Homework #1

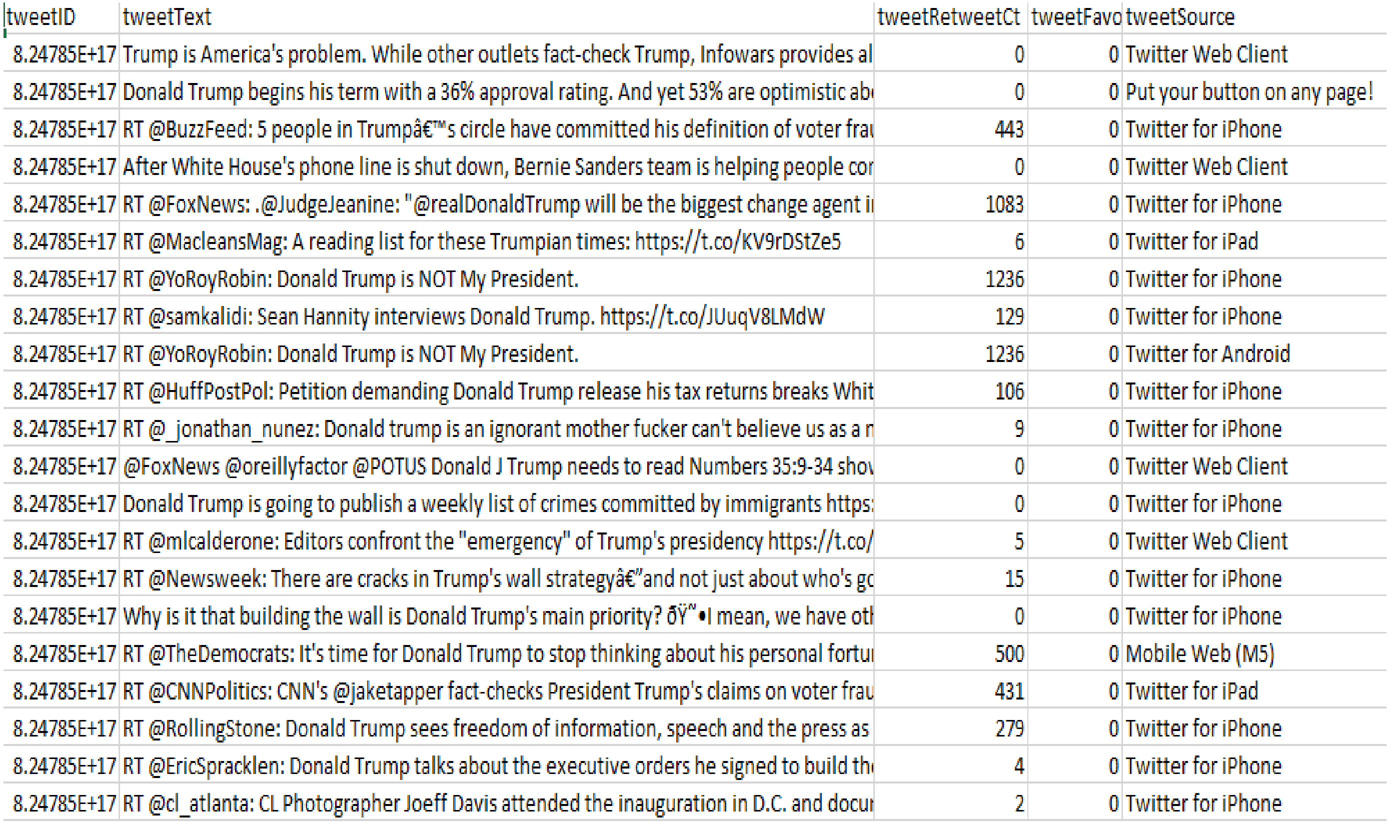
CS 6675, Spring 2016

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

  <u>Answer:</u> 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:



tweetID	tweetText	tweetRetweetCt	tweetFavo	tweetSource
8.24785E+17	Trump is America's problem. While other outlets fact-check Trump, Infowars provides al	0	0	Twitter Web Client
8.24785E+17	Donald Trump begins his term with a 36% approval rating. And yet 53% are optimistic ab	0	0	Put your button on any page!
8.24785E+17	RT @BuzzFeed: 5 people in Trumpâ€™s circle have committed his definition of voter frau	443	0	Twitter for iPhone
8.24785E+17	After White House's phone line is shut down, Bernie Sanders team is helping people cor	0	0	Twitter Web Client
8.24785E+17	RT @FoxNews: .@JudgeJeanine: "@realDonaldTrump will be the biggest change agent i	1083	0	Twitter for iPhone
8.24785E+17	RT @MacleansMag: A reading list for these Trumpian times: https://t.co/KV9rDStZe5	6	0	Twitter for iPad
8.24785E+17	RT @YoRoyRobin: Donald Trump is NOT My President.	1236	0	Twitter for iPhone
8.24785E+17	RT @samkalidi: Sean Hannity interviews Donald Trump. https://t.co/JUuqV8LMdW	129	0	Twitter for iPhone
8.24785E+17	RT @YoRoyRobin: Donald Trump is NOT My President.	1236	0	Twitter for Android
8.24785E+17	RT @HuffPostPol: Petition demanding Donald Trump release his tax returns breaks Whit	106	0	Twitter for iPhone
8.24785E+17	RT @_jonathan_nunez: Donald trump is an ignorant mother fucker can't believe us as a r	9	0	Twitter for iPhone
8.24785E+17	@FoxNews @oreillyfactor @POTUS Donald J Trump needs to read Numbers 35:9-34 shov	0	0	Twitter Web Client
8.24785E+17	Donald Trump is going to publish a weekly list of crimes committed by immigrants https:	0	0	Twitter for iPhone
8.24785E+17	RT @mlcalderone: Editors confront the "emergency" of Trump's presidency https://t.co/	5	0	Twitter Web Client
8.24785E+17	RT @Newsweek: There are cracks in Trump's wall strategyâ€”and not just about who's gc	15	0	Twitter for iPhone
8.24785E+17	Why is it that building the wall is Donald Trump's main priority? ðŸ˜•I mean, we have otl	0	0	Twitter for iPhone
8.24785E+17	RT @TheDemocrats: It's time for Donald Trump to stop thinking about his personal fortu	500	0	Mobile Web (M5)
8.24785E+17	RT @CNNPolitics: CNN's @jaketapper fact-checks President Trump's claims on voter frau	431	0	Twitter for iPad
8.24785E+17	RT @RollingStone: Donald Trump sees freedom of information, speech and the press as	279	0	Twitter for iPhone
8.24785E+17	RT @EricSpracklen: Donald Trump talks about the executive orders he signed to build th	4	0	Twitter for iPhone
8.24785E+17	RT @cl_atlanta: CL Photographer Joeff Davis attended the inauguration in D.C. and docur	2	0	Twitter for iPhone

  Attaching python code used for scrapping:

twitter.py

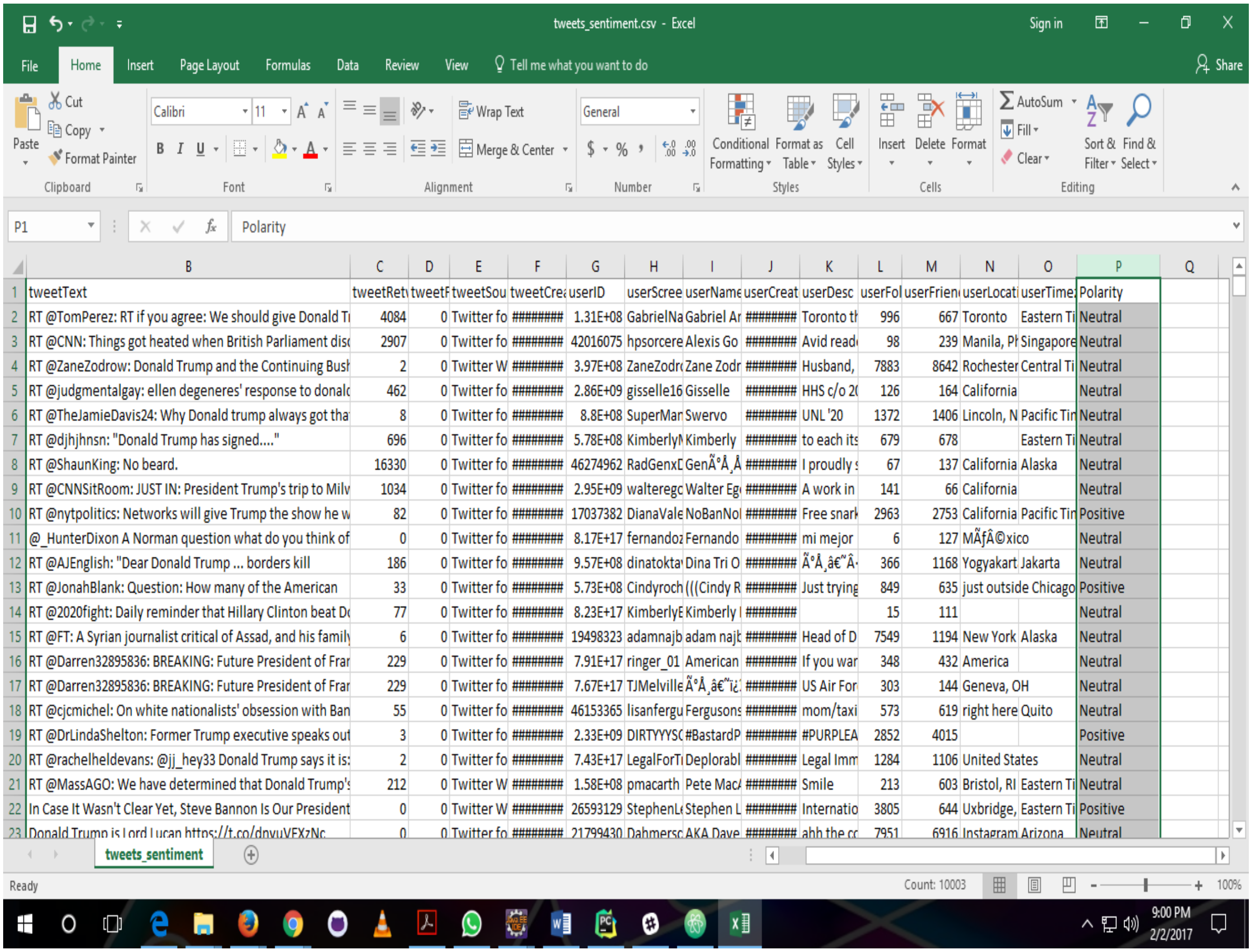
- Tweepy provides the feature to ‘Search’ and ‘Stream’. I have used ‘Search’.
- As per problem statement to fetch data from USA, I have set the geo_search property of API as “USA” and granularity as country.
- To get English tweets, “lang” property has been set to “en”.
- Now for searching the tweets 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)

