Homework #1

CS 7930, Spring 2016

Task 1: Collecting Twitter Data:

- For collecting the tweets I have used **Tweepy** API, which is implemented in Python.
- Tweepy offers both 'Search' and 'Streaming' schemes for searching the tweets. For this assignment I have used 'Search'.
- On making search request to twitter, twitter sends only 100 tweets in one go. So for getting 1000s of tweets I have made a loop which gets tweets in batches of 100.
- To get the tweets specifically from USA, I have set the geo_search property of API as "USA". And for getting only English tweets, "lang" property has been set to "en".
- Now for searching the tweets for Donald Trump, I have used the keyword 'Trump' and 'Donald Trump' which got me about 2400 and 1800 tweets respectively. Now I removed the duplicate tweets from these two sets, which gave me 2502 unique tweets.
- I did the same thing for Hillary Clinton tweets, using keyword "Hillary" and "Hillary Clinton", and got 2400 unique tweets.
- All the tweets are exported in separate CSV files for Trump and Clinton.

```
__author__ = 'Aditya'

import tweepy
import csv

auth = tweepy.OAuthHandler("paHY2eshhxrBAwY6laiZCSeVy", "7lH0BTCPwFtJvcavIRjXWkCeq2NgHLSC7aQ8phTIxDKxbTQH7S")
auth.set_access_token("4493024472-Ip6tGzAID8uq9JznDYwM5ecdiN92obmhqV6yi5c",
                      "dC1UeiZOXgcchMtd3cg2obuNJtqOLwtvSS3PYjdiUph33")

api = tweepy.API(auth)
places = api.geo_search(query="USA", granularity="country")
place_id = places[0].id

max_tweets = 2500

searched_tweets = []
last_id = -1

csvFile = open('D:\\h1.csv', 'a')
csvWriter = csv.writer(csvFile)

while len(searched_tweets) < max_tweets:
    count = max_tweets - len(searched_tweets)
    for tweet in tweepy.Cursor(api.search, q="place:%s trump" % place_id, lang="en", count=count,
                               max_id=str(last_id - 1)).items():
        csvWriter.writerow([tweet.created_at, tweet.text.encode('utf-8', 'ignore'),
                            tweet.author._json['screen_name'].encode('utf-8', 'ignore'),
                            tweet.author._json['lang'].encode('utf-8', 'ignore'), tweet.author._json['friends_count'],
                            tweet.author._json['followers_count'],
                            tweet.author._json['location'].encode('utf-8', 'ignore')])
```

Task 2: Preprocessing the data:

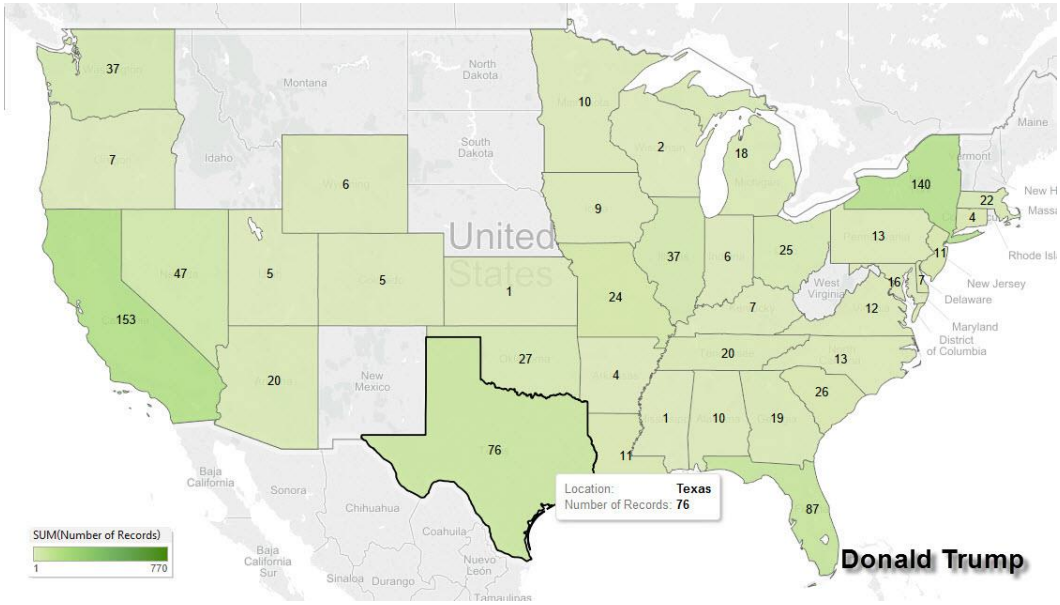
- Tweets collected had lots of inconsistences in the location field. Some of the tweets did not have any location, so had location names which does not exist, like, Earth, Worldwide, -212.23123, etc. Some had city names with states as abbreviation, like, San Diego, CA.
- To make the location consistent, I used Google Refine. I converted all the location to their corresponding state, by using Google Refine's facet and clustering tools (contained various clustering algorithms).
- Most of the tweets had http links embedded to them. I removed all those link using python code.
- Figure below shows the sample of tweets after cleaning:

