“ANALYZING DONALD J. TRUMP’S TWEETS USING TEXT MINING AND SENTIMENT ANALYSIS”

PREPARED BY: DR. AKRAM ALMOHALWAS, YAN KANG AND CITINA LIANG

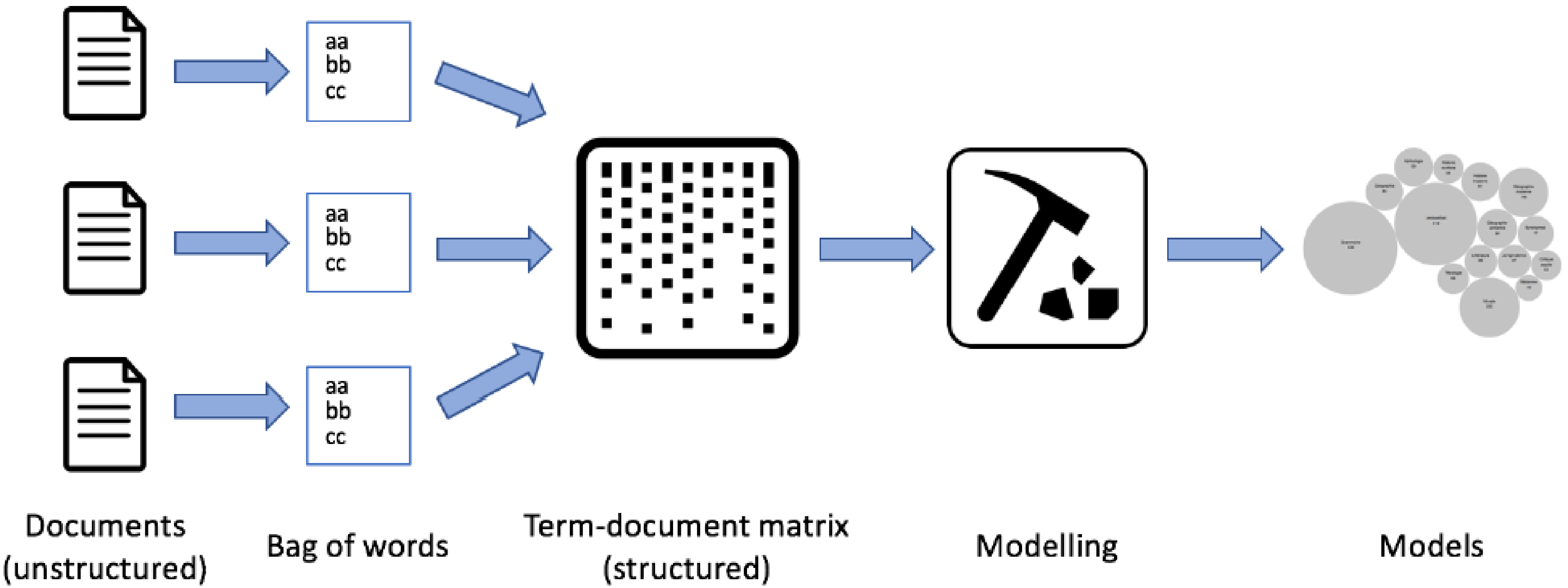
DEPARTMENT OF STATISTICS, UCLA, LOS ANGELES, CA, USA

ICOSDA PRESENTATION, 2019. SATURDAY OCT. 12TH, 10:15 AM – 10:35 AM

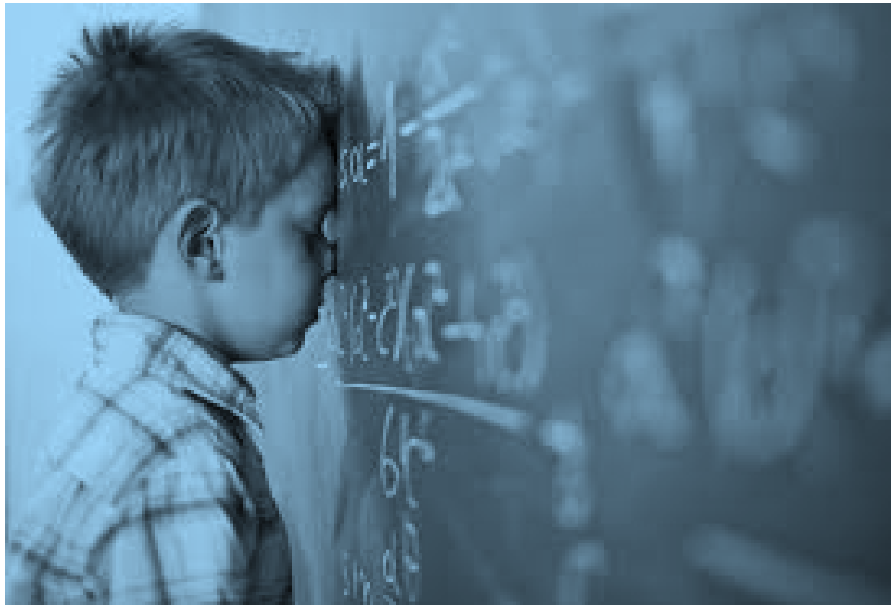
WHAT IS TEXT MINING ALL ABOUT?

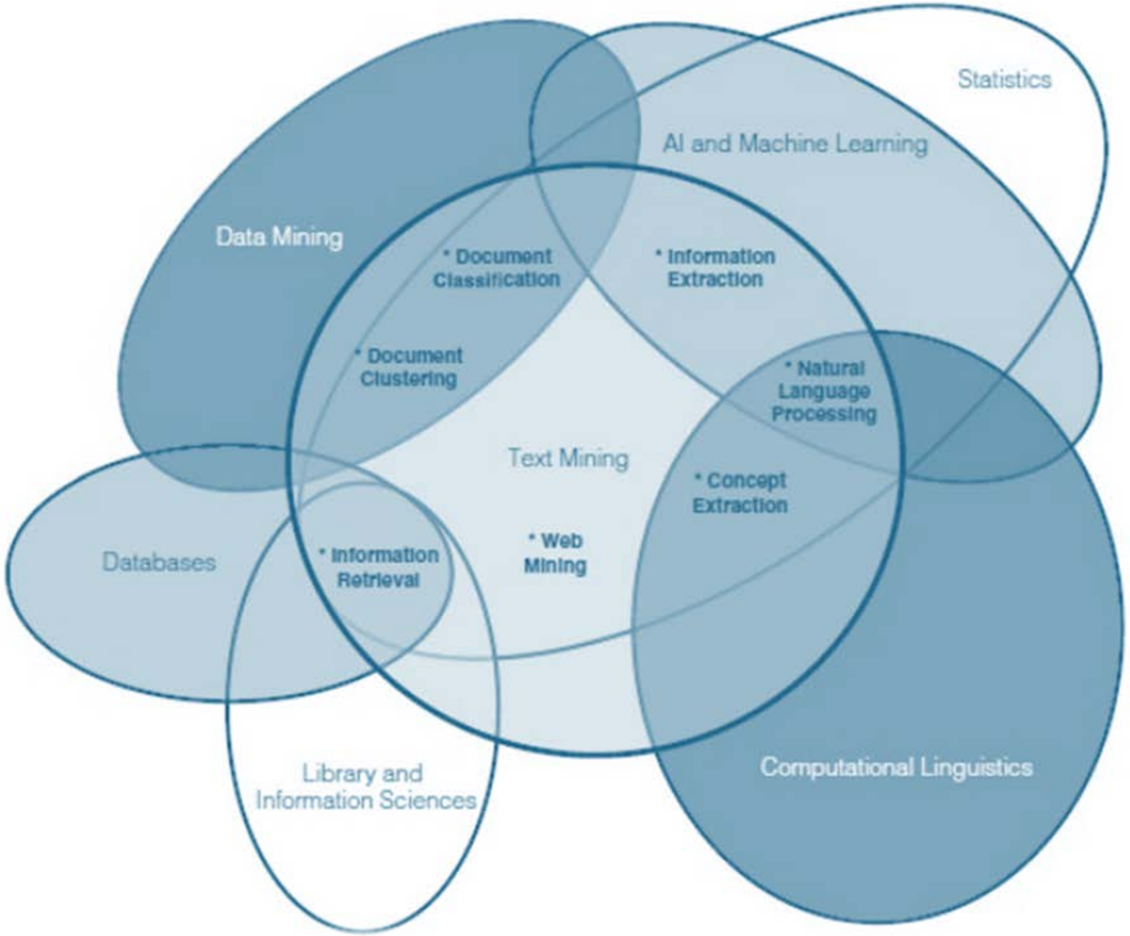
- **Text mining**, also referred to as **text** data **mining**, roughly equivalent to **text analytics**, is the process of deriving high-quality information from **text**. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning.
- Text Mining applies analytic tools to learn from collections of text data, like social media, books, newspapers, emails, etc.
- The goal can be considered to be similar to humans learning by reading such material. However, using automated algorithms we can learn from massive amounts of text, very much more than a human can. The material could consist of millions of newspaper articles to perhaps summarize the main themes and to identify those that are of most interest to particular people.

TYPICAL PROCESS OF TEXT MINING



TEXT MINING IS VERY INVOLVED





GENERAL STEPS OF TEXT MINING; RECOMMENDED R PACKAGES

	R packages	
Operation	example	alternatives
Data preparation		
importing text	*readtext*	*jsonlite, XML, antiword, readxl, pdftools*
string operations	*stringi*	*stringr*
preprocessing	*quanteda*	*stringi, tokenizers, snowballC, tm, etc.*
document-term matrix (DTM)	*quanteda*	*tm, tidytext, Matrix*
filtering and weighting	*quanteda*	*tm, tidytext, Matrix*
Analysis		
dictionary	*quanteda*	*tm, tidytext, koRpus, corpustools*
supervised machine learning	*quanteda*	*RTextTools, kerasR, austin*
unsupervised machine learning	*topicmodels*	*quanteda, stm, austin, text2vec*
text statistics	*quanteda*	*koRpus, corpustools, textreuse*
Advanced topics		
advanced NLP	*spacyr*	*coreNLP, cleanNLP, koRpus*
word positions and syntax	*corpustools*	*quanteda, tidytext, koRpus*

MORE USEFUL R PACKAGES: SO MANY COOKIES

library(tm) # Framework for text mining.
library(qdap) # Quantitative discourse analysis of transcripts.
library(qdapDictionaries)
library(dplyr) # Data wrangling, pipe operator %>%().
library(RColorBrewer) # Generate palette of colors for plots.
library(ggplot2) # Plot word frequencies.
library(scales) # Include commas in numbers.
library(Rgraphviz) # Correlation plots.
library(wordcloud) # [Fellows, 2012] visualizes results.
library(fpc) # [Christian Hennig, 2005] flexible procedures for clustering.
library(igraph) # [Gabor Csardi , 2012] a library and R package for network analysis.
library(knitr) # (Xie, 2014) is a package which provides convenient methods for reproducible research purposes. With knitr one can create HTML, PDF or WORD documents from R analysis scripts.
library(rvest) # (Wickham, 2016) web crawling and scraping
library(topicmodels) # (Gr¨un and Hornik, 2011) For unsupervised clustering
library(LiblineaR) # (Helleputte, 2017) For supervised classification.

WHY TWITTER DATA?

Twitter is:

- An online social networking service that enables users to send and read short 280-character (used to be 140 before November 2017) messages called \tweets". This limit was doubled to 280 for all languages except Chinese, Japanese, and Korean. (Wikipedia)
- Founded: March 21, 2006, San Francisco, CA.
- Headquarters: San Francisco, CA.
- Founders: Jack Dorsey, Noah Glass, Biz Stone, Evan Williams
- Over 321 million active users (as February 2019)
- Creating over 500 million tweets per day

GETTING THE TWEETS: 'HTTP://WWW.TRUMPTWITTERARCHIVE.COM/DATA/REALDONALDTRUMP/%S.JSON

We have 41419 tweets and 8 variables 8 "Four un-useful variables were omitted" (As of Oct. 7th, 2019 @ 1:00 PM) including Source of the tweet: Tweets contents (text), Favorite Count (favorite_count), Post time (created_at),Retweet count (retweet_count)

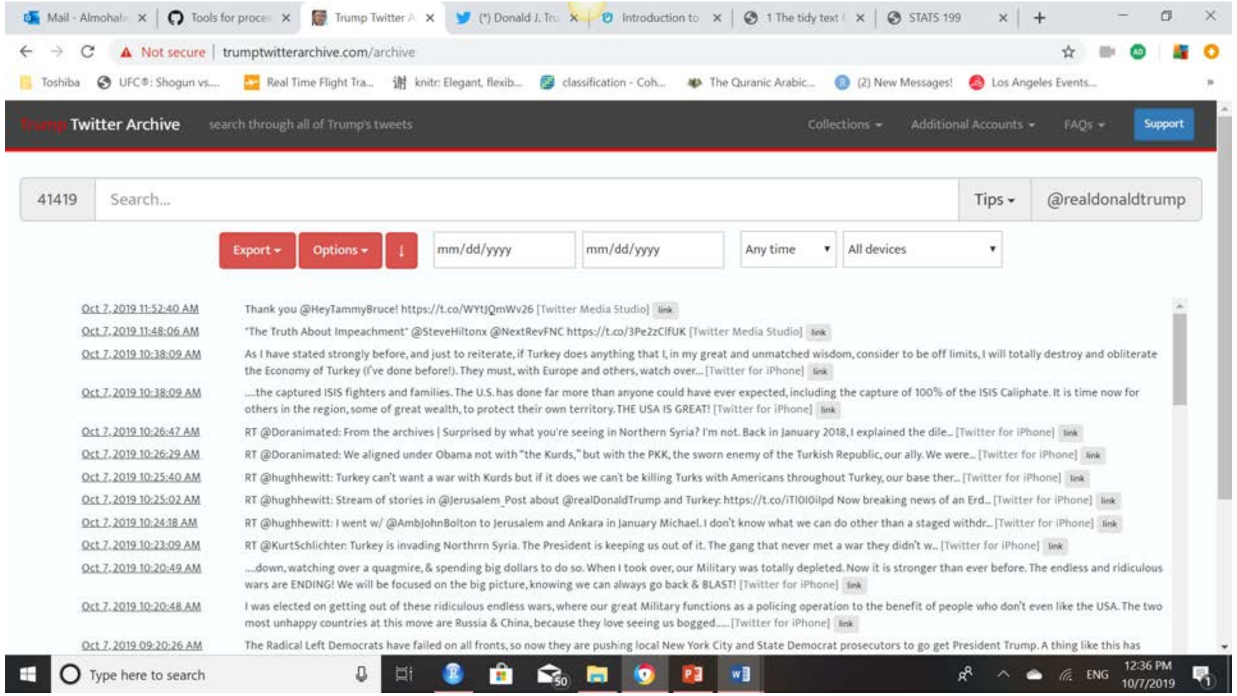
Considering Tweets after June 16, 2015, we have 16657 tweets.

