ECPR Methods Summer School: Automated Collection of Web and Social Data

Pablo Barberá

School of International Relations
University of Southern California
pablobarbera.com

Networked Democracy Lab
www.netdem.org

Course website:

github.com/pablobarbera/ECPR-SC103

Twitter APIs

Two different methods to collect Twitter data:

1. REST API:

Twitter APIs

Two different methods to collect Twitter data:

1. REST API:
   - Queries for specific information about users and tweets

Twitter APIs

Two different methods to collect Twitter data:

1. REST API:
   - Queries for specific information about users and tweets
   - Search recent tweets

Twitter APIs

Two different methods to collect Twitter data:

1. REST API:
   - Queries for specific information about users and tweets
   - Search recent tweets
   - Examples: user profile, list of followers and friends, tweets generated by a given user (“timeline”), users lists, etc.

Twitter APIs

Two different methods to collect Twitter data:

1. REST API:
    - Queries for specific information about users and tweets
    - Search recent tweets
    - Examples: user profile, list of followers and friends, tweets generated by a given user (“timeline”), users lists, etc.
    - R library: netdemR (also twitteR, rtweet)

Twitter APIs

Two different methods to collect Twitter data:

1. REST API:
   - Queries for specific information about users and tweets
   - Search recent tweets
   - Examples: user profile, list of followers and friends, tweets generated by a given user ("timeline"), users lists, etc.
   - R library: netdemR (also twitteR, rtweet)
2. Streaming API:

Twitter APIs

Two different methods to collect Twitter data:

1. REST API:
   - Queries for specific information about users and tweets
   - Search recent tweets
   - Examples: user profile, list of followers and friends, tweets generated by a given user (“timeline”), users lists, etc.
   - R library: netdemR (also twitteR, rtweet)
2. Streaming API:
   - Connect to the “stream” of tweets as they are being published

Twitter APIs

Two different methods to collect Twitter data:

1. REST API:
   - Queries for specific information about users and tweets
   - Search recent tweets
   - Examples: user profile, list of followers and friends, tweets generated by a given user (“timeline”), users lists, etc.
   - R library: netdemR (also twitteR, rtweet)
2. Streaming API:
   - Connect to the “stream” of tweets as they are being published
   - Three streaming APIs:

Twitter APIs

Two different methods to collect Twitter data:

1. REST API:
   - Queries for specific information about users and tweets
   - Search recent tweets
   - Examples: user profile, list of followers and friends, tweets generated by a given user (“timeline”), users lists, etc.
   - R library: netdemR (also twitteR, rtweet)
2. Streaming API:
   - Connect to the “stream” of tweets as they are being published
   - Three streaming APIs:
     2.1 Filter stream: tweets filtered by keywords

Twitter APIs

Two different methods to collect Twitter data:

1. REST API:
   - Queries for specific information about users and tweets
   - Search recent tweets
   - Examples: user profile, list of followers and friends, tweets generated by a given user (“timeline”), users lists, etc.
   - R library: netdemR (also twitteR, rtweet)
2. Streaming API:
   - Connect to the “stream” of tweets as they are being published
   - Three streaming APIs:
     - 2.1 Filter stream: tweets filtered by keywords
     - 2.2 Geo stream: tweets filtered by location

Twitter APIs

Two different methods to collect Twitter data:

1. REST API:
   - Queries for specific information about users and tweets
   - Search recent tweets
   - Examples: user profile, list of followers and friends, tweets generated by a given user (“timeline”), users lists, etc.
   - R library: netdemR (also twitteR, rtweet)
2. Streaming API:
   - Connect to the “stream” of tweets as they are being published
   - Three streaming APIs:
     - 2.1 Filter stream: tweets filtered by keywords
     - 2.2 Geo stream: tweets filtered by location
     - 2.3 Sample stream: 1% random sample of tweets

Twitter APIs

Two different methods to collect Twitter data:

1. REST API:
   - Queries for specific information about users and tweets
   - Search recent tweets
   - Examples: user profile, list of followers and friends, tweets generated by a given user (“timeline”), users lists, etc.
   - R library: netdemR (also twitteR, rtweet)
2. Streaming API:
   - Connect to the “stream” of tweets as they are being published
   - Three streaming APIs:
     - 2.1 Filter stream: tweets filtered by keywords
     - 2.2 Geo stream: tweets filtered by location
     - 2.3 Sample stream: 1% random sample of tweets
   - R library: streamR

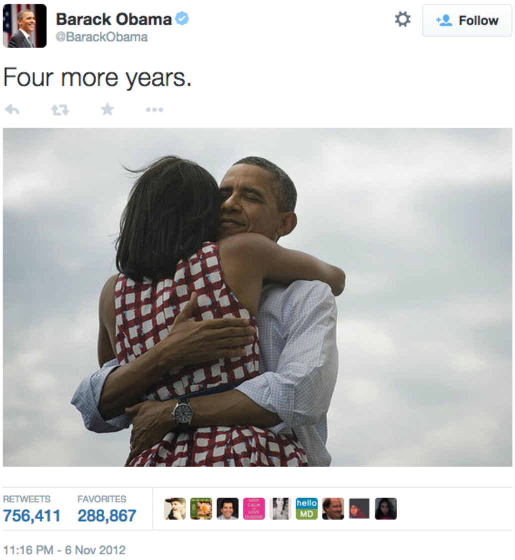
Twitter APIs

Two different methods to collect Twitter data:

