**Homework #1**

**CS 6675, Spring 2016**

**Task 1: Collecting Twitter Data**

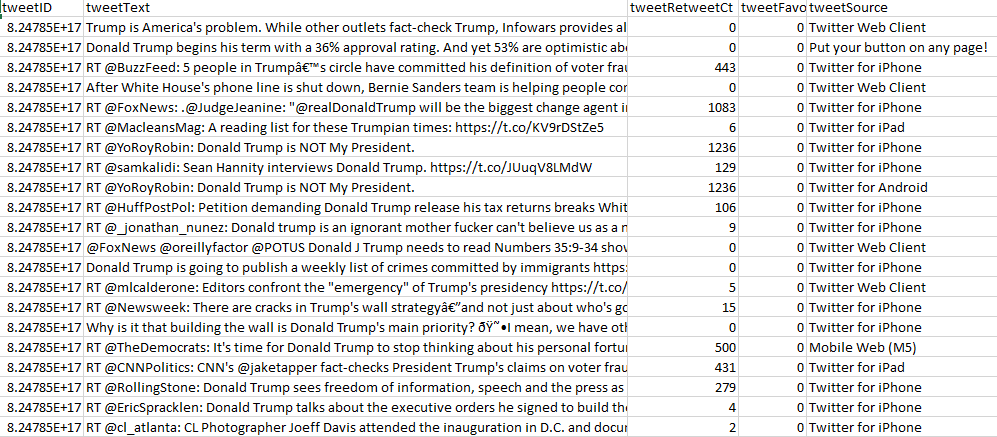
* In this task, you will collect some Twitter data by using Twitter APIs, especially, using Rest APIs.

By using a search API, collect at least 3,000 tweets containing a keyword “donald trump”, and store JSON format. Add following conditions when you use the search API:

* + A tweet's geolocation should be inside USA.
  + Tweets should be posted after election date (November 8, 2016).
  + Tweets should be written in English.

Include a few sample tweets in your report. Also, write the detailed steps that you followed, in order to collect your Twitter data. Report the size of your dataset

**Answer:** For scraping the tweets I have used python script with the use of ‘tweepy’ python library. I saved the data fetched in the form of JSON into CSV file using ‘Pandas’. Below is a screenshot of some sample tweets scrapped based on provided criteria:



Attaching python code used for scrapping:



* Tweepy provides the feature to ‘Search’ and ‘Stream’. I have used ‘Search’.
* As per problem statement to fetch data from USA, I have set the geo\_search property of API as “USA” and granularity as country.
* To get English tweets, “lang” property has been set to “en”.
* Now for searching the tweets for Donald Trump, I have used the keyword “donald trump” and set the number of items and count as 10000.