text	favorite_count	created_at	retweet_count
I would like to wish everyone A HAPPY AND HEALTHY NEW YEAR. WE MUST ALL WORK TOGETHER TO, FINALLY, MAKE AMERICA SAFE AGAIN AND GREAT AGAIN!	16495	2015-12-31 23:21:49	6776
Do you believe that The State Department, on NEW YEAR'S EVE, just released more of Hillary's e-mails. They just want it all to end. BAD!	6824	2015-12-31 23:11:35	2755
THANK YOU ILLINOIS! Let's not forget to get family & friends- out to VOTE IN 2016! https://t.co/lg5kMbNLYK https://t.co/dtMAslq4cf	6047	2015-12-31 18:51:12	2468
HAPPY BIRTHDAY to my son, @DonaldJTrumpJr! Very proud of you! #TBT https://t.co/ULerCEOCGX https://t.co/nbxPVdarJM	8416	2015-12-31 16:52:38	2080
I would feel sorry for @JebBush and how badly he is doing with his campaign other than for the fact he took millions of $'s of hit ads on me	5780	2015-12-31 15:07:18	1875
#MakeAmericaGreatAgain #Trump2016 https://t.co/IEIXos0wh9	5729	2015-12-31 12:51:35	2285

GETTING STARTED: THE CORPUS, LOADING TEXT FILES

- A corpus. is a collection of texts, usually stored electronically, and from which we perform our analysis.
- A corpus might be a collection of news articles from Reuters or the published works of Shakespeare or in our example today the Holy Bible.
- Within each corpus we will have separate documents, which might be articles, stories, or book volumes. Each document is treated as a separate entity or record.
- Documents which we wish to analyze come in many different formats. Quite a few formats are supported by the tm library (Feinerer and Hornik, 2015), and *Quanteda* library, the packages we will illustrate text mining with in this module. The supported formats include text, PDF, Microsoft Word, and XML.

A snap shot of the President's tweets Text File:



GETTING TEXT READY TO BE ANALYZED: TRANSFORMING AND STEMMING TEXT

- Use regular expressions to remove at-tags and urls from the remaining documents (if any)
- Convert all text letters to lowercase.
- Remove white spaces.
- Remove punctuation signs.
- Build a corpus, and specify the source to be character vectors.
- Remove stopwords.
- Replace contractions (ex. "it's" becomes "it is").
- Replace Arabic numbers in words (ex. "2" becomes "two").
- Remove anything other than English letters or space.
- Keep a copy to use later as a dictionary for stem completion.
- *Procedures of stemming words by using stemDocument function, and use stemCompletion function to complete the words. stemming: the process of reducing inflected (or sometimes derived) words to their word stem, base or root form*
- Remove generic and custom stopwords, such as 'dont','didnt','arent','cant','one','two'.
- Removing Sparse Terms. We decide the threshold.
- Use cleaned corpus to create Tokens: **Tokenization** is the act of breaking up a sequence of strings into pieces such as words, keywords, phrases, symbols and other elements called tokens. Tokens can be individual words, phrases or even whole sentences. In the process of **tokenization**, some characters like punctuation marks are discarded.

FROM CORPUS TO DOCUMENT-TERM MATRIX (DTM)

- Creating Document-term matrix: The document term matrix (DTM) is one of the most common formats for representing a text corpus (i.e. a collection of texts) in a bag-of-words format. A DTM is a matrix in which rows are documents, columns are terms, and cells indicate how often each term occurred in each document.
- The advantage of this representation is that it allows the data to be analyzed with vector and matrix algebra, effectively moving from text to numbers.
- Furthermore, with the use of special matrix formats for sparse matrices, text data in a DTM format is very memory efficient and can be analyzed with highly optimized operations.
- Using tidytext package (Silge & Robinson, 2016) or quanteda package.

DTM of President Trump Corpus

<<TermDocumentMatrix (terms: 13152, documents: 16657)>>
Non-/sparse entries: 197106/218875758
Sparsity : 100%
Maximal term length: 38
Weighting : term frequency (tf)

	10841	10944	12209	12808	14199	14598	15161	15226	15770	16644
america	0	0	0	1	0	0	0	0	0	0
country	0	0	0	0	0	0	0	0	0	0
democrat	0	0	0	0	0	0	0	0	0	0
great	0	0	1	0	0	0	0	0	0	1
make	0	0	0	0	0	0	0	0	0	1
people	0	0	0	0	0	1	0	0	0	0
president	0	0	0	0	0	0	0	0	1	1
realdonaldtrump	0	0	0	0	0	0	0	0	0	0
state	0	0	0	0	0	0	1	0	0	0
trump	0	1	1	0	1	0	0	0	0	1

SOME INITIAL RESULTS: LOTS TO DO AND LEARN

The Following Topic have been investigated:

I. Single-word frequency (Word Cloud)

II. Bigram frequency

III. Unsupervised Learning Algorithms (Clustering)

IV. Building Latent Dirichlet Allocation Models

V. Supervised Learning Algorithms (Classifications)

VI. VII. Sentiment Analysis

VII. Social Network Analysis

a. Follower's Location Analysis

b. Tweets: Most Favorites and Most Retweeted

I. SINGLE WORD ANALYSIS:

Below showing the most frequency Single words appear at least 500 times in Trump's tweets

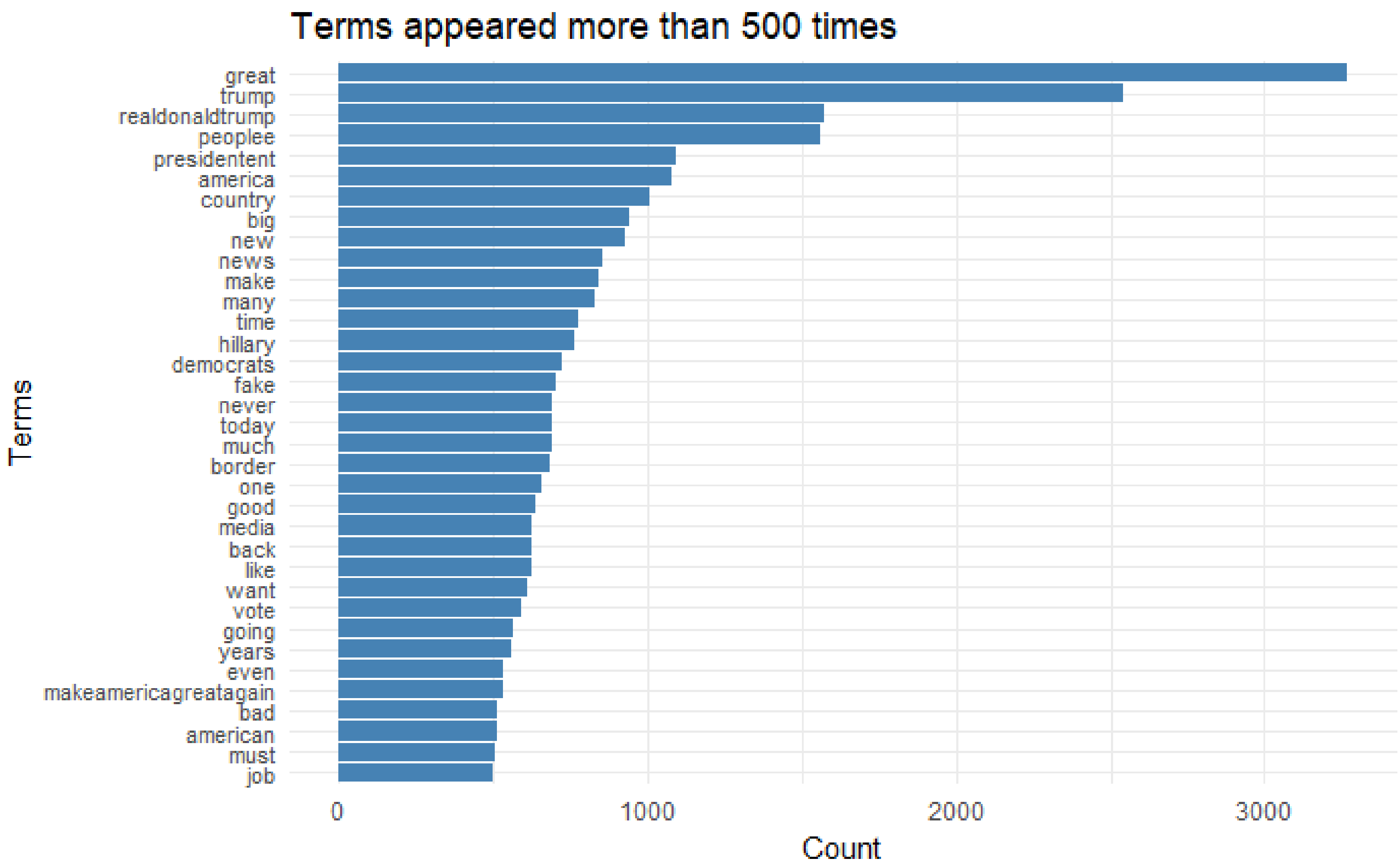
```
[1] "america"            "great"                   "like"
 [4] "make"               "must"                    "new"
 [7] "bad"                "want"                    "years"
[10] "vote"               "makeamericagreatagain" "trump"
[13] "news"               "time"                    "going"
[16] "big"                "peoplee"                 "many"
[19] "realdonaldtrump"    "even"                    "country"
[22] "never"              "much"                    "good"
[25] "one"                "presidentent"            "today"
[28] "hillary"            "democrats"               "border"
[31] "job"                "media"                   "back"
[34] "american"           "fake"
```

A SINGLE WORD STATISTICAL SUMMARIES:

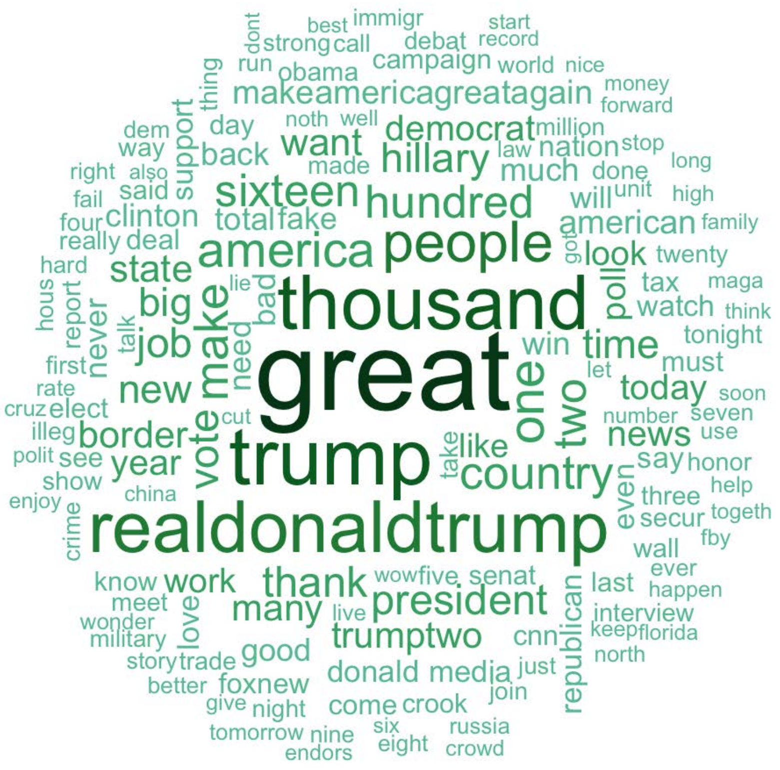
- Top Words Frequencies

Term	Frequency	Probability
great	2681	0.0827265
trump	1791	0.0552641
thousand	1689	0.0521168
realdonaldtrump	1585	0.0489077
people	1172	0.0361639
one	1010	0.0311651