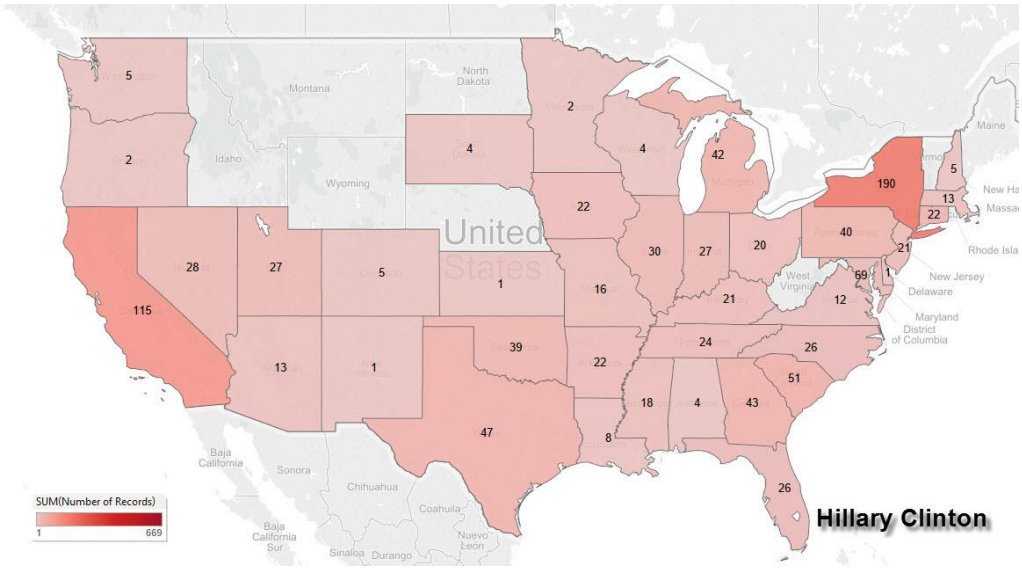
	A	B	C	D	E	F	G	H	I
1	Date/Time	Tweet Id	Tweet	User ID	Name	Language	Friends	Followers	Location
2	2016-01-31	693905046111592000	The big, final Iowa poll is good for Clinton, less so fi	952086582	pbump	en	561	17447	New York
3	2016-01-31	693902534755377000	Donald Trump's sons, Donald JR and Eric, shoot w	2405211005	GingerGibson	en	1922	10118	Washington
4	2016-01-31	693897662114824000	I swiped this from @chrismhood because it's awes	54251487	Die_Obliterator	en	469	648	New York
5	2016-01-31	693897246102802000	RT @FreeJesseJames: Donald Trumpl've met a lot	4747746462	AustinKuti1	en	127	211	
6	2016-01-31	693893061059842000	In almost as sick of the Trump kids as I am of the (	732139172	DonPastor1	en	466	477	California
7	2016-01-31	693892447236001000	RT @pbump: No candidate since 1992, when Gallu	22537143	jjosephwilliams	en	2021	902	Tennessee
8	2016-01-31	693891325523984000	This #Hospitality #job might be a great fit for you: F	16374932	weareteamtrump	en	35	75	Worldwide
9	2016-01-31	693891216836984000	The Donald Trump, up close and personal. #iowaca	4646810601	SoCoGOP	en	773	554	Missouri
10	2016-01-31	693886014738157000	RT @pbump: No candidate since 1992, when Gallu	54251487	ShaneLoew20	en	593	618	Missouri
11	2016-01-30	693884846364557000	@jerseymes no we need a president that isn't Dona	168238759	whyJoe	en	529	594	Orlando, Florida
12	2016-01-30	693884738889719000	@tedcruz sent shaming letters to Iowa voters. Do v	2737178169	Elvisfan1976	en	2881	1299	Oklahoma
13	2016-01-30	693884122498895000	@Queenie051369 , how do you feel about Trump?	54251487	philr260	en	21	1	
14	2016-01-30	693883261894926000	RT @pbump: No candidate since 1992, when Gallu	145312538	Sunshine3324	en	1265	599	Colorado
15	2016-01-30	693883123487080000	If any Veteran thinks trump CARES at all about Vet	145312538	Werdnat	en	2158	586	Texas
16	2016-01-30	693882672993681000	RT @pbump: No candidate since 1992, when Gallu	1528212727	parkerwbriden	en	561	658	Missouri
17	2016-01-30	693882187624636000	RT @pbump: No candidate since 1992, when Gallu	145312538	zefirotorna	en	692	1313	Everywhere
18	2016-01-30	693881867213283000	THE BBQ BRISKET SNADWICH IS SO ðŸ¯¥ðŸ¯¥ð	541284653	NUMP_Trump	en	2482	6746	California
19	2016-01-30	693881321492328000	RT @pbump: No candidate since 1992, when Gallu	118562262	michael_hendrix	en	2051	1836	Washington
20	2016-01-30	693879747298758000	RT @pbump: No candidate since 1992, when Gallu	2464887805	mccanner	en	3231	5376	New York
21	2016-01-30	693879210537021000	This #Hospitality #job might be a great fit for you: Ir	2737178169	weareteamtrump	en	35	75	Worldwide
22	2016-01-30	693878385445965000	RT @pbump: No candidate since 1992, when Gallu	3118335956	WPAResearch	en	3116	4114	Washington
23	2016-01-30	693877530718257000	RT @pbump: No candidate since 1992, when Gallu	783132301	owenblacker	en-gb	3483	3853	London