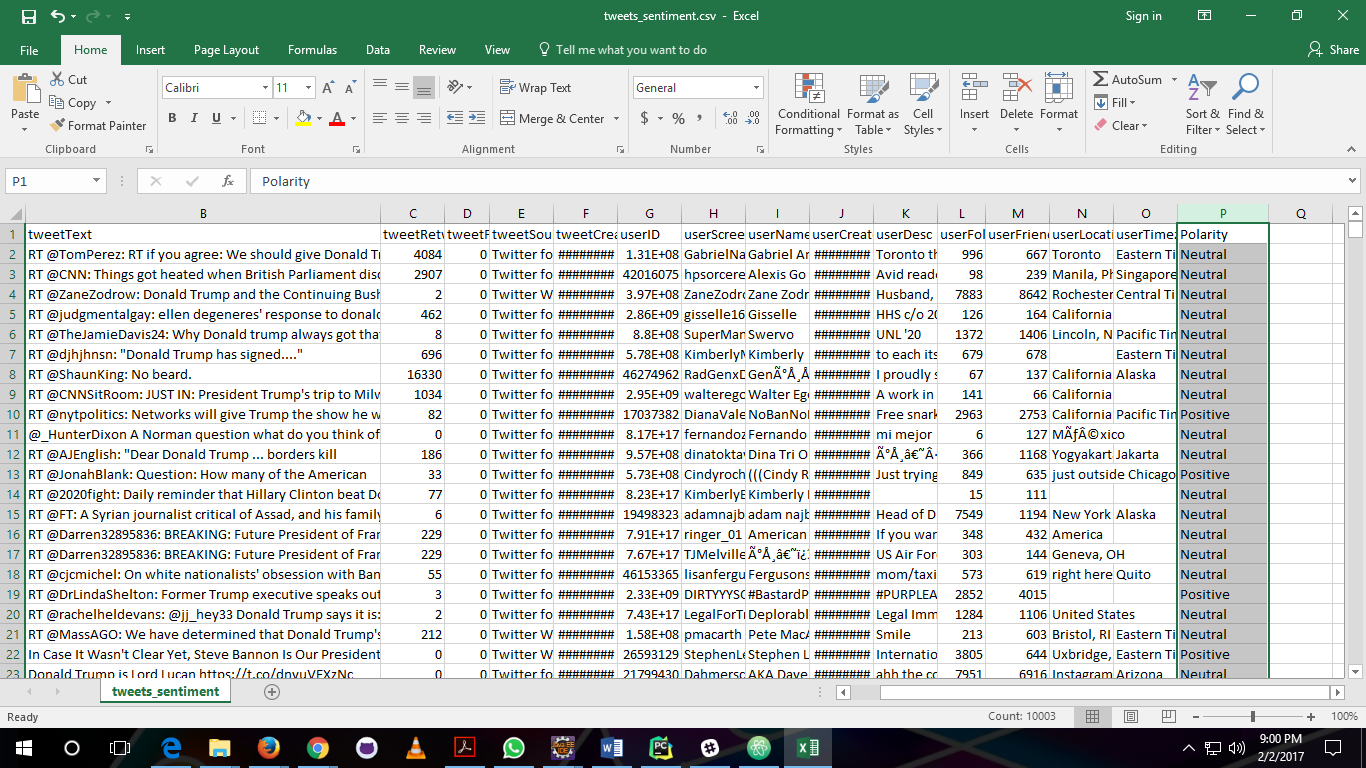
**Task 2: Preprocessing the data**

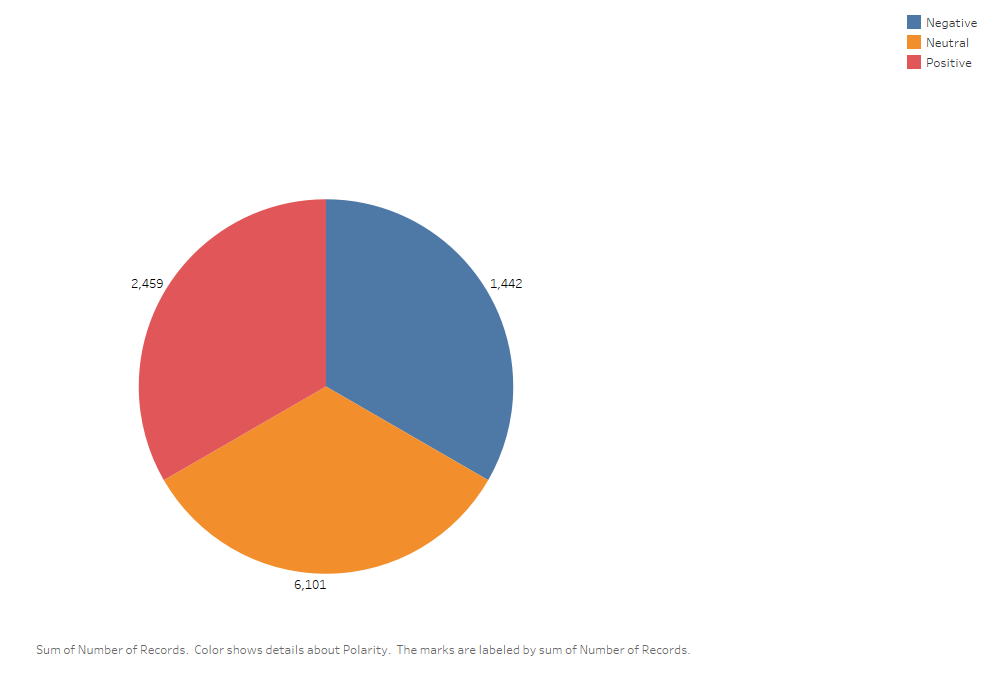
* Extract following properties from each tweet: created\_at, tweet\_id, text, user\_id, geo, coordinates, user\_id, user\_name, user\_location, place, country, friends\_count, followers\_count and language.
* Determine sentiment information of each tweet: sentiment analysis is the process of identifying opinions expressed in a text. It determines whether the attitude towards a particular topic (positive, negative, or neutral)

**Answer:** I extracted all the data and saved in the form of CSV during scraping of data using the python code. I was unable to save place and country as these fields were not available in the JSON obtained.