A SINGLE WORD FREQUENCY OF THE TOP 500 FREQUENT WORDS

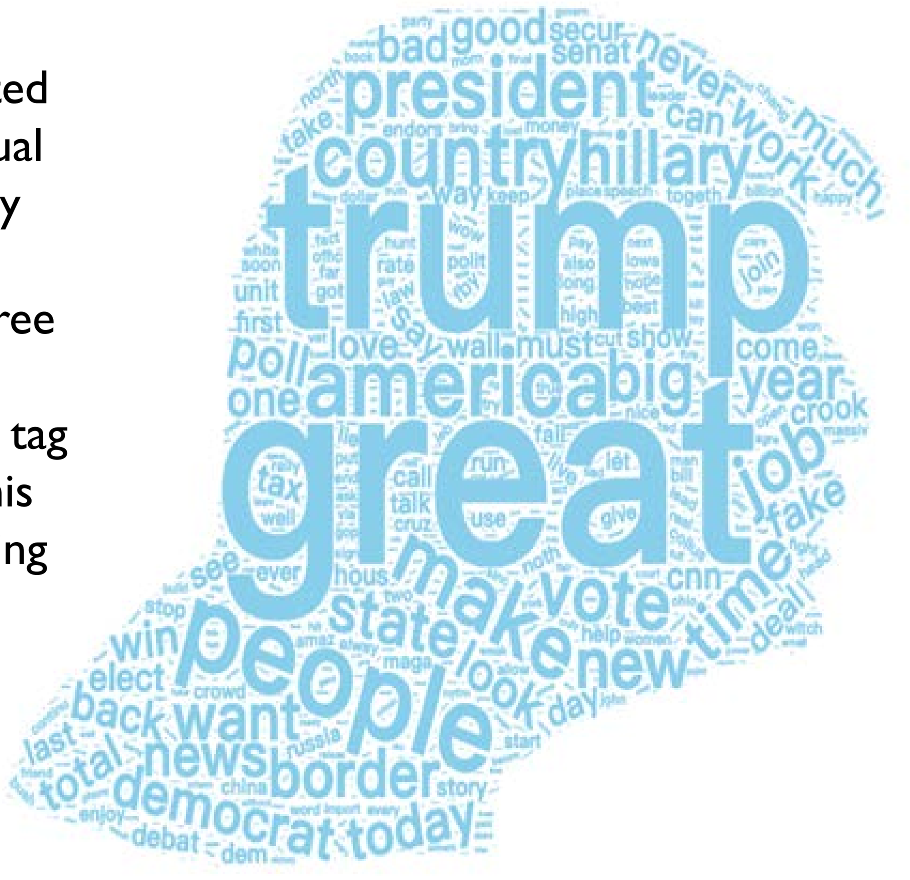


PRESIDENT TRUMP'S TWEETS WORD CLOUD



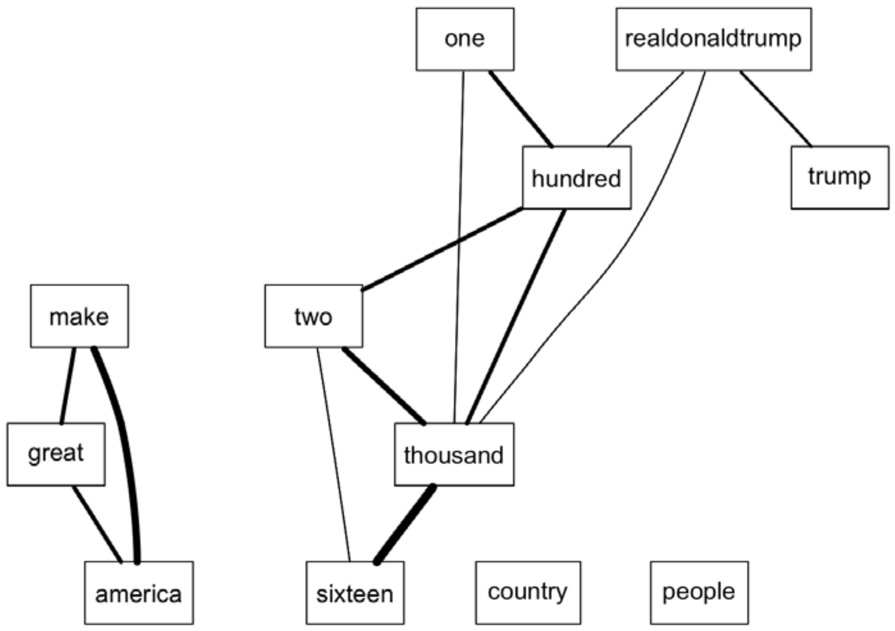
Word Cloud Using Wordcloud library

A tag cloud (word cloud, or weighted list in visual design) is a novelty visual representation of text data, typically used to depict keyword metadata (tags) on websites, or to visualize free form text. Tags are usually single words, and the importance of each tag is shown with font size or color. This format is useful for quickly perceiving the most prominent terms to determine its relative prominence.

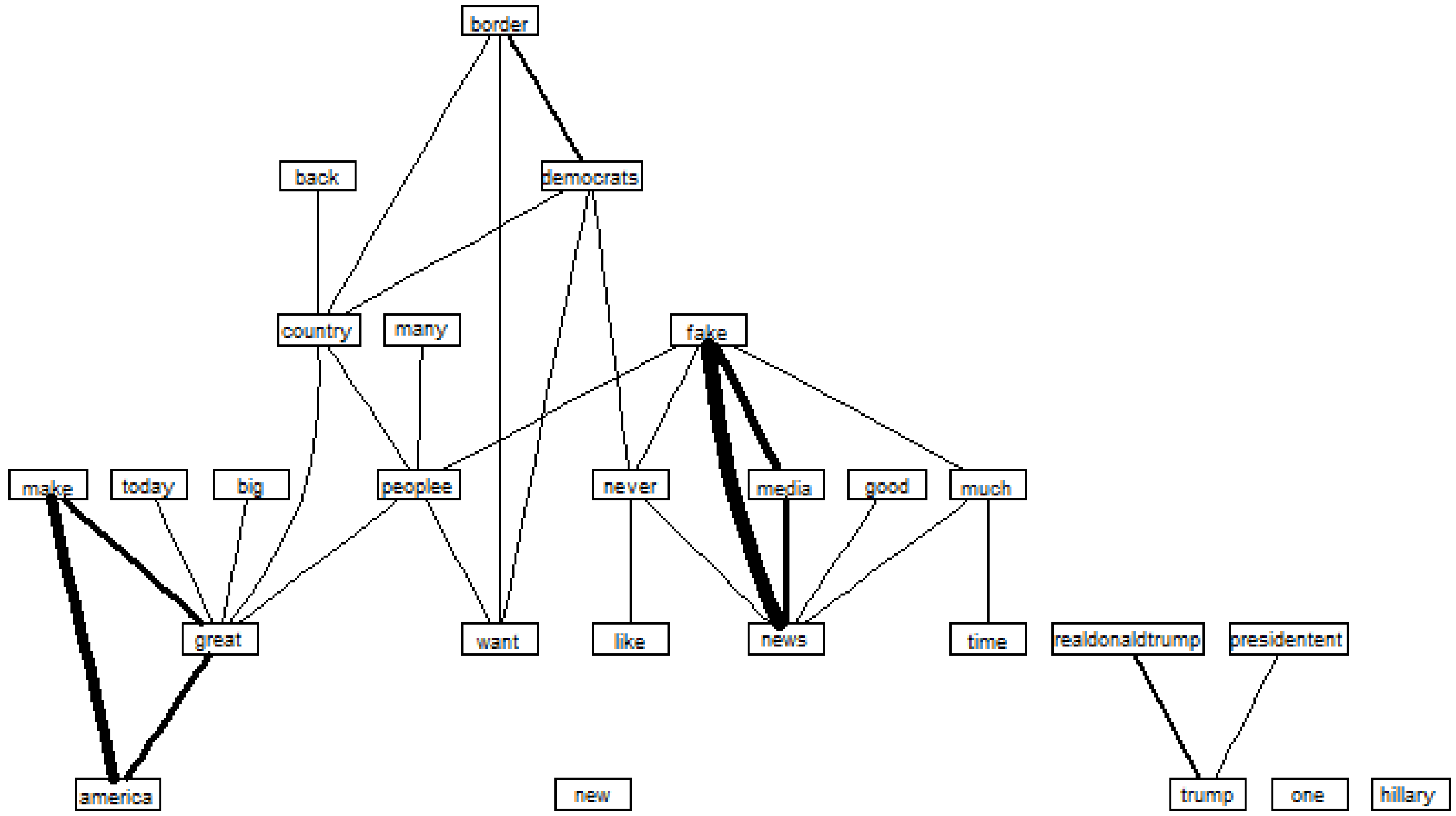


Word Cloud Using Wordcloud2 library

II. DOUBLE WORD FREQUENCIES: BIGRAMS AND WORD ASSOCIATION



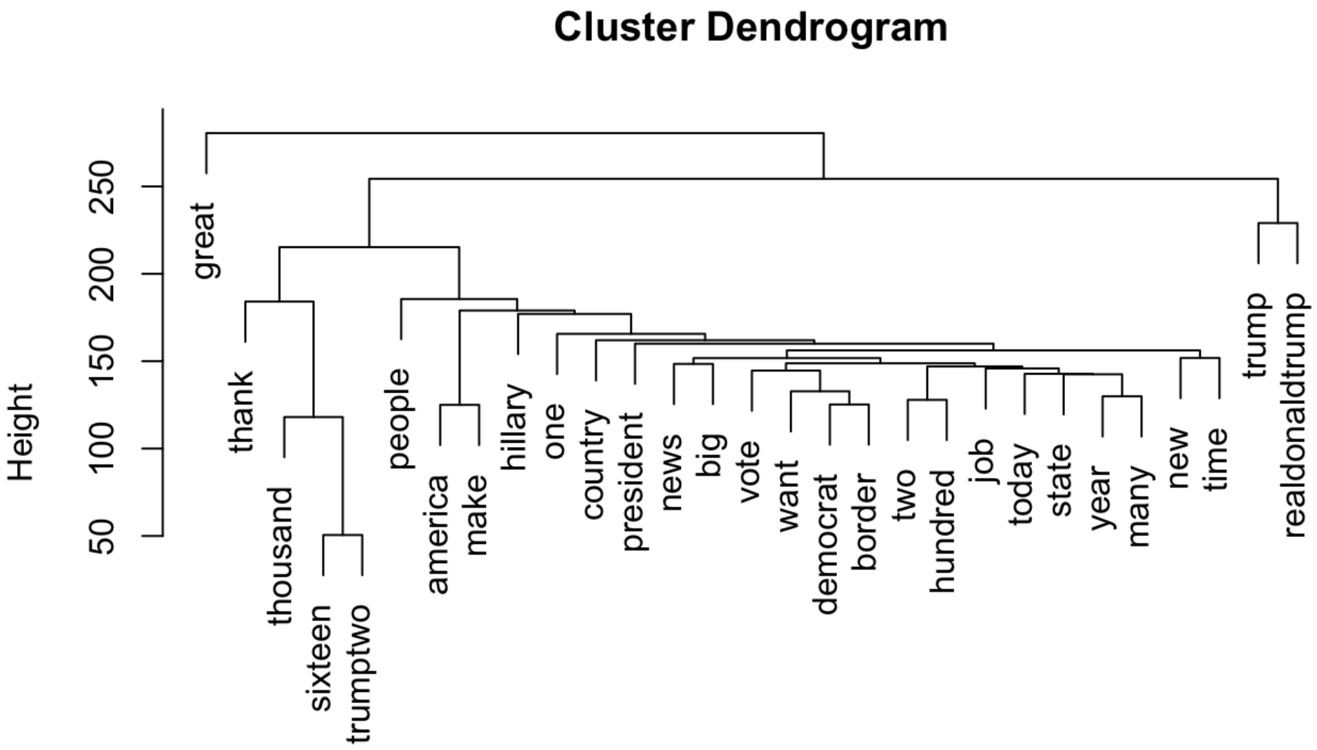
Bigram of top 800 word with 0.1 threshold with Weighting



Bigram of top 600 word with 0.05 threshold with Weighting

III. UNSUPERVISED MACHINE LEARNING APPLIED ON TRUMP'S TWITTER CORPUS

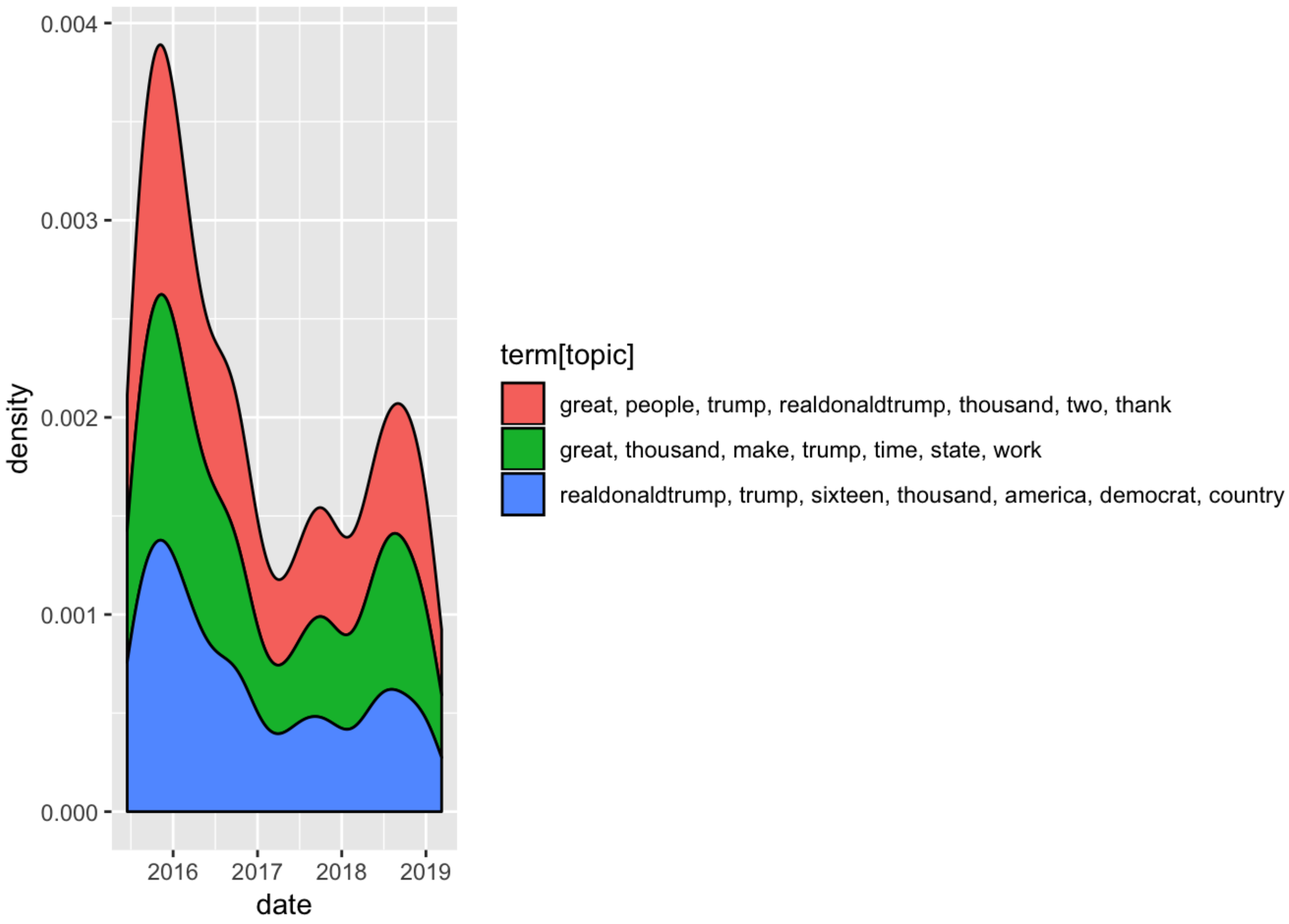
- In unsupervised machine learning approaches, no coding rules are specified and no annotated training data is provided. Instead, an algorithm comes up with a model by identifying certain patterns in text.
- Clustering method to explore topics Since topic modeling is inefficient, we wanted to use clustering method to see if that help with identify topics of trump's tweets.



