Report for Homework 1

CS 7930 Social Media Mining

Spring 2016

Gopal Menon
Computer Science Department
Utah State University
Logan, UT
gopal.menon@aggiemail.usu.edu

1. COLLECTING TWITTER DATA

Tweets were collected by running a Python script containing Twitter RestAPI calls. When the geolocation was specified in the Search API, no tweets were returned. So it was decided to collect tweets without geolocation and later filter out tweets from outside USA. Specifying the end date for tweets in the Search API also resulted in no tweets being returned. So it was decided to search for tweets using only the start date and tweet language. An example is shown below.

```
search_results = twitter_api.search.tweets
(q = 'donald trump
 since:2015-12-16',
  lang='en')
print json.dumps(search_results)
```

As can be seen above, tweets in English, from December 16th, 2016, containing the string *donald trump* were retrieved and then converted to JSON for easier parsing.

Each Search API request returned 15 tweets. Out of these tweets, only around 4 to 5 had user profile location within USA. As a result around 575 set of tweets needed to be collected for each candidate in order to obtain 2500 tweets per candidate with location information. Each set of retrieved tweets had metadata that specified the maximum id of the tweets that were received. This maximum id was used to retrieve the next set of tweets using the Search API by specifying the *since_id* : parameter as the maximum id from the set of tweets received previously. A delay of 62 seconds was added between subsequent retrievals so that the retrieval rate was within the limit required by Twitter.

Some sample tweets are shown in Appendix A.

2. PREPROCESSING THE DATA

The set of tweets returned by the Search API were stored as text files. These text files were picked up by a parsing program that extracted information from tweets and put it into an array.

The first step in the parsing program was to check if the user location in the profile was inside USA. This was done by searching for the state abbreviation and state name inside the user location. Those that were not found to be inside USA were rejected. The JSON tweet data was parsed to extract tweet components. Shown below is an example for extraction of the tweet text.

```
current_tweet.text =
candidate_tweets_json[0]['statuses'][count]['text']
```

The following were extracted from each tweet in a similar manner as shown above.

- User Location
- State Code
- Tweet creation timestamp
- Tweed ID
- Tweet Text
- User Screen Name
- User ID
- Tweet Place
- Number of Friends
- Number of Followers
- Tweet Languge

2.1 Exploratory Analysis

2.2 Winner by State

The number of tweets per candidate were summarized by state. This data is shown in Appendix B along with the computation for the electoral votes per state. The winner is the candidate with the most votes and gets all the electoral votes. In the case of a tie, the votes are split and an integer value is assigned.

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

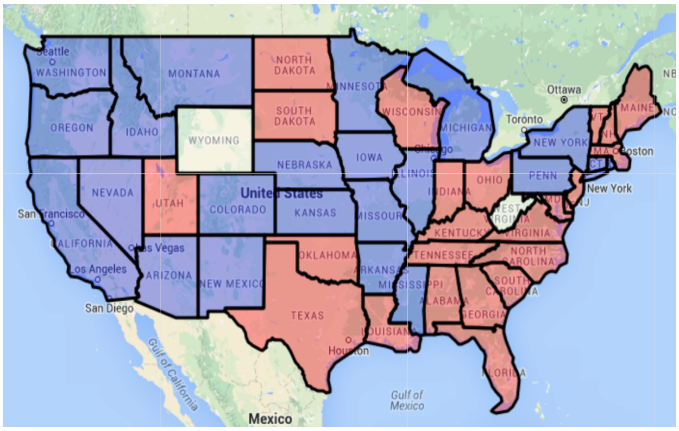

Figure 1: States won by the candidates - red for Trump and Blue for Clinton.

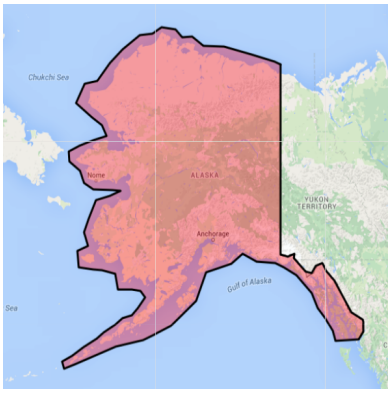

Figure 2: Alaska winner - Trump.

The states won by each candidate are shown in the traditional party colors - red for the Republicans and blue for the Democrats. This is illustrated in figures 1, 2 and 3. The votes were split for Wyoming and West Virginia and this is the reason that these states have not been marked red or blue. The maps were created using Google Maps.

2.3 Tweets per day over time

I did not get tweets that were distributed over the period of study. All the Clinton tweets were for January 29th and 20th, 2016. These were the days I retrieved the tweets and all the tweets retrieved were for those days. All Trump tweets were for January 30th , 31st and February 1st 2016. Due to this reason I do not have data for the trend for tweets per day over time.

2.4 Number of tweets by time of day

Figure 4 shows the number of tweets by time of day. The x-axis shows the hour of the day and the y-axis shows the corresponding number of tweets.

It seems that Trump tweeters are intermittently active from 2 pm till 8 pm. After 8 pm, they remain active till around 3 am. Then they are inactive till 2 pm.