* Tweets collected had some inconsistences in the geo field.
* To make the location consistent, I used Open Refine. Clustered and merged all the location to remove the inconsistencies.
* For sentiment analysis, I used ‘TextBlob’ package in python. I read all the tweets from the CSV file, removed hyperlink, @ and # and passed them to textblob.sentiment.

Sentiment analysis calculated positive, negative, and neutral tweets, based on the individual words. Below I have attached the screenshot of tweets along with the polarity or sentiments:



****

**Task 3: Exploratory Analysis through *k*-means clustering**

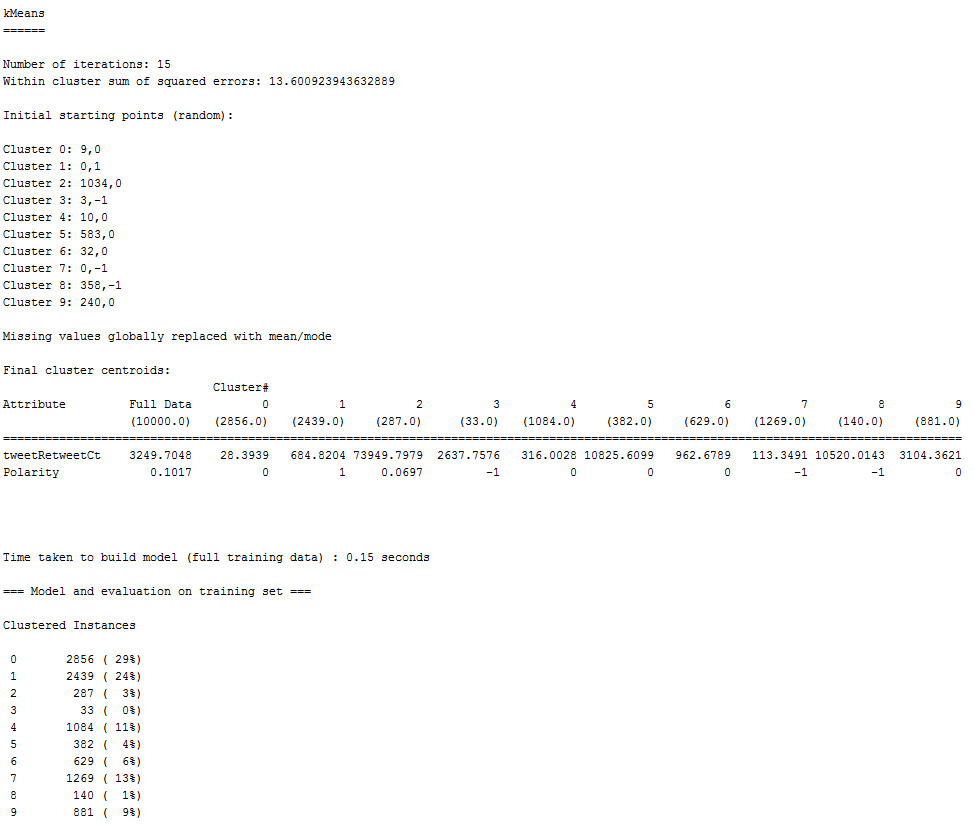
In order to gain clear understanding of the data that you preprocessed, you will conduct exploratory analysis. Based on knowledge regarding on *k*-means clustering algorithms learned in CS5665, you will conduct cluster analysis.

* Explain the reason you selected “k” number for your clustering and how each cluster is different from each other.
* In order to understand the support of Donald Trump in a different state, visualize the tweets distribution based on sentiment information on the map using tools like Tableau, Google maps API, basemap, D3, etc. *Report your findings including some figures like snapshots of the maps.*
* Perform one more interesting analysis of your choice.

**Answer:**

I performed k-means clustering algorithm on the dataset using weka.