IV. LDA MODEL (LATENT DIRICHLET ALLOCATION) :

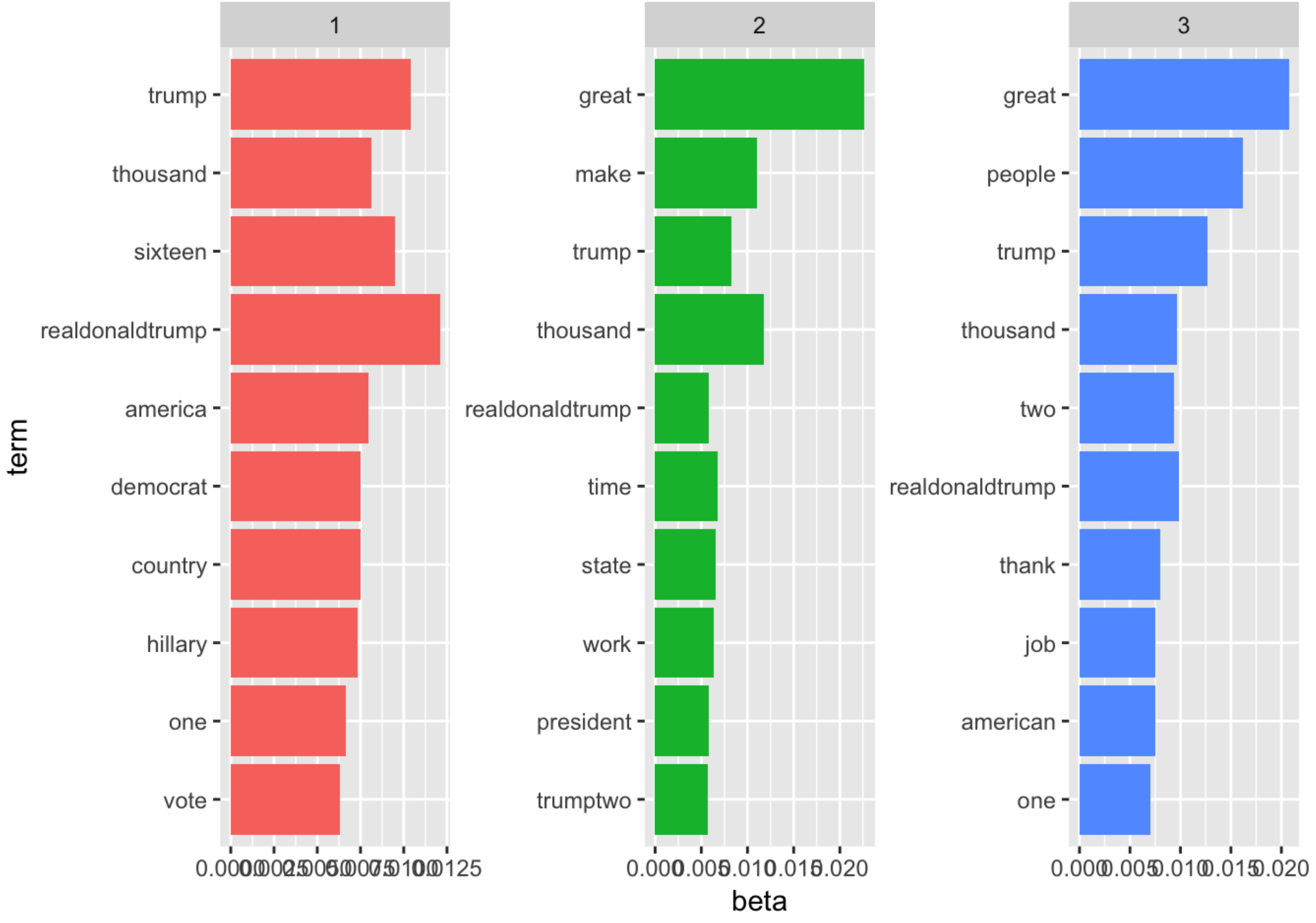
- The LDA (Latent Dirichlet Allocation) model (Blei et al. 2003) from *topicmodels* (Grun & Hornik 2011) package was used to identify major topics in Trump's tweets. To visualize the result, a stacked density graph is shown to express variations of each topic over time.

- Topic models became pretty popular in various fields of the humanities and social sciences (Schmidt, 2012). We apply the initially introduced and most widely used LDA topic model (Blei et al., 2003).

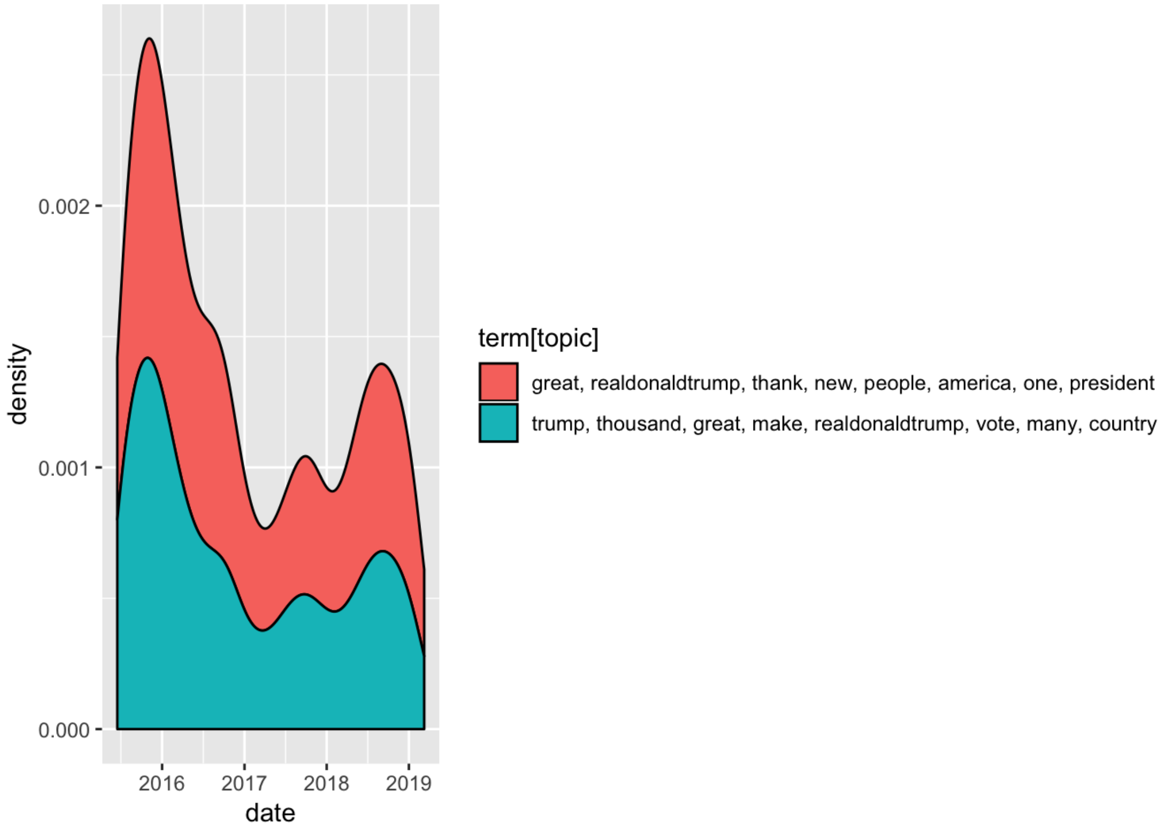


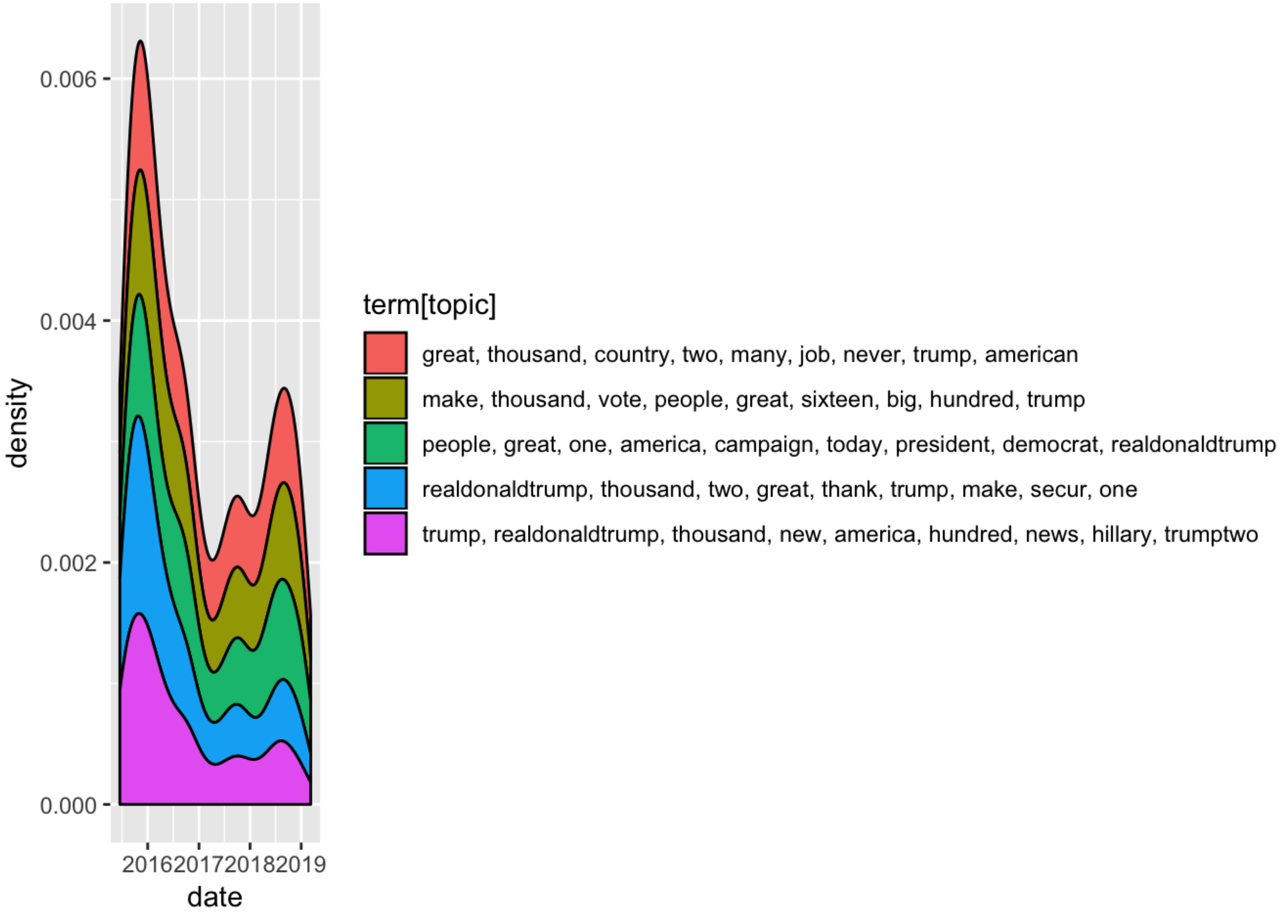
MORE LDA WITH THREE TOPIC SETTINGS:

- We wanted to build a topic model to further investigate our result. And we were able to apply our model to other data. Based on the density plot, there is a peak around the end of January. There are a lot of tweets that is about border wall around that time. That might be the fact that trump were insisting about building border wall so that he might mentioned so many times.



LDA WITH TWO TOPICS AND LDA WITH FIVE TOPICS



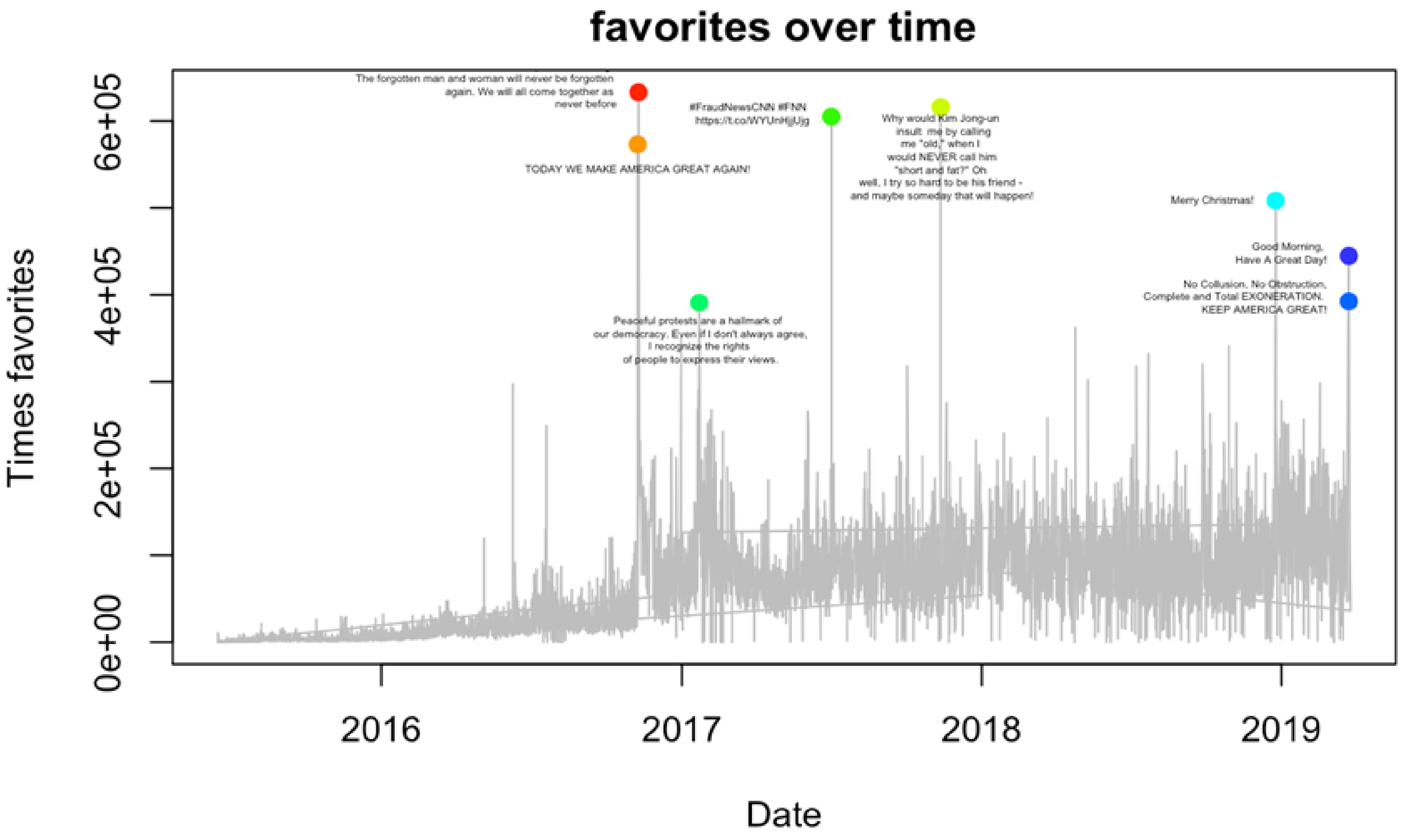


V. SUPERVISED MACHINE LEARNING:

- The supervised machine learning approach refers to all classes of techniques in which an algorithm learns patterns from an annotated set of training data.
- The intuitive idea is that these algorithms can learn how to code texts if we give them enough examples of how they should be coded.
- A straightforward example is sentiment analysis, using a set of texts that are manually coded as positive, neutral, or negative, based on which the algorithm can learn which features (words or word combinations) are more likely to occur in positive or negative texts.

