Collecting and Analyzing Twitter Data
Best Practices

Ramon Villa-Cox
rvillaco@andrew.cmu.edu

The CASOS Center
School of Computer Science, Carnegie Mellon
Summer Institute 2020

Collecting Data on the Web in General

- What platform should I use?
- Should I collect everything?
- How much should I pay?
- Is my collection method ethical?
- Can I share this data?
- Real-time vs. Historical
- API vs. Scraping
Why Twitter?

• One popular social website---more users, more data
• Various ways to collect data---depends on your research purpose.
• Easy to collect, though there are certain limitations to share the data.

Ways to Collect Twitter Data

- Streaming API
  - Following users
  - Following keywords
  - Following locations (geo-bounding boxes)
  - Sampling tweets without filters
  - Get follower ids
  - Get followee ids
  - Get user timeline

- Search by users (certain rate limits)
What format is my data in

- JSON!
- Related question, what is it?
- JSON is a simple format for sharing unstructured data

```
{
    "this_is_a_key" : "This is a value",
    "user_screen_name" : "dancer_geoff_44882",
    "tweet_text" : "Man Kenny's lectures are pretty terrible, amirite? #CASOS"
}
```

- Typically – one JSON “object” per tweet/line of file

---

Tweets to meta-networks

Twitter JSON Structure

- Text
- Coordinates
- Created_at
- favorite_count
- favorited
- id
- Lang
- User (another JSON object)
- ...

Full list of fields at:
https://dev.twitter.com/overview/api/tweets

Networks

- User x User
  - Mention
  - Following
  - Retweet
- Hashtag Graphs
  - Co-occurrence
  - Bipartite graph: user x hash tag
- Node attributes
  - Profile features: following count, creation date,...
  - Language patterns, geo coord., etc
How to do it?

• Option 1: Use some commercial data collecting services
• Option 2: Get the ASU team to do it (TweetTracker)
• Option 3: Do it yourself!
  – What you’ll need:
    • API credentials (https://apps.twitter.com/)
    • Find a programming language you’re comfortable with
      – R - twitteR package
      – Python – tweepy is the most popular tool
      – Java – Hosebird is Twitter’s own tool for connecting to the streaming API

Common approaches

• Track all tweets within the U.S. for 6 months
• Follow 1000 users I think are interesting for 6 months, do a network analysis
• Follow #coronavirus for 6 months, do a network analysis
• …
Common practice 1

1. Hook in to the Streaming API with keywords and/or bounding box for a bit
2. Find users that are “interesting”
3. Use the Search API to collect all of these users’ data
4. Try to get rid of bots, celebrities, etc.

Pros: Relatively easy, fast
Cons: Results are limited to these streaming keywords/locations. The resulting mentioning/retweeting networks are usually sparse.

Common practice 2---snowball sampling

1. Start with a set of seed users of interest
2. Collect timelines for these users
3. Find new users within one-step connection (mentioning, following, retweeting)
4. Repeat step 1.

Pros: Get comprehensive social links for a group of users.
Cons: Time consuming, relies on the choice of seed users.
Demo

- Step 1: Go to https://apps.twitter.com/, and apply for a developer account. The process can take some days to complete.
- Step 2: Install tweepy for Python,
  
  pip install tweepy --user
  Or (if you use anaconda as a package manager)
  conda install -c conda-forge tweepy
- Step 3: Fill the access token and filtering criteria in stream.py
  
  The code takes in a list of strings (queries).
  Elements in the list are searched as an OR query, words in an element constitute an AND query.
- Step 4: Run stream.py
  
  python stream.py