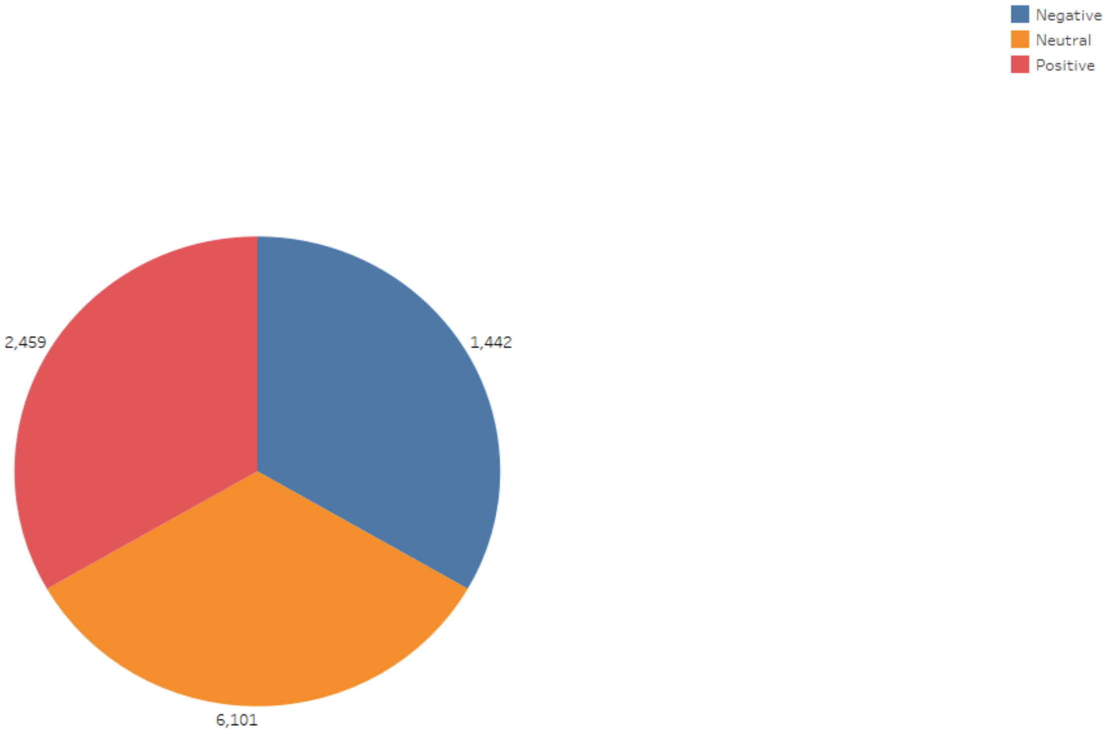
  Answer: I extracted all the data and saved in the form of CSV during scraping of data using the python code. I was unable to save place and country as these fields were not available in the JSON obtained.
- Tweets collected had some inconsistences in the geo field.
- To make the location consistent, 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:



P1 fx Polarity

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	tweetText	tweetRet	tweetF	tweetSou	tweetCre	userID	userScree	userName	userCreat	userDesc	userFol	userFrien	userLocati	userTime	Polarity
2	RT @TomPerez: RT if you agree: We should give Donald Tı	4084	0	Twitter fo	########	1.31E+08	GabrielNa	Gabriel Ar	########	Toronto tl	996	667	Toronto	Eastern Ti	Neutral
3	RT @CNN: Things got heated when British Parliament disc	2907	0	Twitter fo	########	42016075	hpsorcere	Alexis Go	########	Avid read	98	239	Manila, Pł	Singapore	Neutral
4	RT @ZaneZodrow: Donald Trump and the Continuing Busł	2	0	Twitter W	########	3.97E+08	ZaneZodr	Zane Zodr	########	Husband,	7883	8642	Rochester	Central Ti	Neutral
5	RT @judgmentalgay: ellen degeneres' response to donalc	462	0	Twitter fo	########	2.86E+09	giselle16	Gisselle	########	HHS c/o 2(	126	164	California		Neutral
6	RT @TheJamieDavis24: Why Donald trump always got tha	8	0	Twitter fo	########	8.8E+08	SuperMar	Swervo	########	UNL '20	1372	1406	Lincoln, N	Pacific Tin	Neutral
7	RT @djhjhnsn: "Donald Trump has signed...."	696	0	Twitter fo	########	5.78E+08	KimberlyN	Kimberly	########	to each its	679	678		Eastern Ti	Neutral
8	RT @ShaunKing: No beard.	16330	0	Twitter fo	########	46274962	RadGenxD	GenÃ°Å¸Å	########	I proudly s	67	137	California	Alaska	Neutral
9	RT @CNNSitRoom: JUST IN: President Trump's trip to Milv	1034	0	Twitter fo	########	2.95E+09	walteregc	Walter Eg	########	A work in	141	66	California		Neutral
10	RT @nytpolitics: Networks will give Trump the show he w	82	0	Twitter fo	########	17037382	DianaVale	NoBanNol	########	Free snark	2963	2753	California	Pacific Tin	Positive
11	@_HunterDixon A Norman question what do you think of	0	0	Twitter fo	########	8.17E+17	fernandoz	Fernando	########	mi mejor	6	127	MÃƒÂ©xico		Neutral
12	RT @AJEnglish: "Dear Donald Trump ... borders kill	186	0	Twitter fo	########	9.57E+08	dinatokta	Dina Tri O	########	Ã°Å¸â€˜Â·	366	1168	Yogyakart	Jakarta	Neutral
13	RT @JonahBlank: Question: How many of the American	33	0	Twitter fo	########	5.73E+08	Cindyroch	(((Cindy R	########	Just trying	849	635	just outside	Chicago	Positive
14	RT @2020fight: Daily reminder that Hillary Clinton beat D(	77	0	Twitter fo	########	8.23E+17	KimberlyE	Kimberly l	########		15	111			Neutral
15	RT @FT: A Syrian journalist critical of Assad, and his family	6	0	Twitter fo	########	19498323	adamnajb	adam najb	########	Head of D	7549	1194	New York	Alaska	Neutral
16	RT @Darren32895836: BREAKING: Future President of Frar	229	0	Twitter fo	########	7.91E+17	ringer_01	American	########	If you war	348	432	America		Neutral
17	RT @Darren32895836: BREAKING: Future President of Frar	229	0	Twitter fo	########	7.67E+17	TJMelville	Ã°Å¸â€˜ï¿½	########	US Air For	303	144	Geneva, OH		Neutral
18	RT @cjcmichel: On white nationalists' obsession with Ban	55	0	Twitter fo	########	46153365	lisanfergu	Fergusons	########	mom/taxi	573	619	right here	Quito	Neutral
19	RT @DrLindaShelton: Former Trump executive speaks out	3	0	Twitter fo	########	2.33E+09	DIRTYYYS(	#BastardP	########	#PURPLEA	2852	4015			Positive
20	RT @rachelheldevans: @jj_hey33 Donald Trump says it is:	2	0	Twitter fo	########	7.43E+17	LegalForT	Deplorabl	########	Legal Imm	1284	1106	United States		Neutral
21	RT @MassAGO: We have determined that Donald Trump's	212	0	Twitter W	########	1.58E+08	pmacarth	Pete MacA	########	Smile	213	603	Bristol, RI	Eastern Ti	Neutral
22	In Case It Wasn't Clear Yet, Steve Bannon Is Our President	0	0	Twitter W	########	26593129	StephenL	Stephen L	########	Internatio	3805	644	Uxbridge,	Eastern Ti	Positive
23	Donald Trump is Lord Lucan https://t.co/dnvuVFXzNc	0	0	Twitter fo	########	21799430	Dahmersc	AKA Dave	########	ahh the cc	7951	6916	Instagram	Arizona	Neutral



Sum of Number of Records. Color shows details about Polarity. The marks are labeled by sum of Number of Records.