VI. SENTIMENT ANALYSIS:

- In this figure , sentiment scores are computed based on the two time periods of tweets (One period is while he was working on his campaign, and another period is after he became the president).

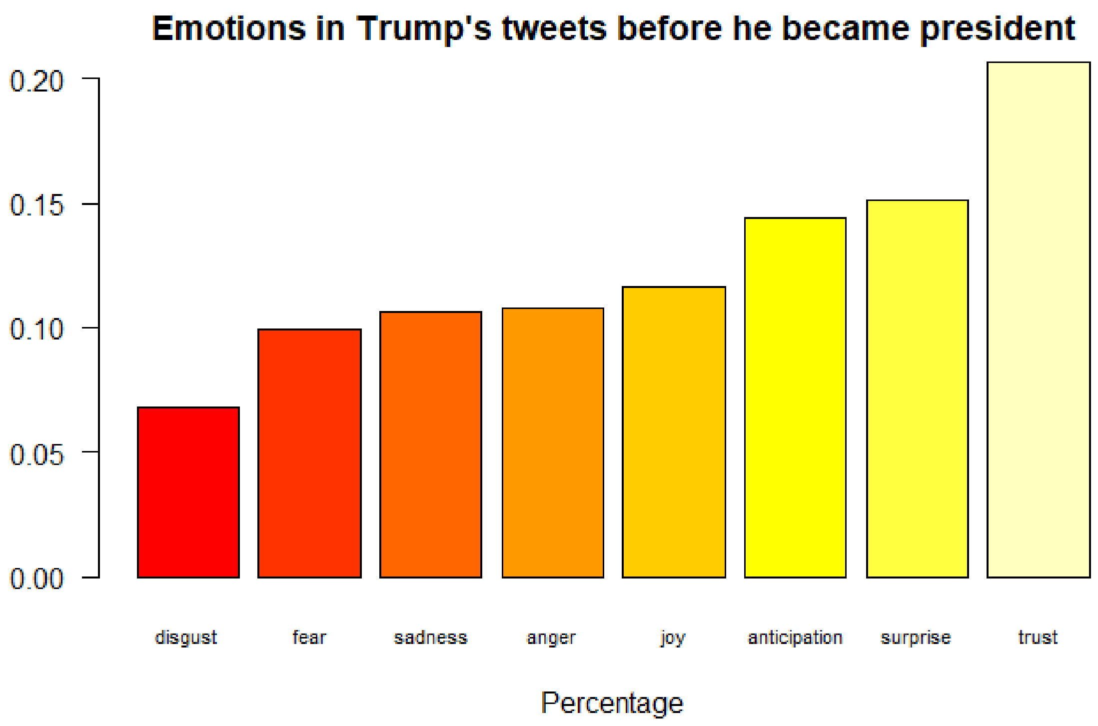


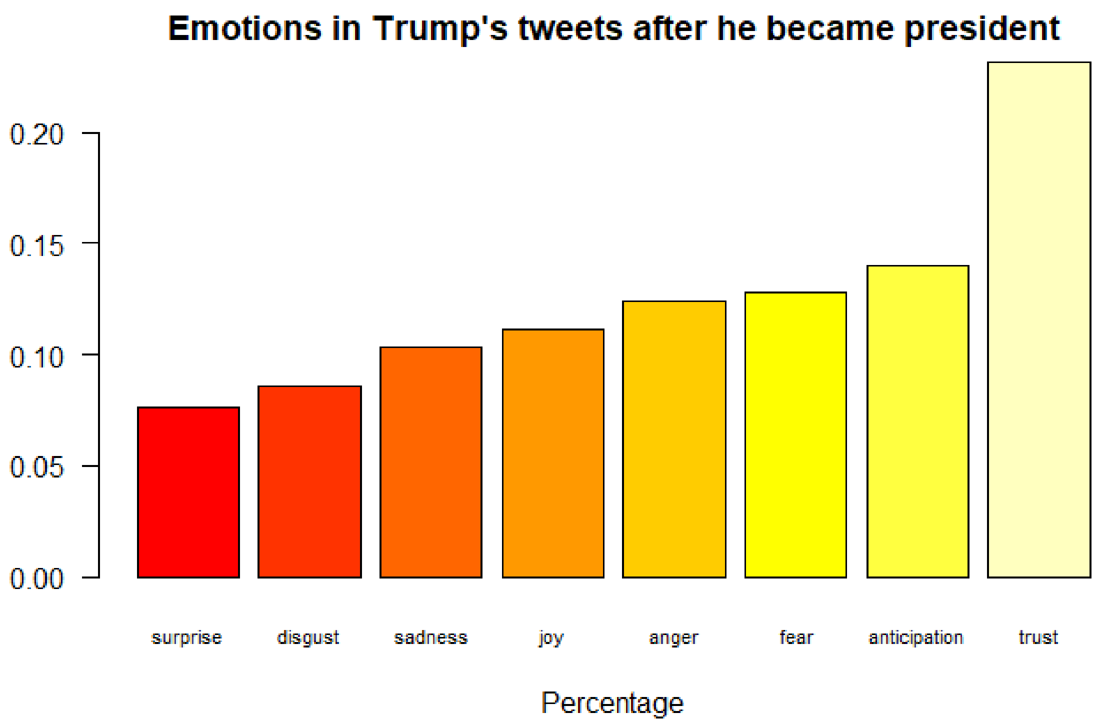
VI. SENTIMENT ANALYSIS:

- The get_nrc_sentiment implements Saif Mohammad's NRC Emotion lexicon. According to Mohammad, "the NRC emotion lexicon is a list of words and their associations with eight emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive)"

- We want to know what is president Trump's mood while he was in campaign. Our first impression while he was in campaign would be anger. He was arguing harshly with Hillary Clinton while debating. We assumed that his anger will be shown on his twitter.

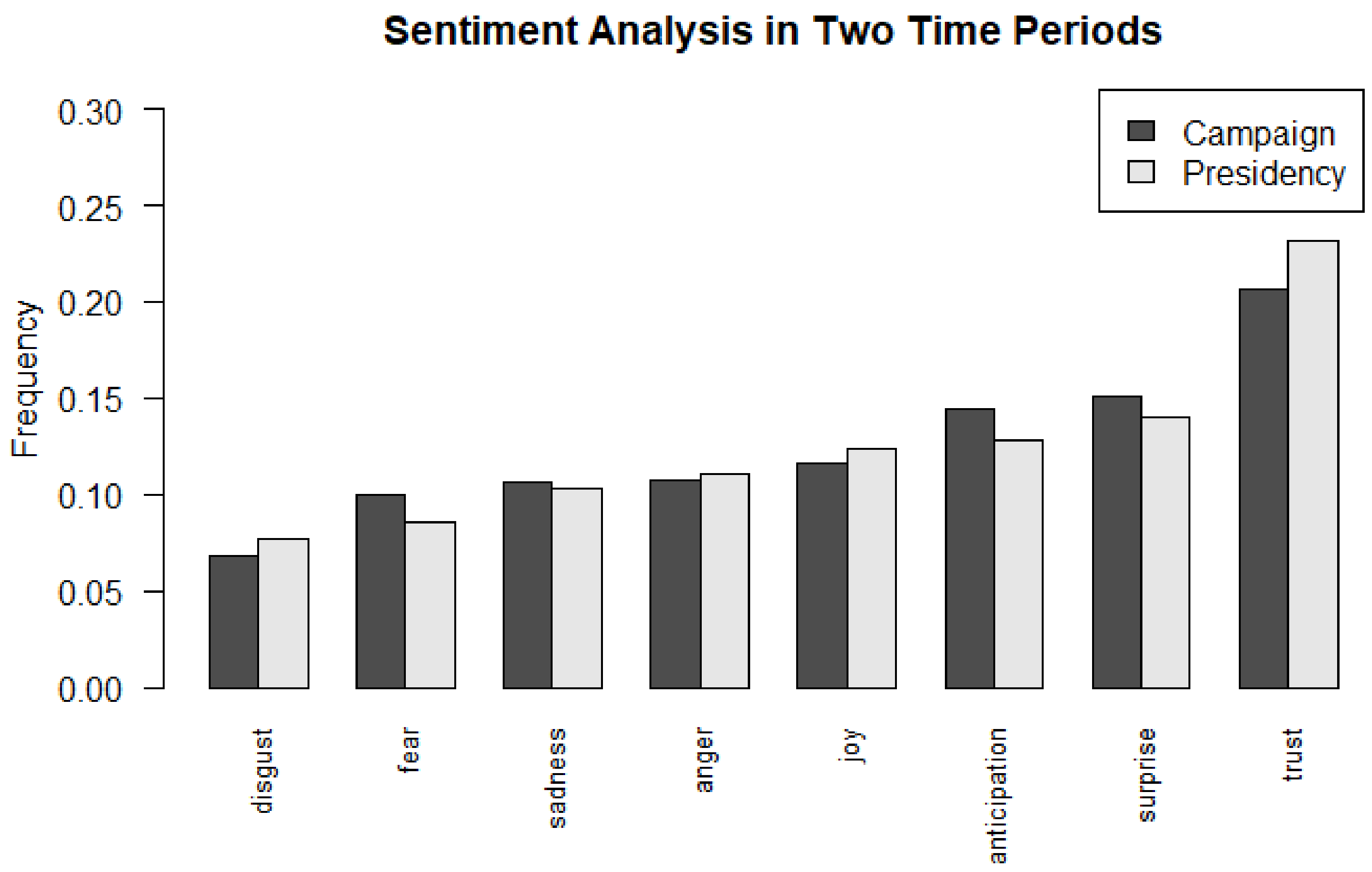
VI. SENTIMENT ANALYSIS: BEFORE AND AFTER ELECTIONS





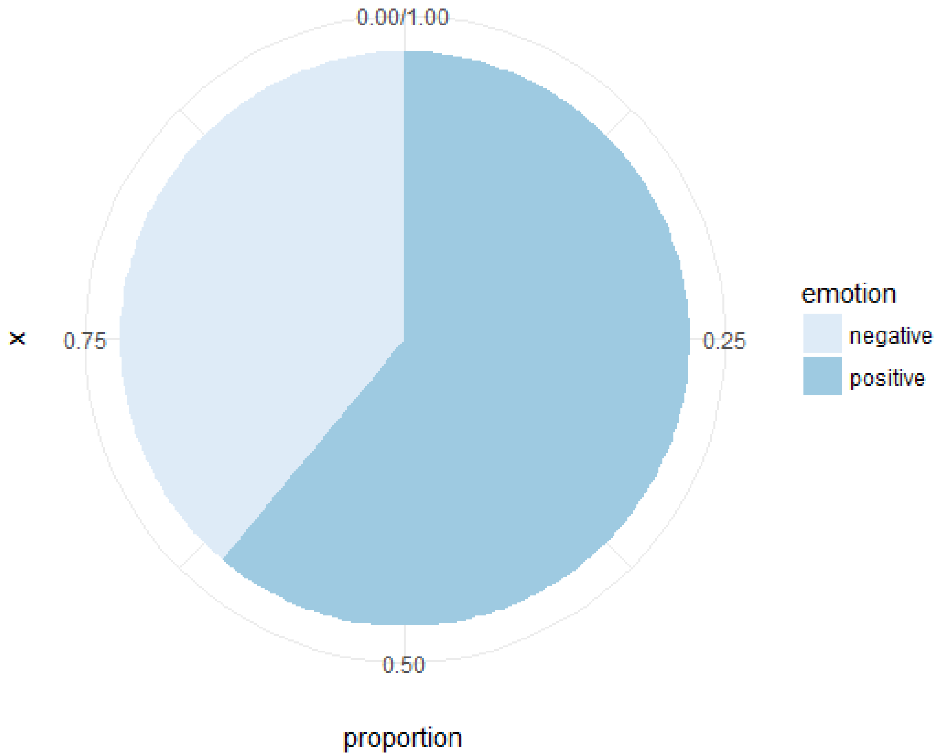
VI. SENTIMENT ANALYSIS: (CONT.)

- There are some minor differences between the two time periods. Generally, after Trump became president, he tended to post more positive tweets.



