Social Media and Political Participation

Lab 6

pablo.barbera@nyu.edu

January 20, 2015

Today

- Summary of lab sessions
- Additional examples of Twitter and Facebook data analysis
- Crowdsourcing and machine learning
- Analysis of Instagram data

- `lab3_collecting_tweets.R`
  - Collecting tweets filtered by keywords
  - Collecting tweets filtered by location
  - Collecting random sample of tweets
  - Collecting tweets by a given user
- `lab3_analyzing_tweets.R`
  - Reading tweets in JSON format
  - Analyzing key variables from tweets
  - Wordcloud of tweet text
  - Map of geolocated tweets
- `lab5-twitter.R`
  - Finding most common hashtags
  - Hashtag wordcloud
  - Plot of number of tweets over time
  - Finding top tweets
- `lab6-examples.R`
  - Counting number of tweets with a picture, or that are retweets
  - Subsetting data from a period of time

- `lab4_collecting_facebook_data.R`
  - Scraping a public Facebook page
  - Find popular posts on a Facebook page
  - Collecting pages' likes data
  - Collecting pages' comments data
- `lab4_analyzing_facebook_data.R`
  - Plot with number of likes over time
  - Visualizing comments on a page with a word cloud
- `lab5-facebook.R`
  - Loading a dataset of Facebook posts
  - Finding posts that mention specific words
  - Wordcloud of posts messages
  - Subsetting data from a period of time
- `lab6-examples.R`
  - Plot with number of posts over time

R code

Advanced Examples of Twitter Data Analysis

The R script `lab6_examples.R` shows how to:

- Count tweets that contain a picture
- Find tweets that mention specific words
- Subsetting tweets by date
- Visualize number of Facebook posts over time

Crowdsourcing and machine learning

Purpose of SMaPP lab: understanding how the use of social media platforms affect political participation.

Two dimensions:

1. Social media as data:
    - Digital traces facilitate measurement of human behavior
2. Social media as a variable
    - Online platforms reduce cost of collective action, facilitate information diffusion and coordination

Both dimensions require classification of large datasets of social media posts into categories.

But human coding is expensive at a large scale. How do we solve this? Crowdsourcing and machine learning.

Crowdsourcing and machine learning

Crowdsourcing

- Individuals code a random sample of posts into categories
- "Wisdom of crowds": multiple coders to increase precision

Machine learning

- "Train" a classifier on this small set of data to learn what words are associated with each latent category
- Apply what we learn to classify the rest of the dataset

Example: coding a set of tweets about the Millions March in NYC, in `lab6-coding-task.R`

Bonus: analyzing Instagram data

What is available through Instagram API:

- Search and download pictures that mention a given hashtag on its caption, or that were sent from a specific location
- Collect information about these pictures (creation date, caption, author, filter, . . .)
- Download pictures sent by a given user
- Count number of photos that mention a specific hashtag