Task 3: Exploratory Analysis:

❖ **In order to understand the popularity of each politician in a different state, visualize the tweets distribution of each politician on the map using tools like Tableau, Google maps API, basemap, D3, etc. *Report your findings including some figures like snapshots of the maps.***

- To visualize the data I have used Tableau.
- The figures below shows the distribution of tweets containing the keywords for Donald Trump and Hillary Clinton on US map:

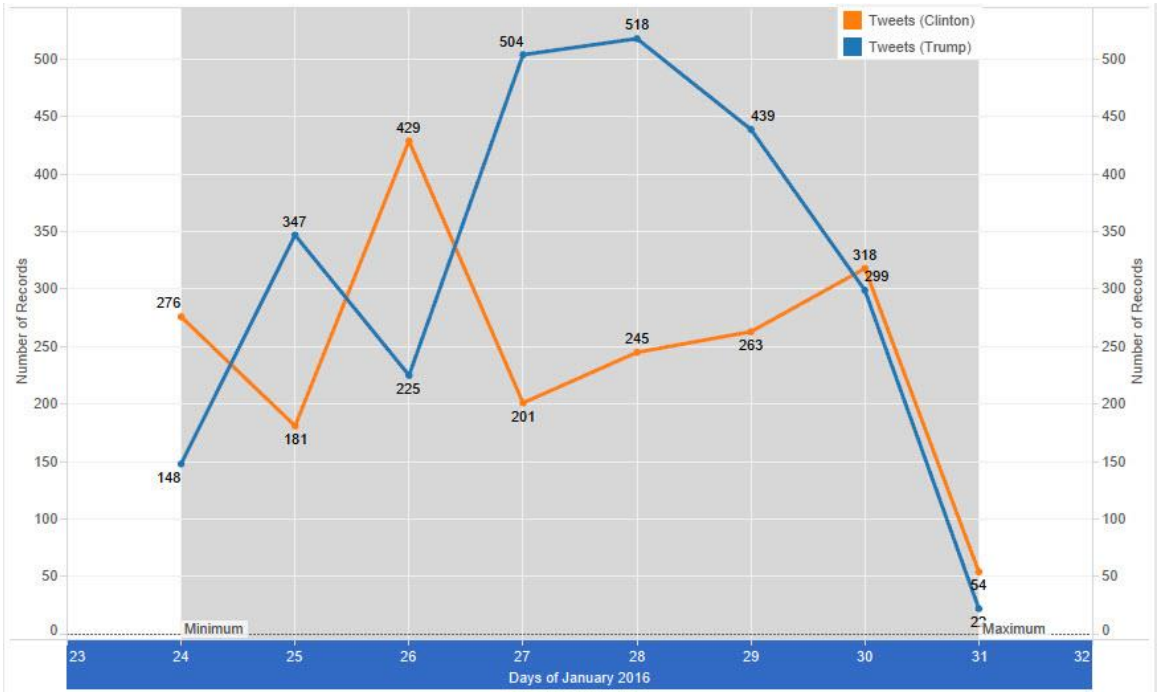




- In the above maps we can see that most number of tweets about Donald Trump has come from California, followed by New York, Texas, Florida and Washington.
- Hillary Clinton seems to be more popular in New York, followed by California, Georgia, South Carolina, and Texas. Basically Clinton seems to be more popular in east USA.
- These two maps only show the number of tweets from different states. This does not mean that Trump/Clinton are going to get more votes in those states. We don't know the sentiment of those tweets. People may have positive or negative sentiments about the candidates, reflected by their tweets. We'll do sentiment analysis next.

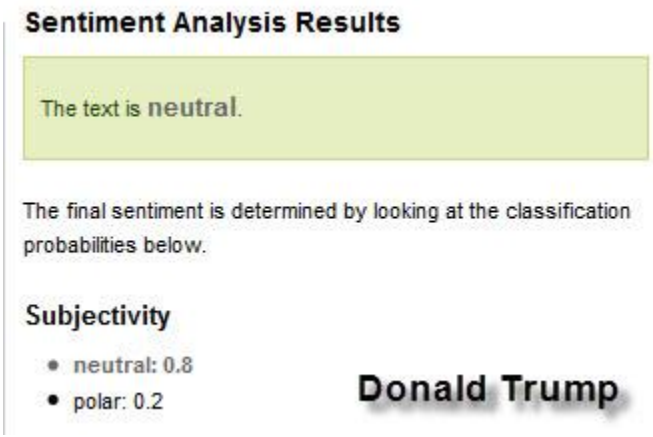
❖ **A time series figure with the number of tweets per day over time for both candidates.**

- The figure below shows the timeline for the number of tweets for each day for Trump and Clinton:

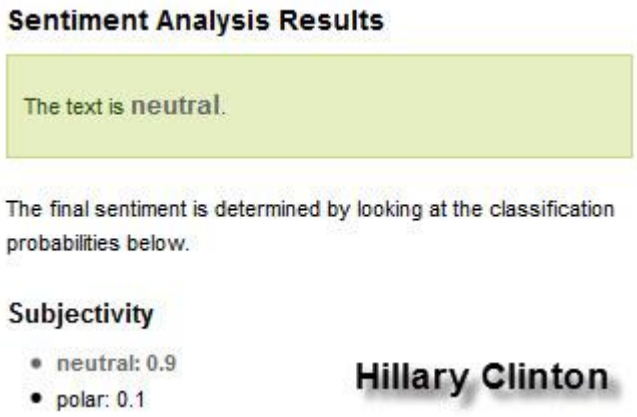


❖ **Perform one more interesting analysis of your choice.**

- For both Trump and Hillary tweets I have performed sentiment analysis using the implementation of NLTK algorithm at http://text-processing.com/demo/sentiment/
- Sentiment analysis calculates positivity and negativity of the sentences, based on the individual words.
- Using the above tool, following were the results of Donald Trump:



- We can observe that the tweets about Donald Trump are neutral with a score of 0.8 and polarity of 0.2.
- And following are the results for Hillary Clinton:



- We can observe that the tweets about Donald Trump are neutral with a score of 0.9 and polarity of 0.1.
- So sentiment for Hillary's tweets have more neutral score and lesser polarity in comparison to Trump. It is more towards positive side for Hillary. So it can be said that people are little more in favor of Hillary than Trump.

❖ **Based on the above analysis, who will be elected in 2016 presidential election? (Just Assuming that Donald and Hillary will be the final candidate from each party)**

- Since the sentiment score for both the candidates is very much similar, it is really hard to predict who will be elected as the next president. But since Hillary has sentiment score slightly more towards positive side, we can say that **Hillary Clinton** has marginally more chance of becoming president in the 2016 presidential election.

------- X --------