VI. SENTIMENT ANALYSIS (CONT.)

- We chose all tweets with emotion score above zero. We surprisingly found out that his major emotion was trust which is not what we expected at the beginning.
- The President's positive emotion took about 60% of the tweets. Then we started to looking at what are some of his emotion after he became the president.
- We used all the president's tweets after he became the president. His anger tweets at this point is about 11% before he became president.

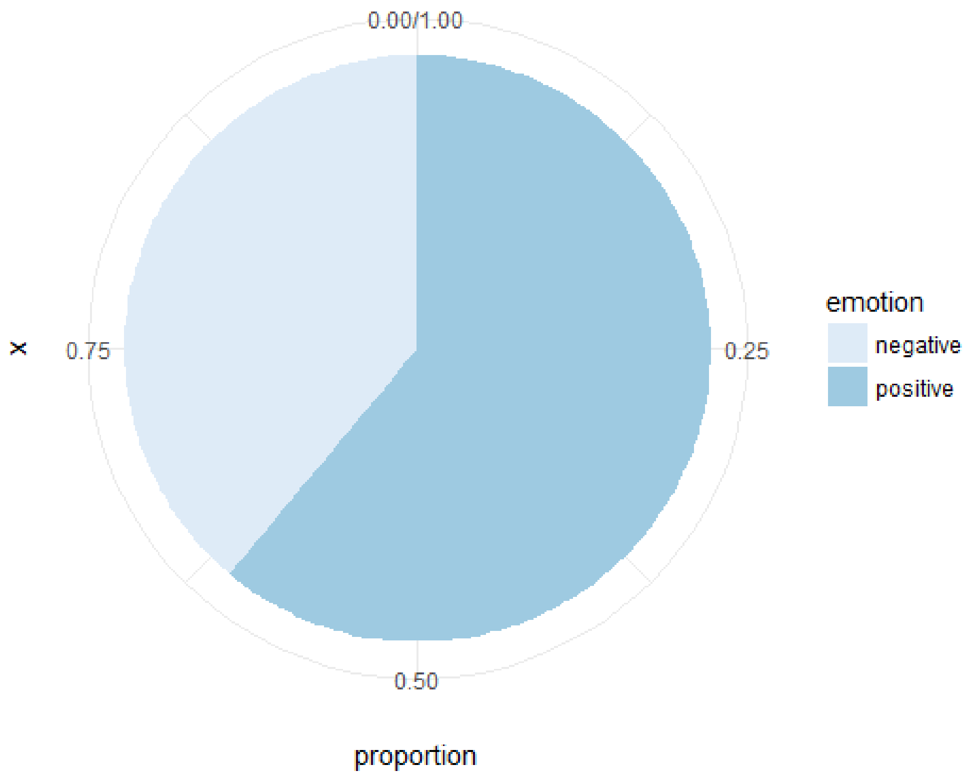


VI. SENTIMENT ANALYSIS: (CONT.)

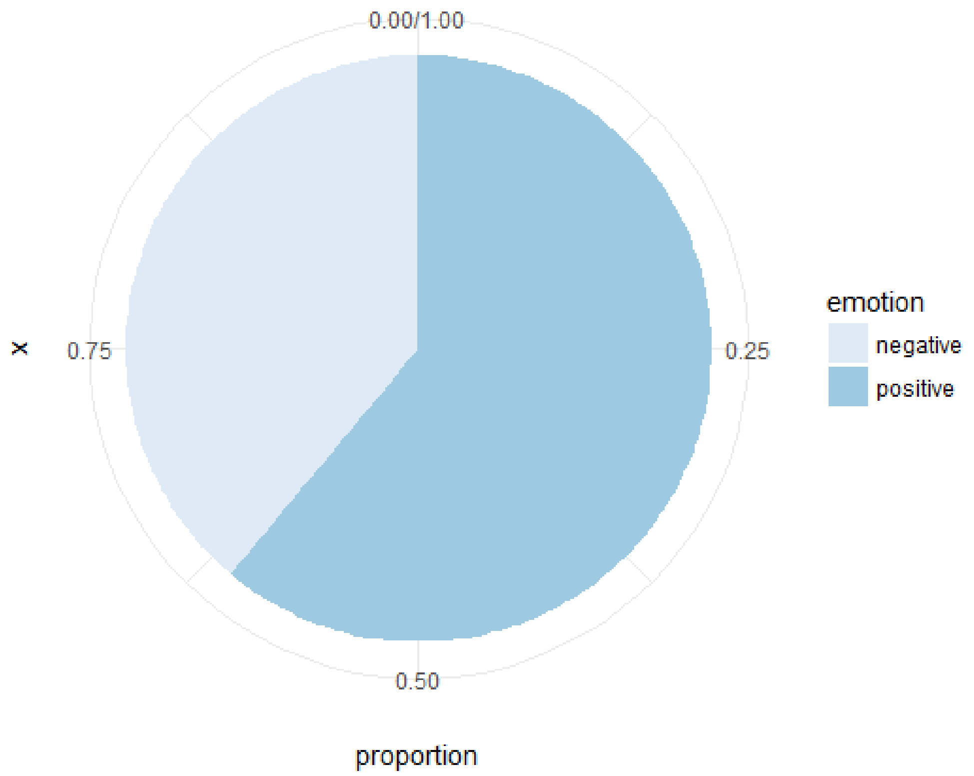
- An online study on showed that author's speculation about Trump only use Andriod cellphone to post tweets. “*The Man Behind the President’s Tweets, Unraveling the mystery of Dan Scavino, the White House social media director, whose job is to help @realDonaldTrump stay un-presidential.*”
- All other devices that are used to post tweets made by his team.
- Sentiment analysis will be used in order to distinguish sentiment differences between two types of devices.
- However, from the dataset, all tweets were posted from IOS or web since March 25th, 2017. Therefore, the sentiment analysis will cover tweets before that date.

VI. SENTIMENT ANALYSIS: (CONT.): ANDROID VS. NON-ANDROID

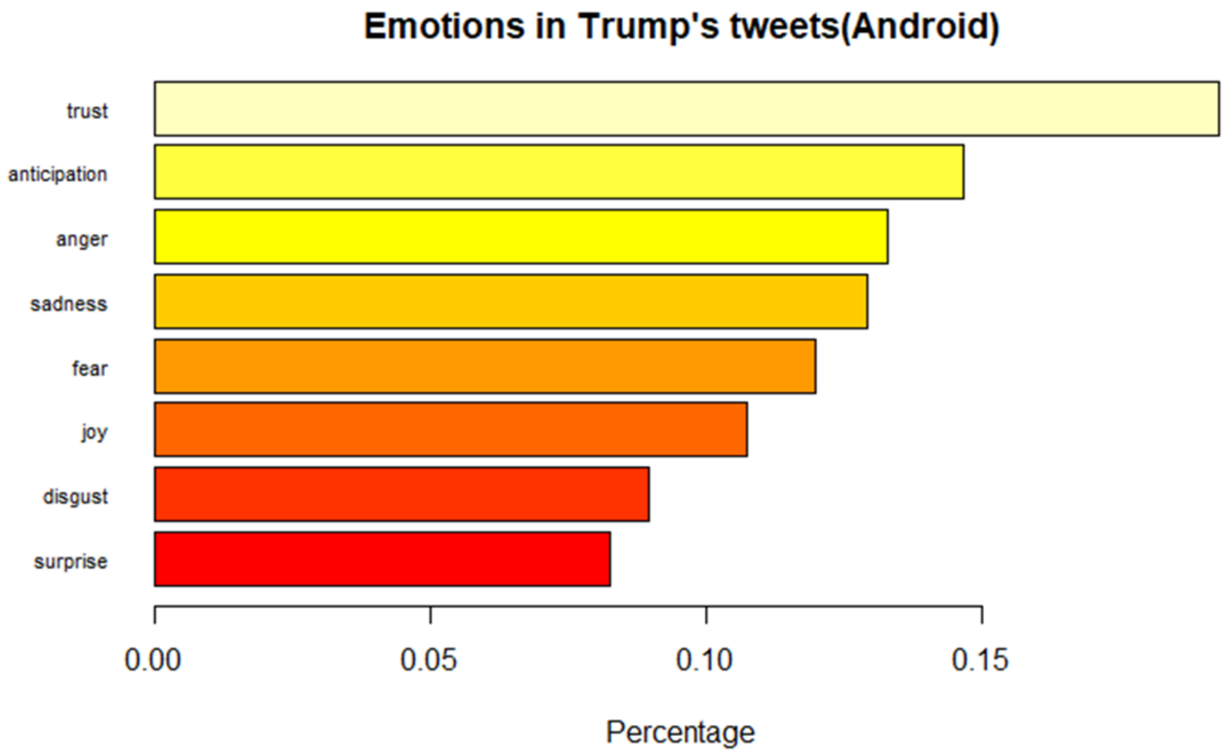
"Emotions in Trump's tweets(Android)"

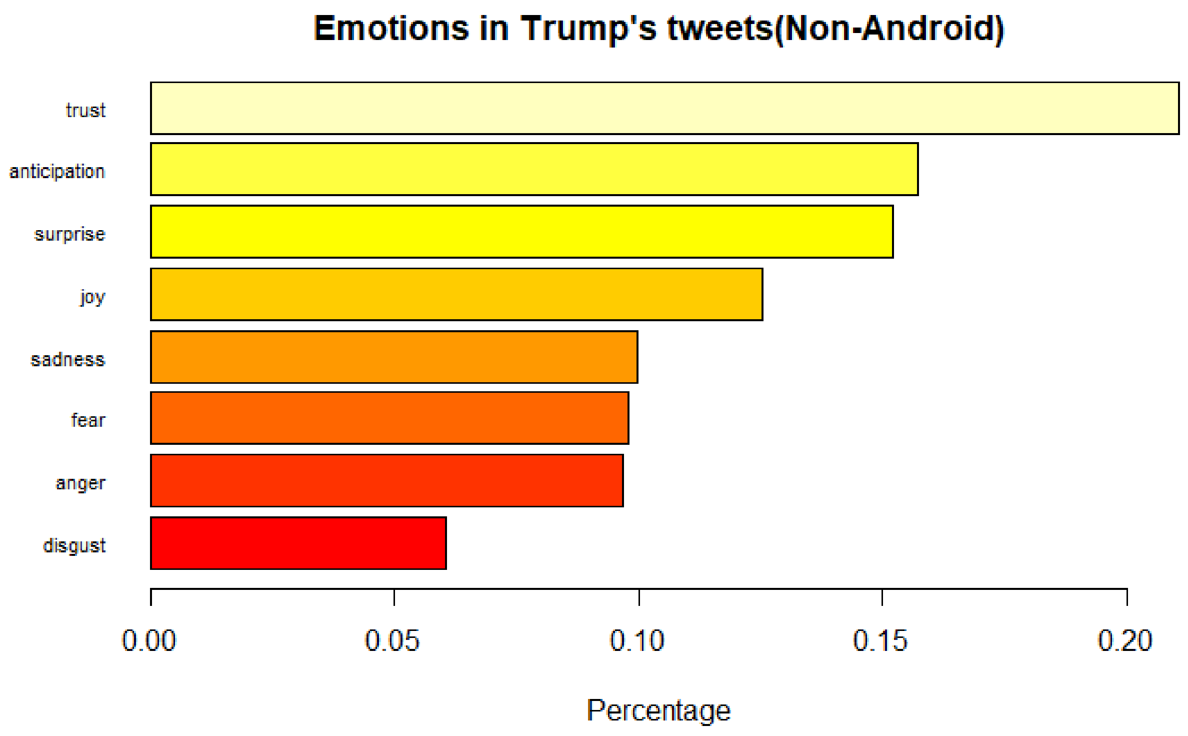


"Emotions in Trump's tweets(Non-Android)"