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

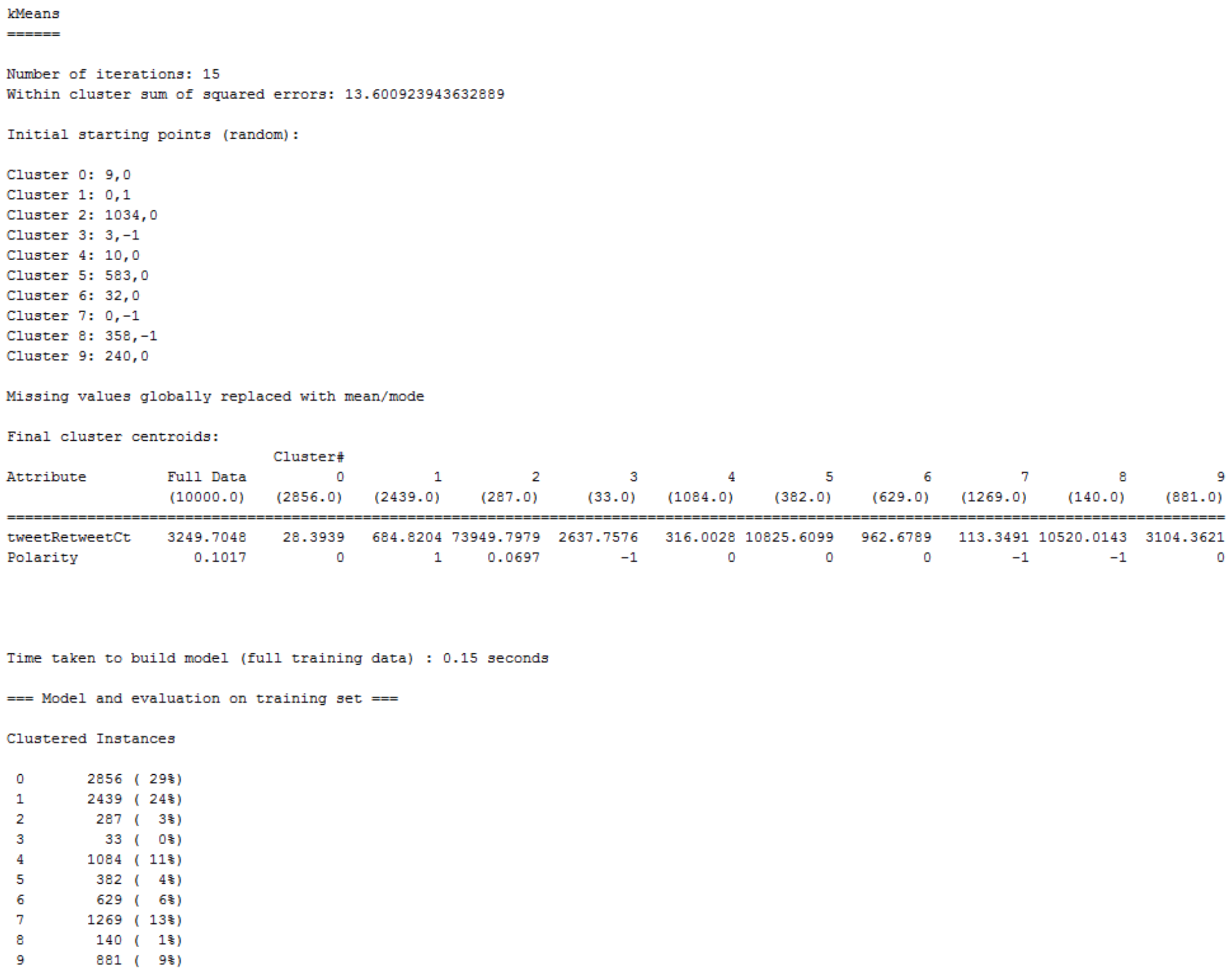
In order to gain clear understanding of the data that you preprocessed, you will conduct exploratory analysis. Based on knowledge regarding on *k*-means clustering algorithms learned in CS5665, you will conduct cluster analysis.

- 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.

<u>Answer:</u>

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.