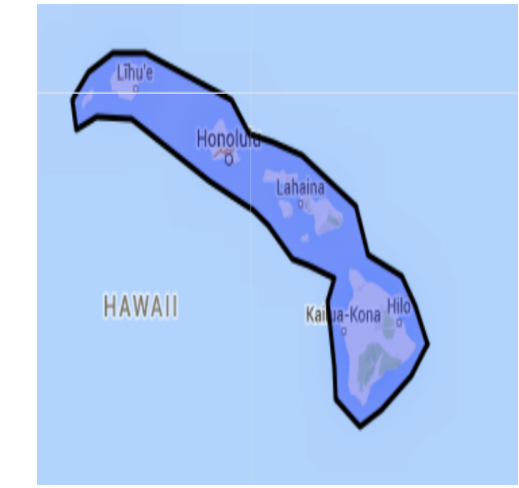

Figure 3: Hawaii winner - Clinton.

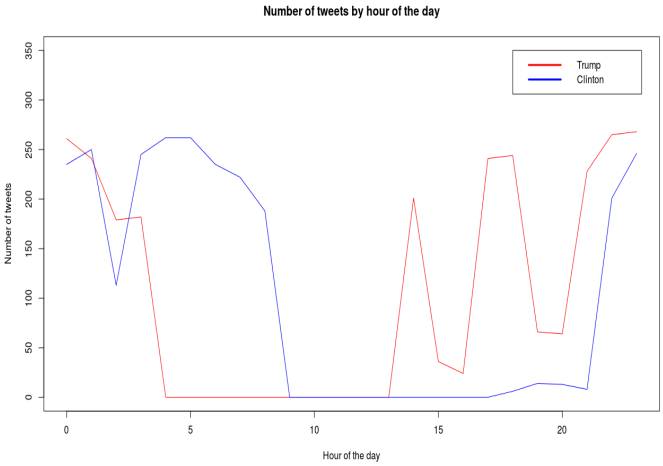

Figure 4: Number of tweets by time of day.

On the other hand Clinton tweeters are inactive from 9 am till 9 pm and active at other times.

Since the tweets that are returned by the Twitter Search API are from around the time the search was run, I have a suspicion that this trend over time of day corresponds to the time the tweets were retrieved. However I am not sure if that is the case or not.

2.5 Projected Winner

According to this analysis, the projected winner is Donald Trump with 269 to 264 electoral votes.

APPENDIX

A. SAMPLE TWEETS

`user_location:`Ohio
`created_at:`Sun Jan 31 17:15:53 +0000 2016
`id_str:`693845165463752705
`text:`RT @samantharonson: Donald Trump has flip-flopped so much that Stephen Colbert hosted a Trump vs. Trump debate https://t.co/Wa8c8WkfTo via ?
`user_screen_name:`christinerod1
`user_id_str:`27403847
`place:`None
`user_friends_count:`286
`user_followers_count:`52
`user_lang:`en

`user_location:`Atlanta, GA
`created_at:`Sun Jan 31 17:25:18 +0000 2016
`id_str:`693847538554789889
`text:`RT @redsteeze: Also a good modern day analogy of Donald Trump and the Republican Party https://t.co/Gv0L0MimV2
`user_screen_name:`WarDamnGunners
`user_id_str:`728693384
`place:`None
`user_friends_count:`533
`user_followers_count:`463
`user_lang:`en

`user_location:`North Carolina
`created_at:`Mon Feb 01 02:17:10 +0000 2016
`id_str:`693981384113831936
`text:`RT @SavageNation: BREAKING POLL: Donald Trump Is winning Latino Republicans: Si Se Puede! In new poll on Latino voters finds Don... https:/?
`user_screen_name:`NewsieSc
`user_id_str:`2946622600
`place:`None
`user_friends_count:`937
`user_followers_count:`489
`user_lang:`en

B. SUMMARIZED DATA

State	Tweets		Electoral Votes	Electoral Votes	
	Clinton	Trump		Clinton	Trump
alabama	31	36	9		9
alaska	6	8	3		3
arizona	53	52	11	11	
arkansas	22	16	6	6	
california	313	258	55	55	
colorado	36	26	9	9	
connecticut	17	15	7	7	
delaware	14	15	3		3
florida	171	201	29		29
georgia	66	83	16		16
hawaii	17	11	4	4	
idaho	16	5	4	4	
illinois	80	57	20	20	
indiana	81	119	11		11
iowa	43	33	6	6	
kansas	27	12	6	6	
kentucky	23	24	8		8
louisiana	20	41	8		8
maine	13	23	4		4
maryland	22	33	10		10
massachuset	44	51	11		11
michigan	68	64	16	16	
minnesota	31	22	10	10	
mississippi	21	9	6	6	
missouri	47	45	10	10	
montana	6	3	3	3	
nebraska	21	16	5	5	
nevada	45	36	6	6	
new hampsh	11	19	4		4
new jersey	52	64	14		14
new mexico	20	9	5	5	
new york	233	229	29	29	
north carolin	51	64	15		15
north dakota	0	4	3		3
ohio	66	73	18		18
oklahoma	27	28	7		7
oregon	55	28	7	7	
pennsylvania	77	62	20	20	
rhode island	5	4	4	4	
south carolin	22	40	9		9

Figure 5: Electoral Votes won by state

south dakota	1	3	3		3
tennessee	27	43	11		11
texas	216	263	38		38
utah	11	12	6		6
vermont	4	8	3		3
virginia	47	57	13		13
washington	190	140	12	12	
west virginia	5	5	5	2	2
wisconsin	26	31	10		10
wyoming	0	0	3	1	1
				264	**269**

Figure 6: Electoral Votes won by state (contd.)