1. REST API:
   - Queries for specific information about users and tweets
   - Search recent tweets
   - Examples: user profile, list of followers and friends, tweets generated by a given user ("timeline"), users lists, etc.
   - R library: netdemR (also twitteR, rtweet)
2. Streaming API:
   - Connect to the "stream" of tweets as they are being published
   - Three streaming APIs:
     - 2.1 Filter stream: tweets filtered by keywords
     - 2.2 Geo stream: tweets filtered by location
     - 2.3 Sample stream: 1% random sample of tweets
   - R library: streamR

Important limitation: tweets can only be downloaded in real time (exception: user timelines, $\sim$ 3,200 most recent tweets are available)

Anatomy of a tweet



Barack Obama
@BarackObama

Four more years.

RETWEETS 756,411

FAVORITES 288,867

11:16 PM - 6 Nov 2012

Anatomy of a tweet

Tweets are stored in JSON format:

```
{ "created_at": "Wed Nov 07 04:16:18 +0000 2012",
  "id": 266031293945503744,
  "text": "Four more years. http://t.co/bAJE6Vom",
  "source": "web",
  "user": {
    "id": 813286,
    "name": "Barack Obama",
    "screen_name": "BarackObama",
    "location": "Washington, DC",
    "description": "This account is run by Organizing for Action staff.
        Tweets from the President are signed -bo.",
    "url": "http://t.co/8aJ56Jcemr",
    "protected": false,
    "followers_count": 54873124,
    "friends_count": 654580,
    "listed_count": 202495,
    "created_at": "Mon Mar 05 22:08:25 +0000 2007",
    "time_zone": "Eastern Time (US & Canada)",
    "statuses_count": 10687,
    "lang": "en" },
  "coordinates": null,
  "retweet_count": 756411,
  "favorite_count": 288867,
  "lang": "en"
}
```

Streaming API

- Recommended method to collect tweets

Streaming API

- Recommended method to collect tweets
- Potential issues:

Streaming API

- Recommended method to collect tweets
- Potential issues:
  - Filter streams have same rate limit as spritzer: when volume reaches 1% of all tweets, it will return random sample

Streaming API

- Recommended method to collect tweets
- Potential issues:
  - Filter streams have same rate limit as spritzer: when volume reaches 1% of all tweets, it will return random sample
  - Stream connections tend to die spontaneously. Restart regularly.

Streaming API

- Recommended method to collect tweets
- Potential issues:
  - Filter streams have same rate limit as spritzer: when volume reaches 1% of all tweets, it will return random sample
  - Stream connections tend to die spontaneously. Restart regularly.
  - Lots of invalid content in stream. If it can’t be parsed, drop it.

Streaming API

- Recommended method to collect tweets
- Potential issues:
  - Filter streams have same rate limit as spritzer: when volume reaches 1% of all tweets, it will return random sample
  - Stream connections tend to die spontaneously. Restart regularly.
  - Lots of invalid content in stream. If it can’t be parsed, drop it.
- My workflow:

Streaming API

- Recommended method to collect tweets
- Potential issues:
  - Filter streams have same rate limit as spritzer: when volume reaches 1% of all tweets, it will return random sample
  - Stream connections tend to die spontaneously. Restart regularly.
  - Lots of invalid content in stream. If it can’t be parsed, drop it.
- My workflow:
  - Amazon EC2, cloud computing

Streaming API

- Recommended method to collect tweets
- Potential issues:
  - Filter streams have same rate limit as spritzer: when volume reaches 1% of all tweets, it will return random sample
  - Stream connections tend to die spontaneously. Restart regularly.
  - Lots of invalid content in stream. If it can't be parsed, drop it.
- My workflow:
  - Amazon EC2, cloud computing
  - Cron jobs to restart R scripts every hour.

Streaming API

- Recommended method to collect tweets
- Potential issues:
    - Filter streams have same rate limit as spritzer: when volume reaches 1% of all tweets, it will return random sample
    - Stream connections tend to die spontaneously. Restart regularly.
    - Lots of invalid content in stream. If it can't be parsed, drop it.
- My workflow:
    - Amazon EC2, cloud computing
    - Cron jobs to restart R scripts every hour.
    - Save tweets in .json files, one per day.

Streaming API

- Recommended method to collect tweets
- Potential issues:
  - Filter streams have same rate limit as spritzer: when volume reaches 1% of all tweets, it will return random sample
  - Stream connections tend to die spontaneously. Restart regularly.
  - Lots of invalid content in stream. If it can't be parsed, drop it.
- My workflow:
  - Amazon EC2, cloud computing
  - Cron jobs to restart R scripts every hour.
  - Save tweets in .json files, one per day.
  - For large .json files, preprocess with python (see: github.com/pablobarbera/pytwools)

Sampling bias?