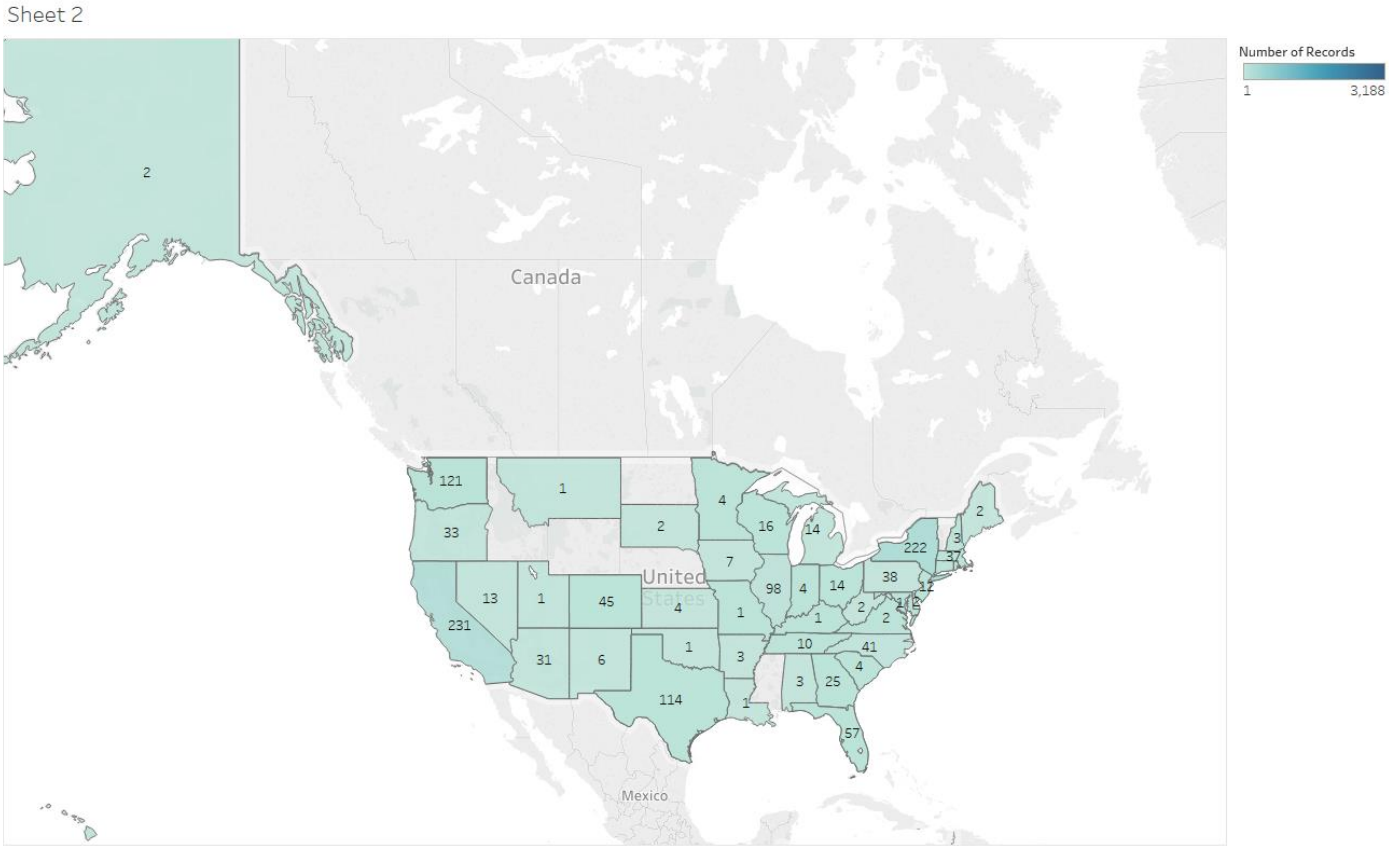


```
kMeans
======

Number of iterations: 15
Within cluster sum of squared errors: 13.600923943632889

Initial starting points (random):

Cluster 0: 9,0
Cluster 1: 0,1
Cluster 2: 1034,0
Cluster 3: 3,-1
Cluster 4: 10,0
Cluster 5: 583,0
Cluster 6: 32,0
Cluster 7: 0,-1
Cluster 8: 358,-1
Cluster 9: 240,0

Missing values globally replaced with mean/mode

Final cluster centroids:
                              Cluster#
Attribute         Full Data          0          1          2          3          4          5          6          7          8          9
                  (10000.0)   (2856.0)   (2439.0)    (287.0)     (33.0)   (1084.0)    (382.0)    (629.0)   (1269.0)    (140.0)    (881.0)
=========================================================================================================================================
tweetRetweetCt    3249.7048    28.3939   684.8204 73949.7979  2637.7576   316.0028 10825.6099   962.6789   113.3491 10520.0143  3104.3621
Polarity             0.1017          0          1     0.0697         -1          0          0          0         -1         -1          0




Time taken to build model (full training data) : 0.15 seconds

=== Model and evaluation on training set ===

Clustered Instances

 0      2856 ( 29%)
 1      2439 ( 24%)
 2       287 (  3%)
 3        33 (  0%)
 4      1084 ( 11%)
 5       382 (  4%)
 6       629 (  6%)
 7      1269 ( 13%)
 8       140 (  1%)
 9       881 (  9%)
```