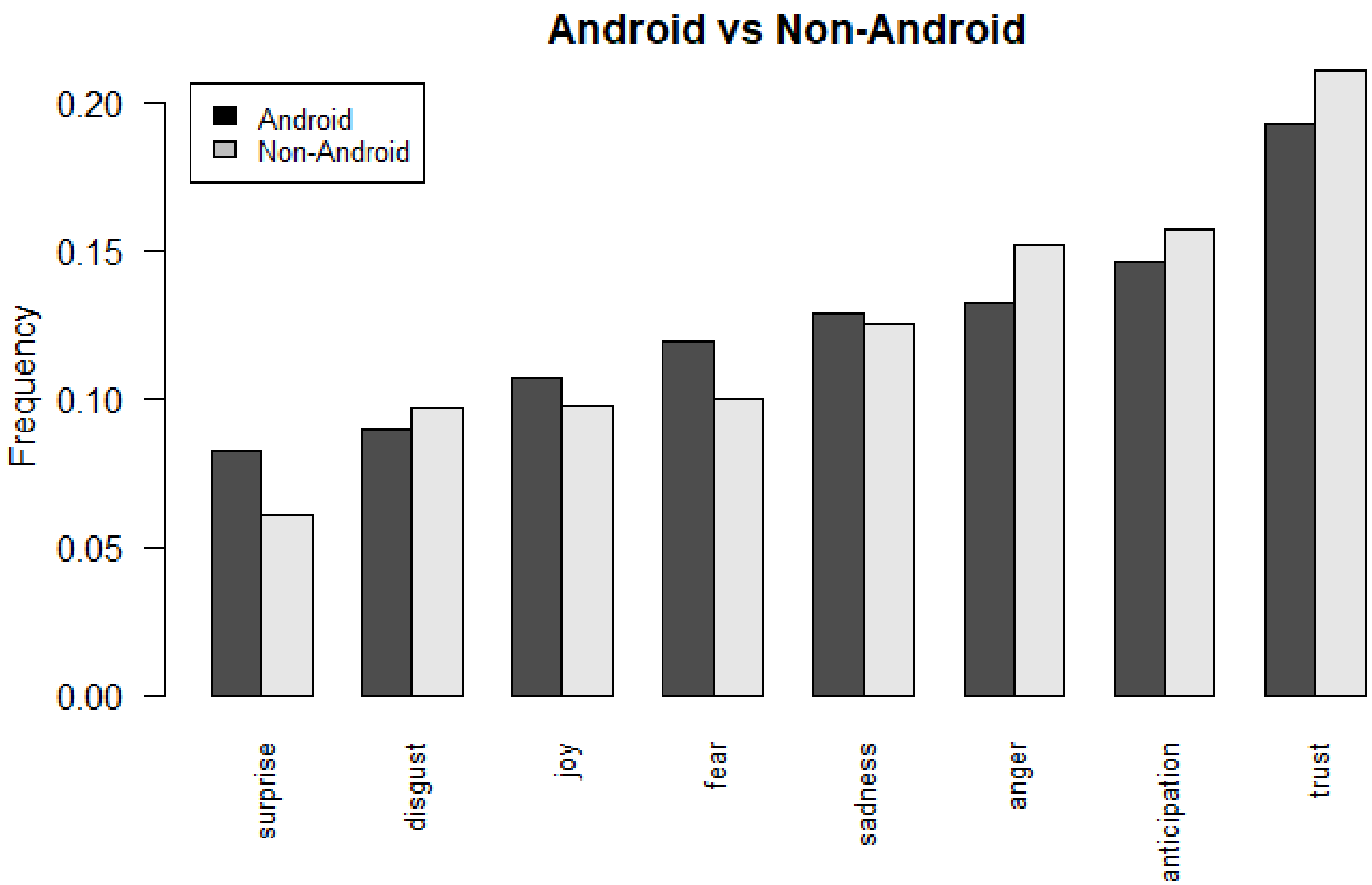
VI. SENTIMENT ANALYSIS: (CONT.): ANDROID VS. NON-ANDROID





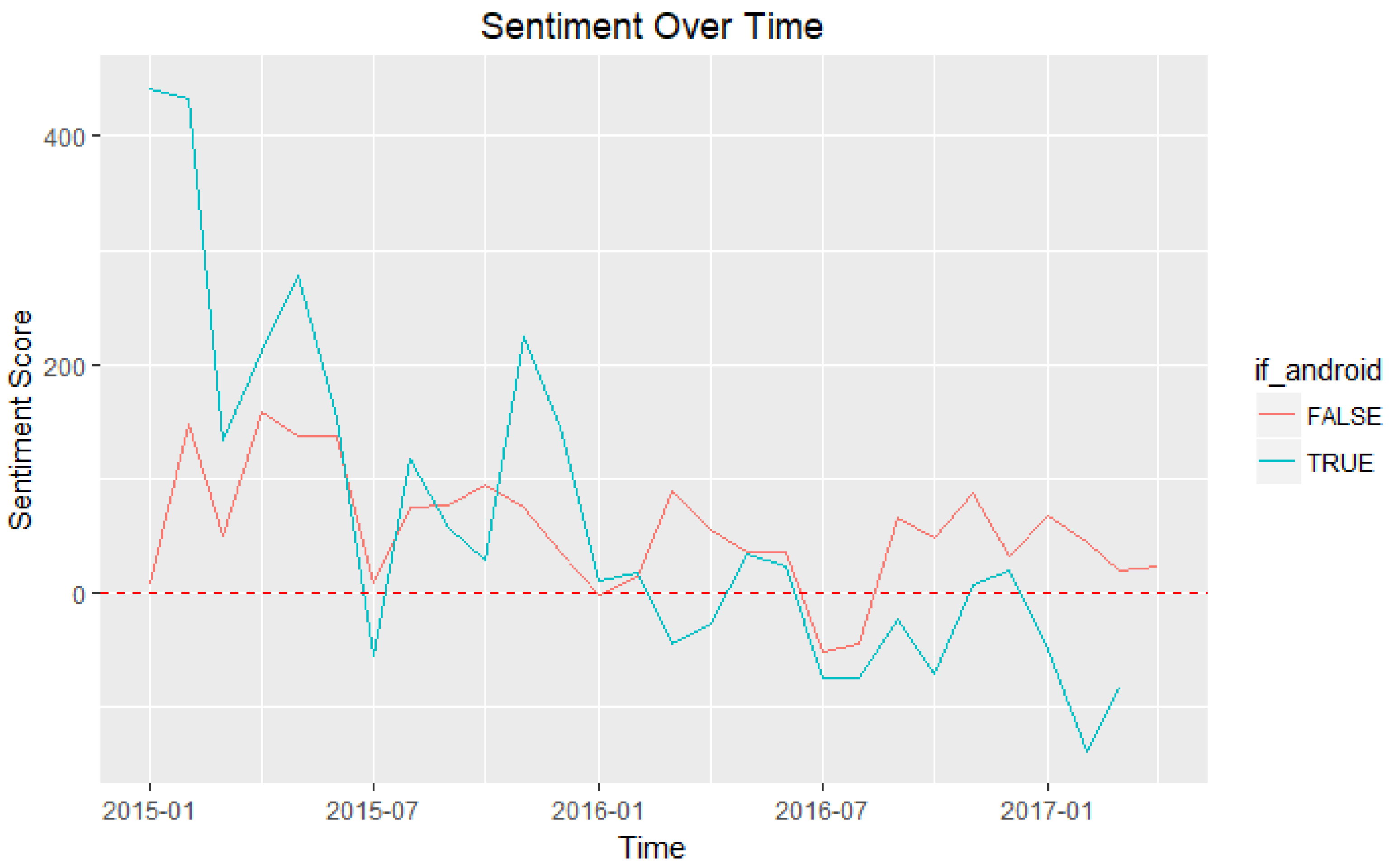
VI. SENTIMENT ANALYSIS: (CONT.): ANDROID VS. NON-ANDROID

- Sentiments of tweets from Andriod device are slightly more negative than sentiments of tweets from non-Android devices.
- The speculation might be true. However, there are still no clear differences in sentiments between the two types of posting.



VI. SENTIMENT ANALYSIS: SENTIMENT OVER TIME

- Sentiment scores over time of two different types of posting methods are graphed. In this case, *afinn* lexicon was used to simply determine if a tweet is positive or negative.
- The sentiment of tweets from Andriod device are more negative starting in 2016, but sentiment scores of tweets from non-Android devices are relatively stable.
- There are some slight differences between tweets from the Android device and non-Android devices. The speculation mentioned before might be real. However, more significant results are needed to prove the speculation.



VII. SOCIAL NETWORK ANALYSIS

- This part is mainly focusing on

A. Who and how many followers are following @RealDonaldTrump : Followers' locations.

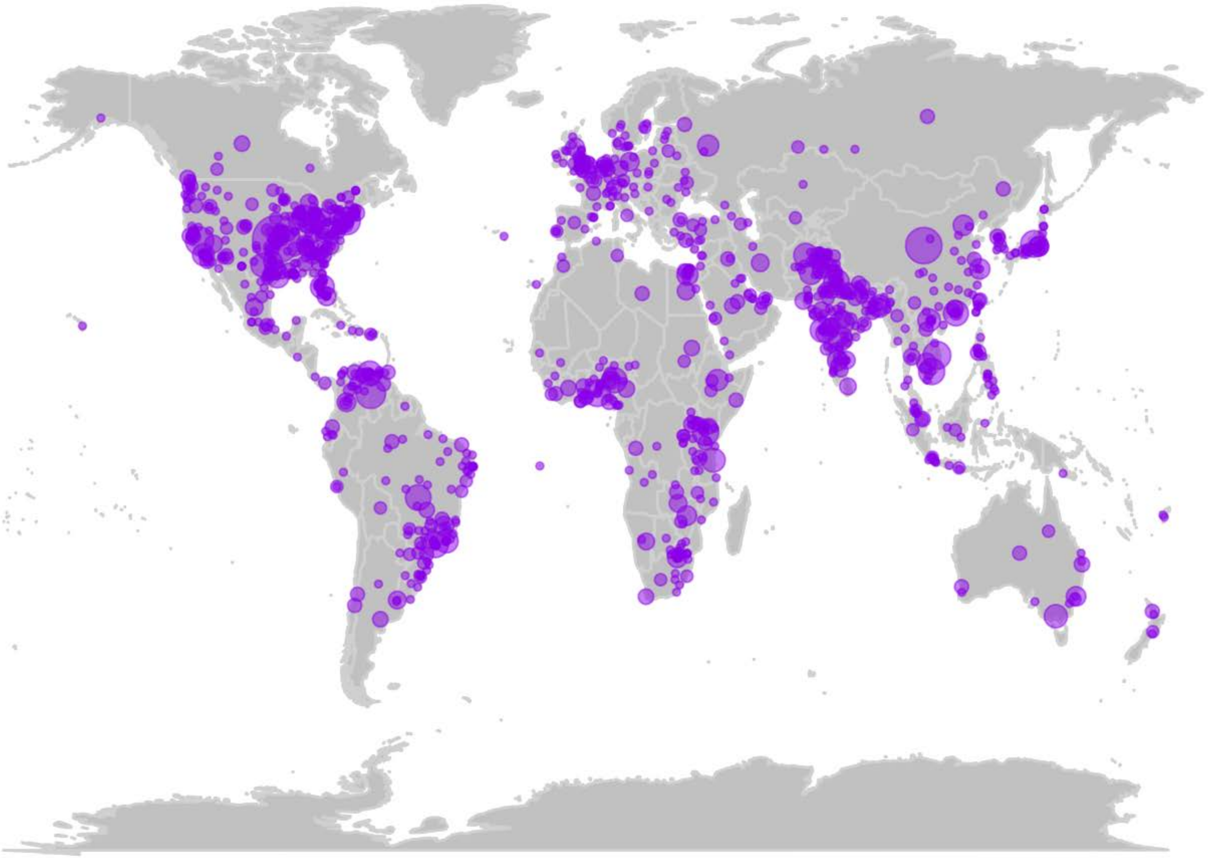
- There are 59.3 millions of twitter account who are currently following President Trump.
- In order to obtain information about followers, we used the Twitter API to access the information about followers on the website. The Twitter API enables us to obtain geographic information about Trump's followers.
- A dataset included Twitter account ID, location, IP Address etc. is downloaded from the Twitter website.

B. Which tweets Trump tweeted have large counts of favorites and large counts of retweeting.

- This section analyzes the most popular and influential tweets from Donald J. Trump via the retweeting and favorite features from Twitter.
- A graph will show tweets posted by Trumps that have been retweeted. Tweets with counts of retweeted over 120000 are selected in the graph (forms top 8 retweeted tweets).
- The most retweeted tweet was tweeted on July 20th, 2016 about Hillary Clinton's emails. The second highest retweeted tweet was tweeted on November 8th, 2016 about his slogan, "Make America great again".

A. PRESIDENT TRUMP TWITTER FOLLOWERS WORLDWIDE

Based on a random Sample of size 10,000 from all the followers.

