

EITM Europe Summer Institute: Social Media Research

Pablo Barberá

London School of Economics

`www.pablobarbera.com`

Course website:

pablobarbera.com/eitm

Social media data

Twitter data

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: tweetscores (also twitterR, 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, ~ 3,200 most recent tweets are available)

Anatomy of a tweet



Barack Obama ✓
@BarackObama



 Follow

Four more years.

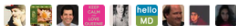


RETWEETS

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
- ▶ 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.
- ▶ My workflow:
 - ▶ Amazon EC2, cloud computing
 - ▶ Cron jobs to restart R scripts every hour.
 - ▶ Save tweets in .json files, one per day.
 - ▶ Will show some examples later

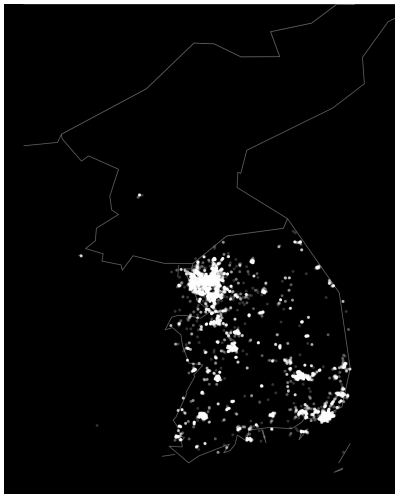
Sampling bias?

[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?



The image shows a screenshot of a Twitter profile page. At the top, there is a profile picture of the North Korean flag. Below it, the name 'North Korea English' is displayed in bold, followed by the handle '@uriminzok_engl'. A bio reads: 'An English translation of @uriminzok - the official North Korea Twitter feed' and the website 'uriminzokkiri.com' is listed. The profile statistics show 671 tweets, 940 following, and 129 followers. A 'Follow' button is visible. Below the profile information, a tweet is shown with the same profile picture and handle, dated 13 hours ago. The tweet text is: 'Beloved Comrade Kim Jung-eun to stay in the national light industry competition attended by Code speeches do was goo.gl/eJWsJ'. There is an 'Expand' link below the tweet.

North Korea English
@uriminzok_engl
An English translation of @uriminzok - the official North Korea Twitter feed
uriminzokkiri.com

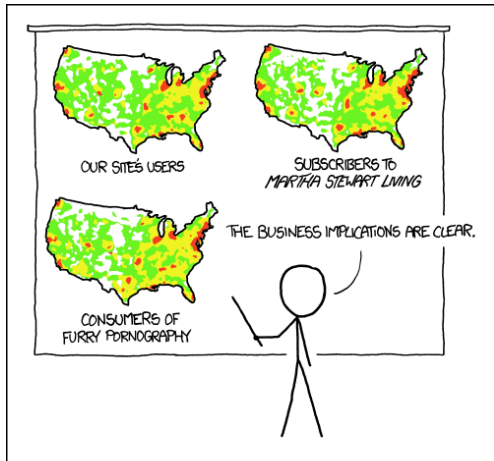
671 TWEETS 940 FOLLOWING 129 FOLLOWERS  

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
 Expand

Twitter user: @uriminzok_engl

But remember...



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

EITM Europe Summer Institute: Social Media Research

Pablo Barberá

London School of Economics

`www.pablobarbera.com`

Course website:

pablobarbera.com/eitm