- 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

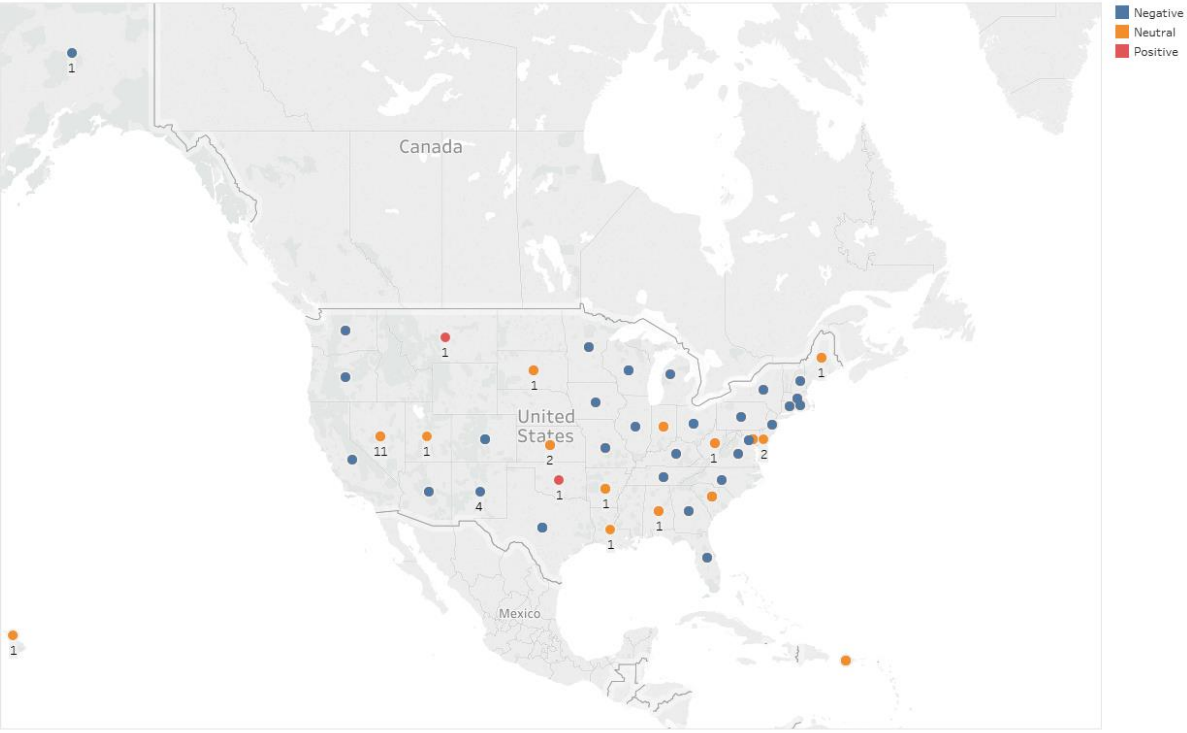


Map based on Longitude (generated) and Latitude (generated). Color shows sum of Number of Records. The marks are labeled by sum of Number of Records. Details are shown for User Location.

- 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.

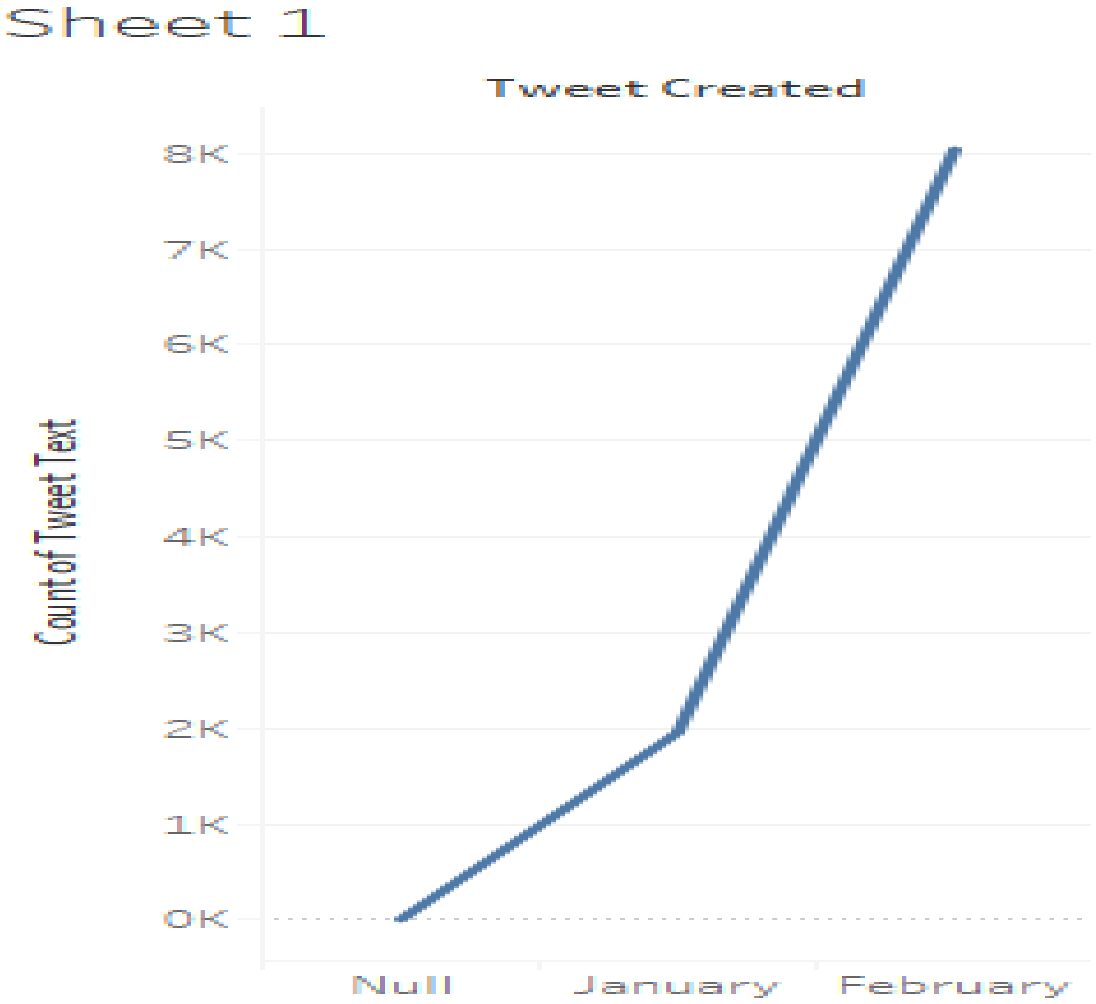
- To visualize the tweets with sentiments in each state