Morstatter et al, 2013, *ICWSM*, “Is the Sample Good Enough? Comparing Data from Twitter’s Streaming API with Twitter’s Firehose”:

- 1% random sample from Streaming API is not truly random
- Less popular hashtags, users, topics... less likely to be sampled
- But for keyword-based samples, bias is not as important

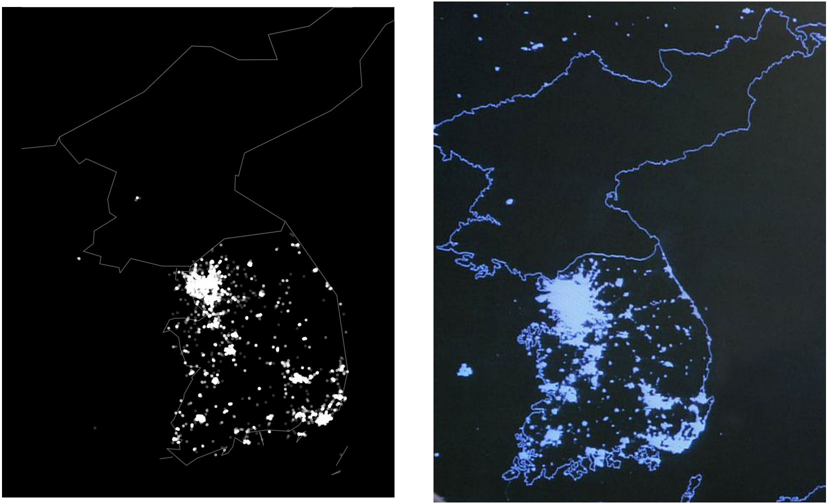
Sampling bias?

Morstatter et al, 2013, *ICWSM*, “Is the Sample Good Enough? Comparing Data from Twitter’s Streaming API with Twitter’s Firehose”:

- 1% random sample from Streaming API is not truly random
- Less popular hashtags, users, topics... less likely to be sampled
- But for keyword-based samples, bias is not as important

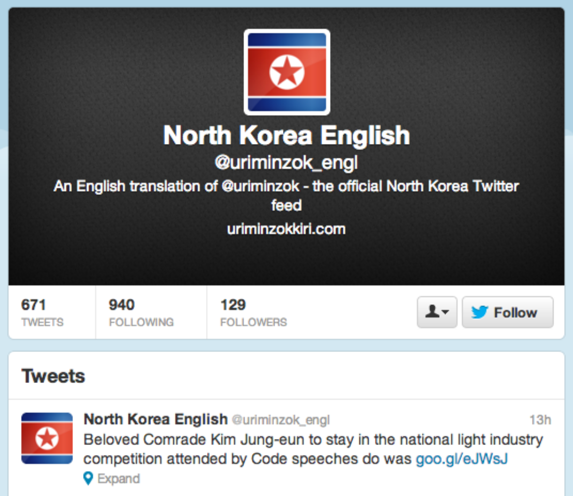
González-Bailón et al, 2014, *Social Networks*, “Assessing the bias in samples of large online networks”:

- Small samples collected by filtering with a subset of relevant hashtags can be biased
- Central, most active users are more likely to be sampled
- Data collected via search (REST) API more biased than those collected with Streaming API



Tweets from Korea: 40k tweets collected in 2014 (left)
Korean peninsula at night, 2003 (right). Source: NASA.

Who is tweeting from North Korea?



North Korea English

@uriminzok_engl

An English translation of @uriminzok - the official North Korea Twitter feed

uriminzokkiri.com

671	940	129
TWEETS	FOLLOWING	FOLLOWERS

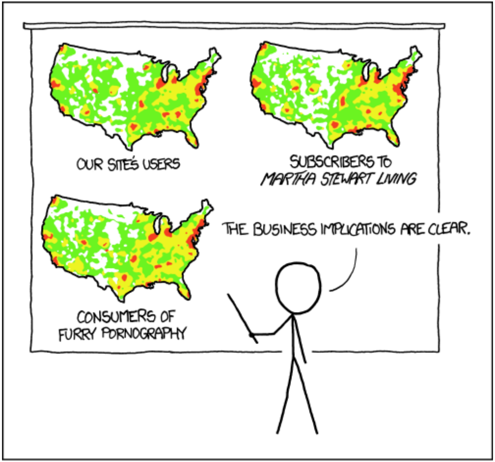
Tweets

North Korea English @uriminzok_engl 13h

Beloved Comrade Kim Jung-eun to stay in the national light industry competition attended by Code speeches do was goo.gl/eJWsJ

Twitter user: @uriminzok_engl

But remember...


OUR SITE'S USERS
SUBSCRIBERS TO *MARTHA STEWART LIVING*
CONSUMERS OF FURRY PORNOGRAPHY
THE BUSINESS IMPLICATIONS ARE CLEAR.

PET PEEVE #208:
GEOGRAPHIC PROFILE MAPS WHICH ARE BASICALLY JUST POPULATION MAPS