Alexandra Plassaras

QMSS GR 5999

1/24/17

Assignment 1 - Data Source and Description

For my thesis I am interested in conducting sentiment analysis on emojis and comparing this with sentiment analysis on text within a single tweet. The data I will use will be from Twitter. More specifically this data consists of Tweets with specific hashtags collected from February 7th, 2016 to April 2nd, 2016 that referred to the presidential primaries that were taking place earlier last year. This data set was collected for QMSS's Data Processing and Data Visualization class (QMSS G4063) during the Spring 2016 semester. Below are links to Data Processing and Data Visualization's github page as well as a link to the original JSON data containing all tweets scraped.

- https://github.com/hassanpour/QMSS_G4063
- https://www.dropbox.com/sh/zyy9tsvibrl4d63/AAAQ6D3h0Kksxb8EeVH2RSSAa/tweets_geo_all.json?dl=0#

No experiments or surveys will be created and implemented, nor will any field research be conducted. Instead I have chosen to use already available data for this project due to its low (as in non-existent) cost and because I have prior access to this large data set.

Approximately 1.37 GB of tweets were collected during this time period. Of the tweets that were collected, 1,816,475 (approximately 318.4 MB) were captured that contained geo-location data. The geo-location attribute will be useful to filter out all tweets that were not sent from the United States. As this study is looking at American sentiment in both text and emoji making this assumption ensures a higher likelihood that

the tweets we will be looking at are coming from Americans or at the very least tweets within the United States. Further analysis is needed to determine 1) how many tweets contain emojis and 2) what kinds of emojis are being used. The tweets used in this study were scraped via Twitter's Streaming API during a 56-day period and information regarding both meta data about the tweets as well as the tweet and user data were collected. A total of 50 unique variables associated for each tweet have been collected but this research will only look at a small subset of these variables. The full list of variables collected for each tweet instance are shown below in Figure 1.

Figure 1 – All variable names within a single tweet

```
> colnames(tweetsUS)
[1] "X.1"                       "id_str"
[3] "idx"                       "text"
[5] "created_at"                "screen_name"
[7] "user_lang"                 "truncated"
[9] "retweeted"                 "favorite_count"
[11] "verified"                  "user_id_str"
[13] "source"                    "followers_count"
[15] "in_reply_to_screen_name"   "location"
[17] "retweet_count"             "favorited"
[19] "utc_offset"                "statuses_count"
[21] "description"               "friends_count"
[23] "user_url"                  "geo_enabled"
[25] "in_reply_to_user_id_str"   "lang"
[27] "user_created_at"           "favourites_count"
[29] "name"                      "time_zone"
[31] "in_reply_to_status_id_str" "protected"
[33] "listed_count"              "place_lon"
[35] "expanded_url"              "place_id"
[37] "full_name"                 "lat"
[39] "country_code"              "place_name"
[41] "url"                       "country"
[43] "lon"                       "place_type"
[45] "place_lat"                 "X"
[47] "Y"                         "STATEFP"
[49] "NAME"                      "COUNT"
```

The variables that will be used in this study are unique identifiers for each tweet, usernames, tweet contents and locations of each tweet (*idx, screen_name, text, place_lon, place_lat*). The tweets were selected from Twitter's API based on their reference to one of any of the following candidates – Hillary Clinton, Bernie Sanders, Ted Cruz, Donald Trump, and Marco Rubio. Nicknames and references to particular candidates were also

included such as Trumpf, Hillary and Cruz. Figure 2 below shows the full list of identifiers used to pull tweets from Twitter's API.

Figure 2 – Breakdown of Identifiers for each Candidate

Candidate's Full Name	**Key words associated to locate Tweets**
Hillary Clinton	Clinton, clinton, Hillary, hillary, Hillaryclinton, hillaryclinton, Hillary Clinton, hillary clinton
Bernie Sanders	Berniesanders, berniesanders, Bernie Sanders, bernie sanders, Bernie, Bernie, Sensanders, sensanders
Ted Cruz	Cruz, cruz, Ted, ted, Tedcruz, tedcruz, Ted Cruz, ted cruz
Donald Trump	Donaldtrump, donaldtrump, Donald Trump, donald trump, Trump, trump, Donald, Donald, Trumpf, trumpf
Marco Rubio	Marcorubio, marcorubio, Marco Rubio, marco rubio