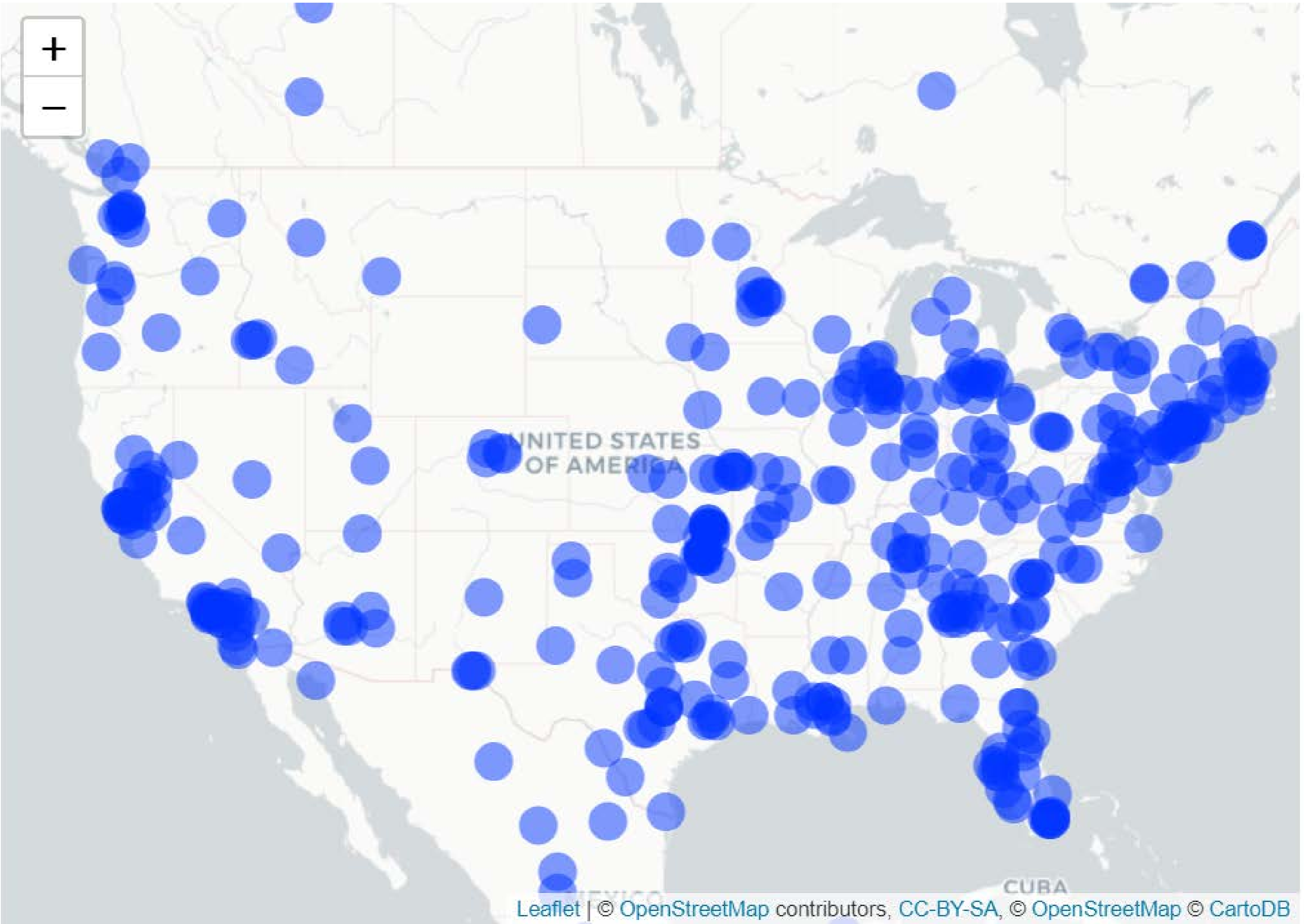


Country	Frequency
USA	33.41%
India	8.69%
Brazil	4.86%
UK	3.03%
Canada	2.31%

A. PRESIDENT TRUMP TWITTER FOLLOWERS IN THE UNITED STATES

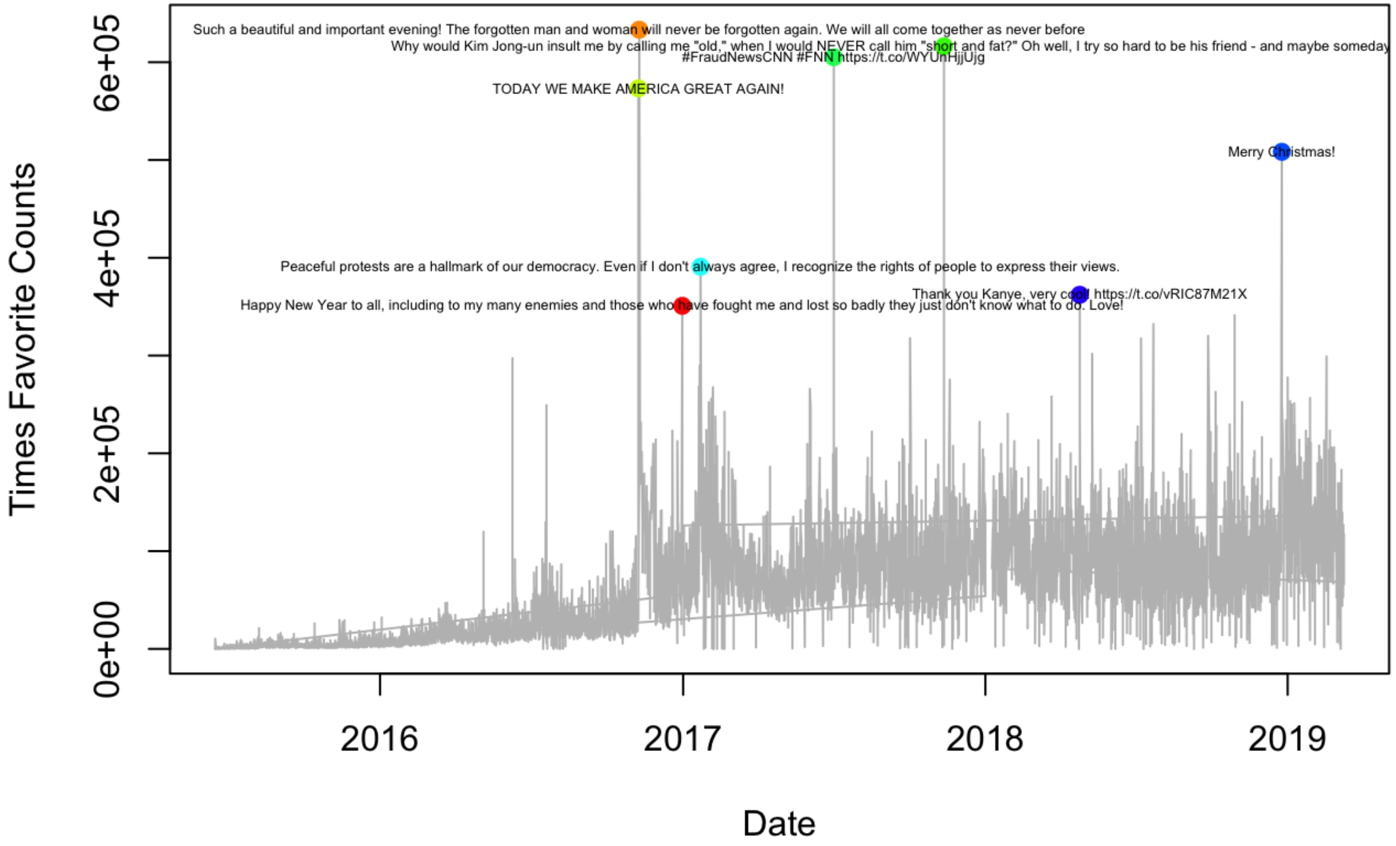
- Based on a random Sample of size 10,000 from all the followers.
- Therefore, only 1254 out of 10000 followers are graphed.

States	Frequency
CA	7.39%
TX	6.92%
FL	5.49%
NY	4.77%
NC	2.15%



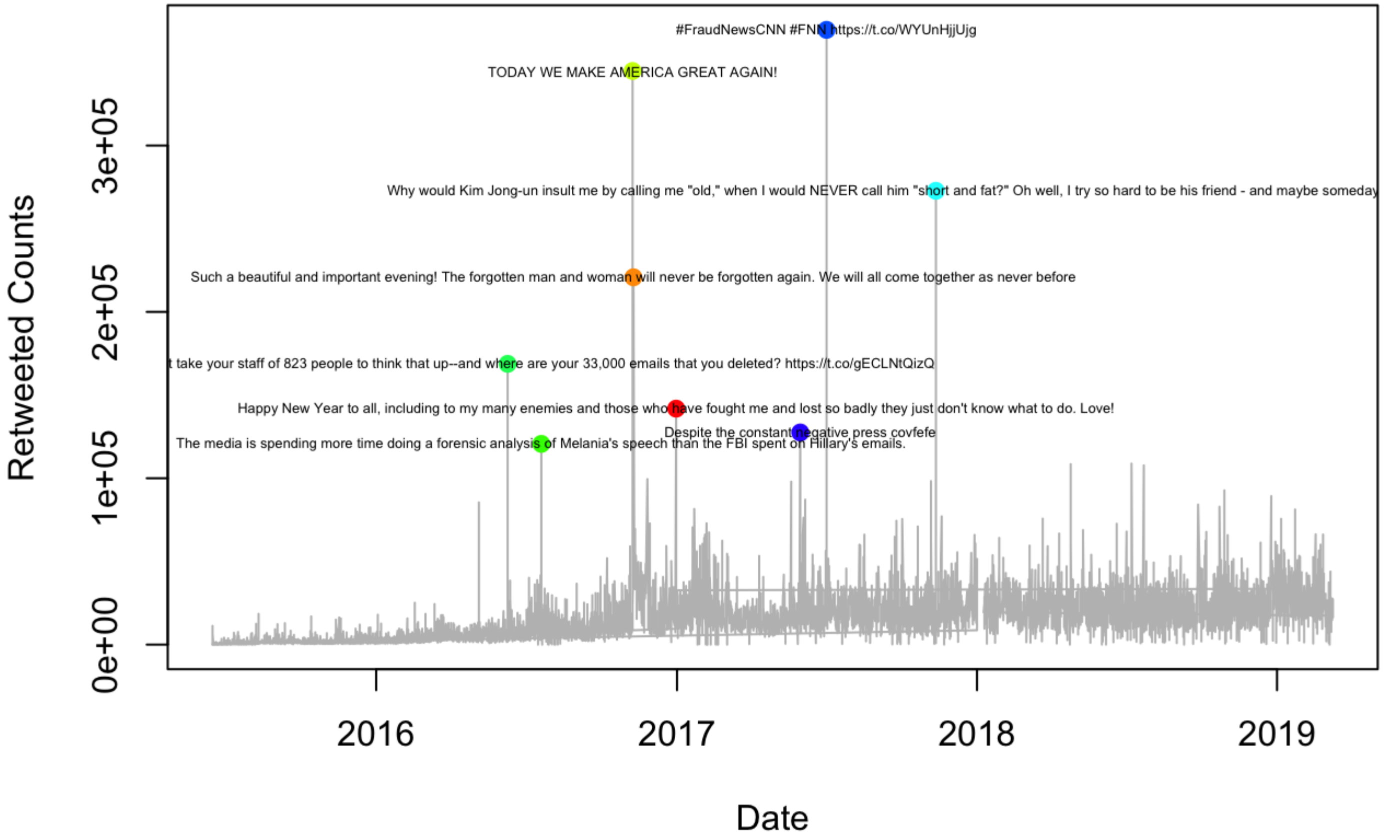
B. TWEETS: MOST FAVORITES AND MOST RETWEETED

- Happy New Year to all, including to my many enemies and those who have fought me and lost so badly they just don’t know what to do. Love!
- Such a beautiful and important evening! The forgotten man and woman will never be forgotten again. We will all come together as never before
- TODAY WE MAKE AMERICA GREAT AGAIN!
- Why would Kim Jong-un insult me by calling me “old,” when I would NEVER call him “short and fat?” Oh well, I try so hard to be his friend - and maybe someday that will happen!
- #FraudNewsCNN #FNN https://t.co/WYUnHjjUjg
- Peaceful protests are a hallmark of our democracy. Even if I don’t always agree, I recognize the rights of people to express their views.
- Merry Christmas!
- Thank you Kanye, very cool! https://t.co/vRIC87M21X
- According to the two plots above, the retweeted counts and favorite couonts are consistent. The top favorite tweets are pretty identical to the top retweeted tweets. The timeline displays that Trump’s tweets have been paid much higher attention than his other tweets before campaign, whcih makes sense.



B. TWEETS: MOST FAVORITES AND MOST RETWEETED

- Happy New Year to all, including to my many enemies and those who have fought me and lost so badly they just don't know what to do. Love!
- Such a beautiful and important evening! The forgotten man and woman will never be forgotten again. We will all come together as never before
- TODAY WE MAKE AMERICA GREAT AGAIN!
- The media is spending more time doing a forensic analysis of Melania's speech than the FBI spent on Hillary's emails.
- How long did it take your staff of 823 people to think that up–and where are your 33,000 emails that you deleted? https://t.co/gECLNtQizQ
- Why would Kim Jong-un insult me by calling me "old," when I would NEVER call him "short and fat?" Oh well, I try so hard to be his friend - and maybe someday that will happen!
- #FraudNewsCNN #FNN https://t.co/WYUnHjjUjg
- Despite the constant negative press covfefe



FUTURE STUDIES:

- Although our study is informative and intellectually rewarding, there are many potential areas for future study:
- Incorporating more text data from Twitter of former presidents of the United States. One clear purpose of text mining is to help humans understand social contexts better. Also, comparing US president's tweets with other world leaders tweets.
- Statistical programs currently have limited support for foreign languages. It will be a great future research to compare between multiple world presidents with multiple languages. Our results provide quantitative evidence to further investigate texts in other languages.
- Further investigation of markers or predictors. For example, although President Trump has few strong markers that were identified we could further break it down into specific syntax and sentence structures, such as classifying tweets sentences based on some characteristics to be able to classify tweet based on the Author markers. This would increase the accuracy of predictions.

REFERENCES:

- Jackers, M. (2017 December), 'Introduction to the Syuchet Package', *cran-r project*
- **URL:** https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html
- Puente L. (2016), 'Geo-Mapping Twitter User ', *R pubs*,
- **URL:** https://rpubs.com/cosmopolitanvan/geotwitter
- Zhang, Y. (2015 October), 'R and Data Mining: Examples and Case Studies', *RDataMining.com.*
- Murrell, H. & Zhang, Y. (2015), 'Data Mining with R', *RDataMining.com.*
- Lang, D. (2018 January), 'Wordcloud2 Introduction', *cran-r project.org.*
- **URL:** https://cran.r-project.org/web/packages/wordcloud2/vignettes/wordcloud.html
- Davis, M. (2017), 'Introduction to Wordclouds in R', *R pubs,*
- **URL:** http://rpubs.com/ecedavis/intro-to-wordclouds
- Prabhakaran, S. (2016), 'The complete ggplot2 Tutorial - Part 1 | Introduction to ggplot2 (Full R code)', r-statistics.co
- **URL:** http://r-statistics.co/Complete-Ggplot2-Tutorial-Part1-With-R-Code.html
- Deckmyn, A. (2018), map, R package version 3.3
- **URL:** https://cran.r-project.org/web/packages/maps/maps.pdf
- Feinerer, I. & Hornik, K. (2013), tm: Text Mining Package. R package version 0.5-8.3.
- **URL:** http://CRAN.R-project.org/package=tm
- Fellows, I. (2013), wordcloud: Word Clouds. R package version 2.4.
- **URL:** http://CRAN.Rproject.org/package=wordcloud
- Gentry, J. (2013), twitteR: R based Twitter client. R package version 1.1.6.
- **URL:** http://CRAN.Rproject.org/package=twitteR
- Lang, D. T. (2013), RCurl: General network (HTTP/FTP/...) client interface for R. R package version 1.95-4.1.
- **URL**: http://CRAN.Rproject.org/package=RCurl
- Wickham, H. (2018), ggplot2. R package version 3.1
- **URL:** http://CRAN.Rproject.org/package=ggplot2
- Jocker, M. (2017), syuzhet. R package version 1.0.4.
- **URL:** http://CRAN.Rproject.org/package=syuzhet
- Wickham, H. (2019), dplyr, R package version 0.8.1.
- **URL:** http://CRAN.Rproject.org/package=dplyr
- Spinu, V. (2019), lubridate, R package version 1.7.4.
- **URL:** http://CRAN.Rproject.org/package=lubridate
- Grun, B. (2018), topicmodelsr, R package version 0.2-8.
- **URL:** http://CRAN.Rproject.org/package=topicmodels

REFERENCES:

- Wickham, H. (2017), tidyverse, R package version 1.2.1.
- **URL:** http://CRAN.Rproject.org/package=tidyverse
- Quieroz, G. (2018), tidytext, R package version 0.2..
- **URL:** http://CRAN.Rproject.org/package=tidytext
- Goodrich, B. (2019), qdapr, R package version 2.3.2.
- **URL:** http://CRAN.Rproject.org/package=qdap
- Lang, D. (2019), wordcloud2, R package version 0.2.1.
- **URL:** http://CRAN.Rproject.org/package=wordcloud2
- Cheng, J. (2019), leaflet, R package version 2.0.2.
- **URL:** http://CRAN.Rproject.org/package=leaflet
- Dowle, M. (2019), data.table, R package version 0.8.1.
- **URL:** http://CRAN.Rproject.org/package=data.table

THANK YOU

STILL SO MUCH TO LEARN….