Sheet 2



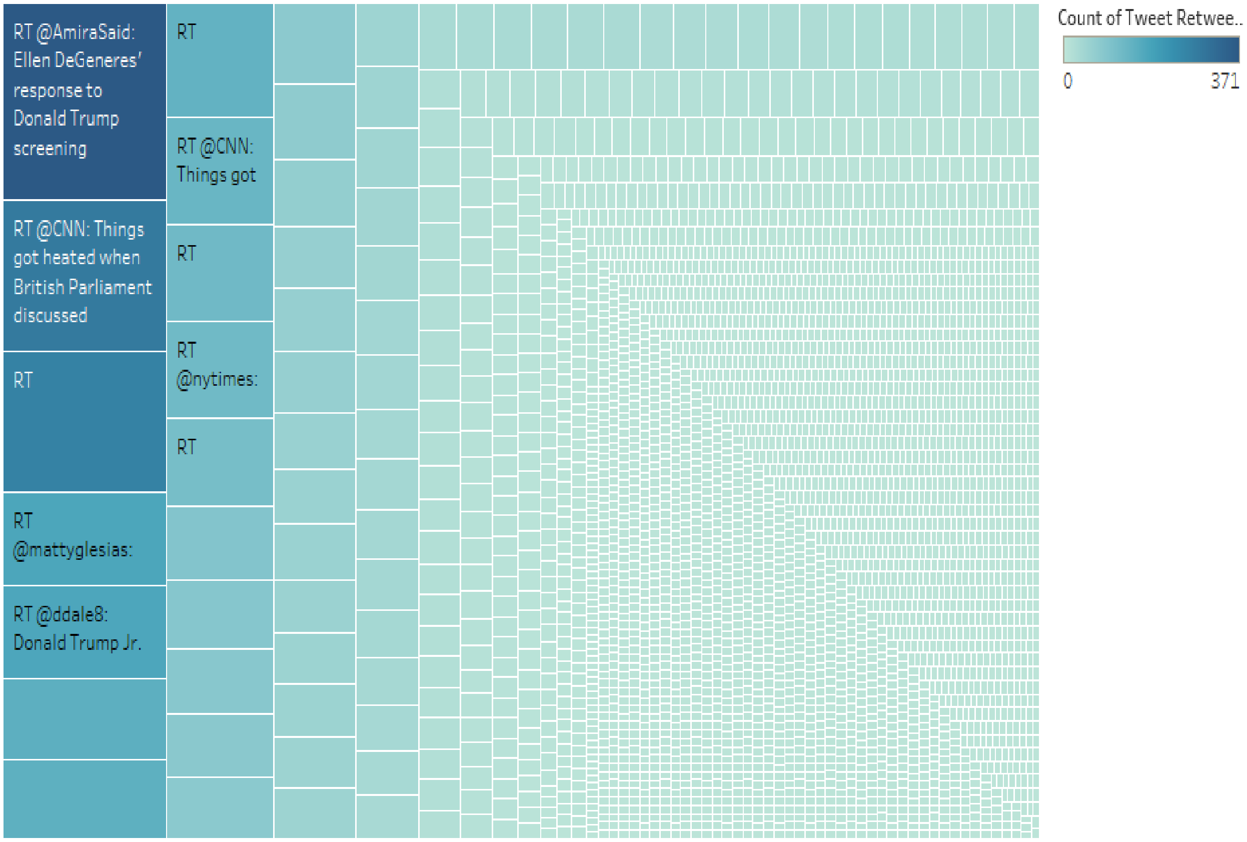
Map based on Longitude (generated) and Latitude (generated). Color shows details about Polarity. The marks are labeled by sum of Number of Records. Details are shown for User Location.

- 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:

Sheet 3



Tweet Text. Color shows count of Tweet Retweet Ct. Size shows count of Tweet Retweet Ct. The marks are labeled by Tweet Text.