TestyMcTestpage --levels 10
```

More: https://github.com/justinlittman/fbarc

Questions?

More:

- http://go.gwu.edu/gwsfm
- @SocialFeedMgr
- libdata@gwu.edu

Or:

- @justin_littman | @DanKerchner
- justinlittman@gwu.edu | kerchner@gwu.edu