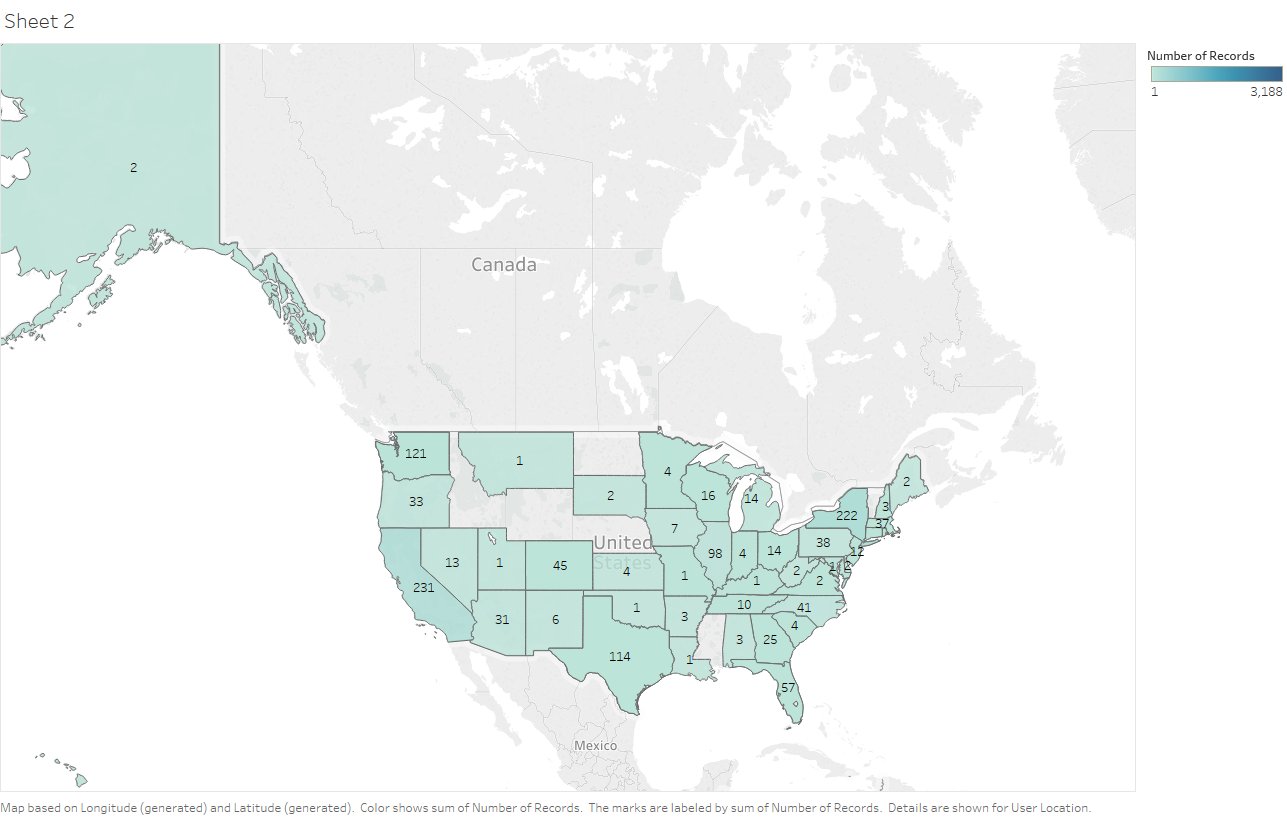
* I selected ‘tweetRetweetCt’ and ‘Polarity’ as 2 features for clustering.
* I selected k as 10, because if we choose very small number of clusters like 2, then the data is not separated well enough to make any business decisions, while if it is too high the data is separated out too much that each data cluster is too small that we cannot make interpretation. Therefore, for 10k instances I took k as 10.



