Collecting Twitter Data

Binxuan Huang
binxuanh@cs.cmu.edu

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?

It's easy to collect and it's useful for some things

Ways to Collect Twitter Data

• Questions you have to ask:
  - Do I want this in real-time?
  - Do I want to collect historical data?
  - Am I interested in particular users?
  - Am I interested in particular keywords?
  - Am I interested in a particular location?
Collecting Twitter Data

**Streaming API**
- Post statuses/filter
  - Following users
  - Following terms
  - Following Geo-bounding boxes
- Get statuses/sample (Firehose, 1% random sample)

**Search API (Snowball searches)**
- User following ties
- User timeline

Collection Gotchas - Bots

Collection Gotchas - Is the 1% unbiased?

Probably not.


Collection Gotchas - Snowball Sampling

- Who is the most central node in a snowball search?
- What nodes are you likely to miss in a snowball search?
- What nodes are you likely to not miss in a snowball search?
- What does this tell you about, e.g., the degree distribution of your network?
Collection Gotchas - Retweets

- Retweets are connected to the original tweet
- This means RTs of RTs get lost (maybe not anymore?)

How to do it?

- Option 1: Pay a lot of money
- Option 2: Get the ASU team to do it
- Option 3: Do it yourself!
  - What you’ll need:
    - API credentials (https://apps.twitter.com/, show how...)
    - Find a programming language you’re comfortable with
      - R - TwittR package (only for search API, AFAIK)
      - Python - tweepy is the most popular tool
      - Python - twitter_dm is Kenny’s tool for the search API
      - Java - Hosebird is Twitter’s own tool for connecting to the streaming API
What format is my data in

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

```json
{
    "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**
- coordinates
- Created_at
- favorite_count
- favorited
- id
- Lang
- ...

Full list of fields at: [https://dev.twitter.com/overview/api/tweets](https://dev.twitter.com/overview/api/tweets)

**Networks**
- User x User
  - Mention
  - Following
  - Semantic
- Hashtag Graphs
  - Co-occurrence
  - Bipartite graph: user x hash tag
- Node attributes
  - Profile features: following count, creation date,...
  - Language patterns, geo coord., etc
One approach

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 if I can help it


Problems?

Two approach

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.


Problems?
Some Made-up 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 #ferguson for 6 months, do a network analysis
• …