Collecting and Analyzing Social Media Data with R

Pablo Barberá

pablo.barbera@nyu.edu

Social media and Political Participation Lab
New York University

February 25, 2014

Collecting and Analyzing Social Media Data with R

The tools I have developed:

1. `smappR` package: R tools for the analysis of Twitter data
   - Access the lab database and run most common tasks.
2. `streamR` package: Access to Twitter Streaming API via R.
   - Capture tweets in real time and read tweets in R
3. `Rfacebook` package: Access to Facebook API via R.
   - Download public information about Facebook users, search public posts, scrape pages, retrieve network of friends...

Additional materials:

- "Social Media and Political Participation" NYU-AD J-Term lab sessions
  - Introduction to the analysis of social media data using R
- `smappPy` and other Python tools

smappR package

Functions to access Twitter data in lab server and run most common tasks:

1. Count the number of tweets that mention a set of keywords or were sent within a certain time period, and extract them as a data frame in R
2. Find the most retweeted tweets and the most popular hashtags in a collection
3. Prepare a plot showing volume of tweets over time (days/hours/minutes)
4. Download timeline, followers, and friends of any Twitter user.
5. Access users' profile information and parse location into geographic coordinates

Documentation: github.com/SMAPPNYU/smappR

streamR package

Functions to collect Twitter data from the Streaming API:

1. `filterStream` function will return tweets in real time that:
   - Mention a set of keywords or hashtags
   - Mention a set of users
   - Are sent within a geographic "bounding box"
2. `sampleStream` function will return a 1% random sample of tweets in real time

Tweets are downloaded in JSON format, but can be parsed to data frames in R (or exported to the lab server)

Documentation: github.com/pablobarbera/streamR

Rfacebook package

Functions to collect Facebook data from the Graph API:

1. Public Facebook posts that mention a certain keyword
2. User information (gender, location, language, full name, profile picture)
3. Posts on public pages (content of post, as well as list of likes and comments)
4. Friends information (network, likes, checkins, newsfeed...)

Data can be saved as `.csv` files or R objects.

Documentation: github.com/pablobarbera/Rfacebook

J-term materials

Introduction to analysis of social media data using R:

- **Lab 3**. Collecting and analyzing Twitter data: using the Streaming API to download tweets, finding most popular tweet, doing a word cloud and a map of tweets, etc.
- **Lab 4**. Collecting and analyzing Facebook data: searching public Facebook posts, getting user information, scraping pages, plotting number of likes over time...
- **Labs 5 & 6**. More advanced examples.

Materials: github.com/SMAPPNYU/NYU-AD-160J

smappPy

Python library developed by SMaPP programmers Peihong Chai and Duncan Penfold-Brown.

Similar functionality as `smappR`, but using python.

Additional functions:

- Extract URLs and images shared on Twitter
- Utilities for text parsing and cleaning
- Build and display retweet networks
- Measure naive sentiment in tweets

In general python is faster than R, specially when working with large collections.
Documentation: github.com/SMAPPNYU/smappPy