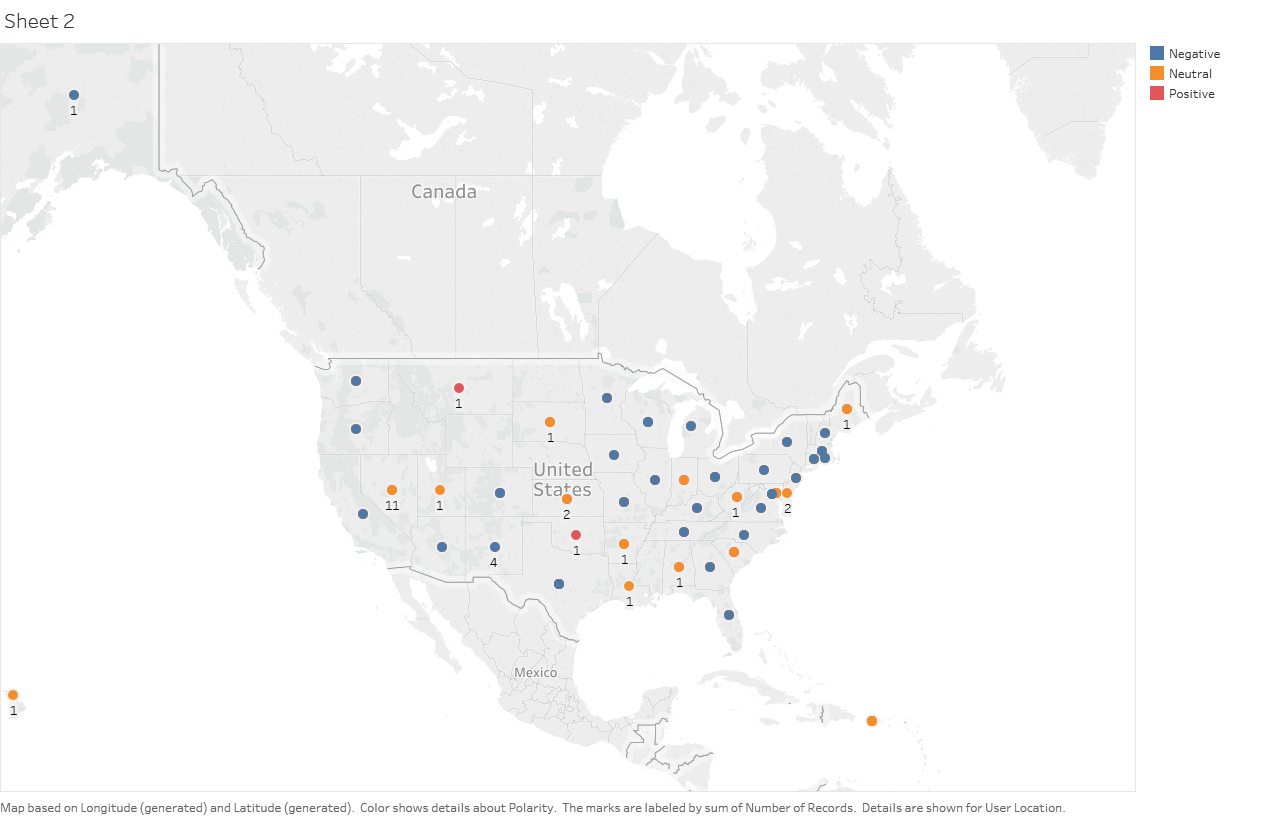
* From the image above, we will exclude cluster 2, 3, 5, 6, 8, 9 because very less instances are assigned to these clusters which is very less to make any meaningful interpretations.

To visualize the tweets, I have used Tableau tool. In which I imported the CSV saved in previous task.

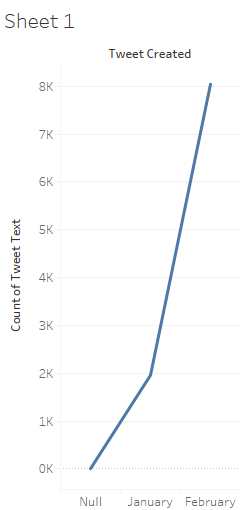
* The map below demonstrates the distribution of count of tweets containing the keywords for Donald Trump in US



* From the above map, we can infer that most number of tweets about Donald Trump has come from California (231), followed by New York (222), Washington (121).
* But only from this we cannot conclude that Trump is going to get more votes in those states. We don’t know the sentiment of those tweets in the map. People may have positive or negative sentiments. Therefore, I used “Sentiment Analysis” calculated in the previous task to find the emotions of the tweets as shown in the below map.
* To visualize the tweets with sentiments in each state



* The line chart below shows the month-wise distribution of number of tweets for Trump:



* The chart below shows the number of times each tweet is retweeted:

