Analyzing Donald J. Trump's Tweets with Text Mining and Sentiment Analysis

Yan Kang, Citina Liang
Supervisor: Dr. Akram M. Almohalwas
University of California, Los Angeles - Statistics Department

Abstract

Twitter, as one of the most popular social network media and microblogging platform, is an optimal place for celebrities to build quality audiences, and connect to their defenders. One of the most typical examples is Donald J. Trump, who continuing being active on Twitter since his official declaration of candidacy in June 2015. This research project performs text mining, sentiment analysis and social network analysis on Donald J. Trump's tweets data and followers' information. First, Donald J. Trump's tweets context is studied with text mining techniques and unsupervised modeling to identify inner relations of terms and to distinguish topics. Second, social network analysis was done to Donald J. Trump's favorited tweets counts, retweeted tweets counts and followers' information. Third, Donald J. Trump's tweets sent from IOS source and Android source during the campaign period. The methodology and libraries used are done with R, and they can be applied to any analysis of Twitter account in general.

Keywords: Twitter, Donald J. Trump, social media, text mining, sentiment analysis, unsupervised learning

I. Introduction

The growing power of social media brings inestimable opportunities for companies, institutions, and politicians to build quality audiences. The social media platforms can be one of the most economical and accessible ways to promote their ideas and build up their charisma. Twitter[1], as one of the most popular social media, people are allowed to post and interact directly with each other with messages called "tweets", and there are 321 million active monthly users as of February 2019. Therefore, it is a choice platform for promotions. Donald J. Trump[2]'s success in the election is one of the most symbolic cases.

The crafty strategies of the way Donald J. Trump uses social media inducing expanding interests in his Twitter data. This research aims to analyze the Donald J. Trump's Twitter data to answer the most trendy and intriguing inquiries. To gain better insight into his tweets context, and furthermore, the sentiment variation between his tweets; text mining, social network analysis, and sentiment analysis are included in the rest of the paper. All the analysis in this report is done with R[3].

The following of the report is established in five sections. The second section introduces where and how Donald J. Trump's twitter data were approached. The third section shows procedures of data cleaning and text mining including term network and topic modeling. The fourth section switched from text mining to social network analysis, Donald J. Trump's followers' data are used in this section. The Fifth section provides sentiment analysis. The last section concludes the results of the report and proposes future work can be done.

II. Twitter Data

A. Donald J. Trump's Twitter Account

[1] http://twitter.com

[2] The 45th president of the United State

[3] https://www.r-project.org/

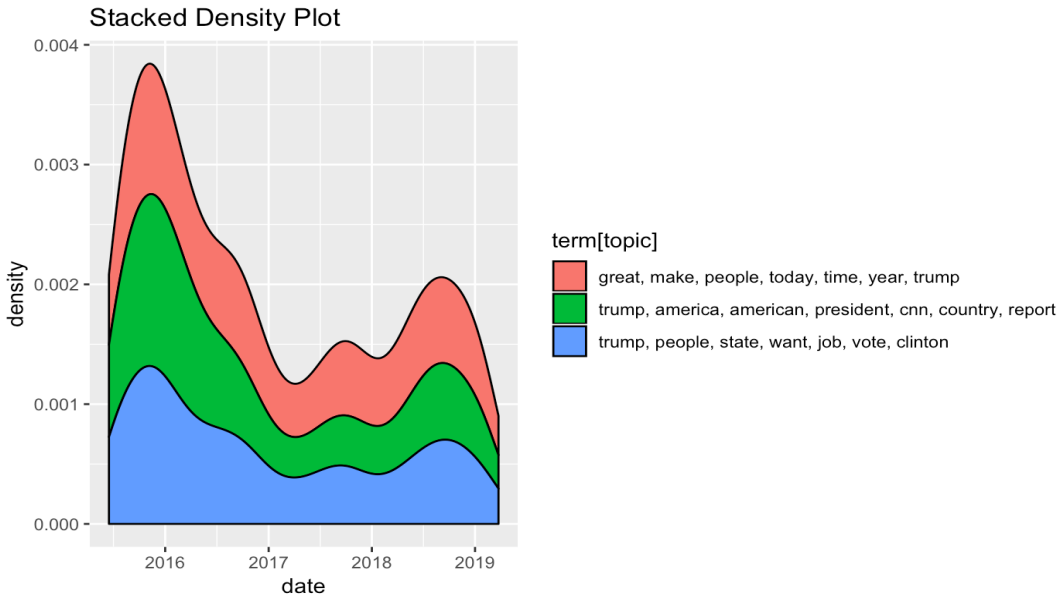

Figure 3: Stacked Density Plot

The official Twitter account owned by Donald J. Trump is @realDonalTrump, this account gets more than 59.3 million followers and over 41,000 tweets as for March 2019. Considering this account being active for political propose from Trump's official declaration of candidacy in June 2015, text data in section three and section four part B are from June 16, 2015, up to date (March 25, 2019).

B. Accessing Twitter Data

Data used in Section three are obtained from the open source website TrumpTwitterArchive. The website provides 8 attributes of Trump's twitter data including the tweet context, created time, favorite count, retweeted count, id string and if retweeted. The tweet contents after June 16, 2015, are extracted and are used in section three and section four part B.

III. Text Mining

At the time (March 25, 2019) output of this section was generated, there were 13,909 tweets in total, therefore the following results are based on these 13,909 tweets.

The text contents of the tweets are processed and analyzed with the *tm* package (Feinerer & Hornik 2018) and the *topicmodels* package (Grün & Hornik 2018). There are eight steps in total to construct a clean corpus, these steps mainly follow an example on text mining (Zhao 2015), and some improvements are done with using *qdap* package (Goodrich et al. 2019). First, built a corpus based on the content of the tweets. Second, converted all text to lower cases for later text cleaning convenience. Third, removed all URLs since Trump's tweets contained many website links. Fourth, replace contractions, words like "it's" were displaced by "it is". Fifth, removed stop words, such as "but", "would", "by" were removed in this step. In addition, words like "the", "very" and "realdonaldtrump" were not included in the function library and appeared with high frequencies, these words were removed manually. sixth, remove anything other than English letters or space, so punctuations, numbers were removed. seventh, stem and stem completion. Last, some words were not able to be completed by stem complete function automatically, for example, "peopl" were replaced manually by "people."

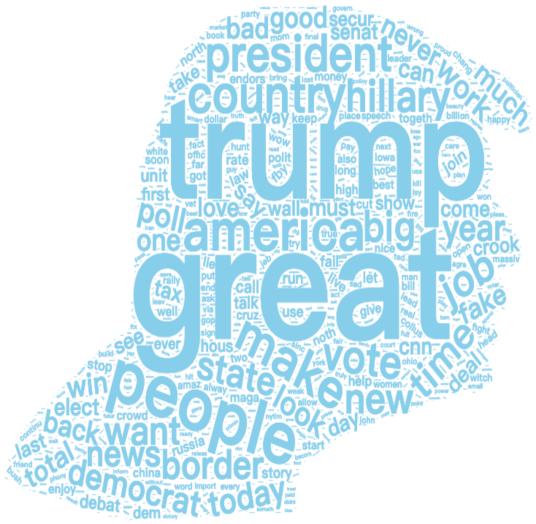

Figure 1: Word Cloud with Trump's Face

A. Frequent Terms and Term Network

A term-document matrix was generated from the clean corpus. The rows are words appeared in the tweets, and the column names are the document names. To give insight into what were the words appeared the most in Trump's tweets, a word cloud (see Figure 1) was plotted used *worldcloud2* package (Lang & Chien 2018). The shape of the word cloud is a cartoon Trump's face. The top 1 frequent word is "great". And the second highest frequent word is "trump". During Trump's campaign, he mentioned numerous times the term "Make America Great Again" which could be the reason that the word "great" is the top 1 frequent word.

Figure 2 shows the graph of term network. The graph basically tells what are some highly related words and how strong their relationships are. The thicker the line is, the stronger relationship between two words it implies. "Make", "America" and "great" are the strongest related three terms showed in the graph. Also, "sixteen" and "thousand" are highly related which means this combination appears a massive number of times. The following formula is used to compute the correlation between words.

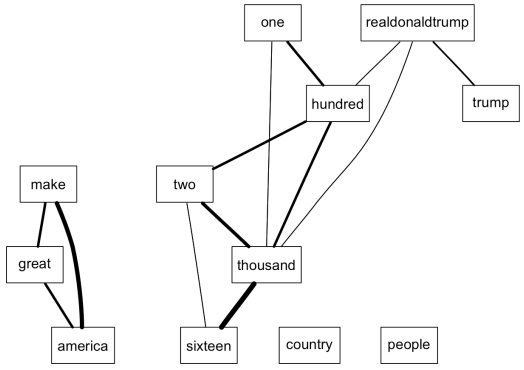

Figure 2: Term Network

In the next section, these words will be used to determine topics and further deeper analysis in order to obtain a better understanding of Trump's tweet

B. Topic Modelling

1. LDA Model

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.

In Figure 3, the stacked density plot showed that there are no clear topics in Trump's tweets. Topics result from the LDA model are quite similar. Some major words in topics are great, trump, people and realdonaldtrump. During his campaign, he mentioned numerous times of the term "Make America Great Again." That could be the reason that the word "great" appears numerous times in the model. In addition, he also mentioned himself many times. Each list of words in the model is telling the same topic about politics in Trump's tweets. Moreover, from each portion of density, we see the same trend for all three topics which also tells topics from the model are the same topic.

Topics appear the most in 2016 when he was working on his campaign and at the end of 2018. Lastly, several other pair of parameters was tried in

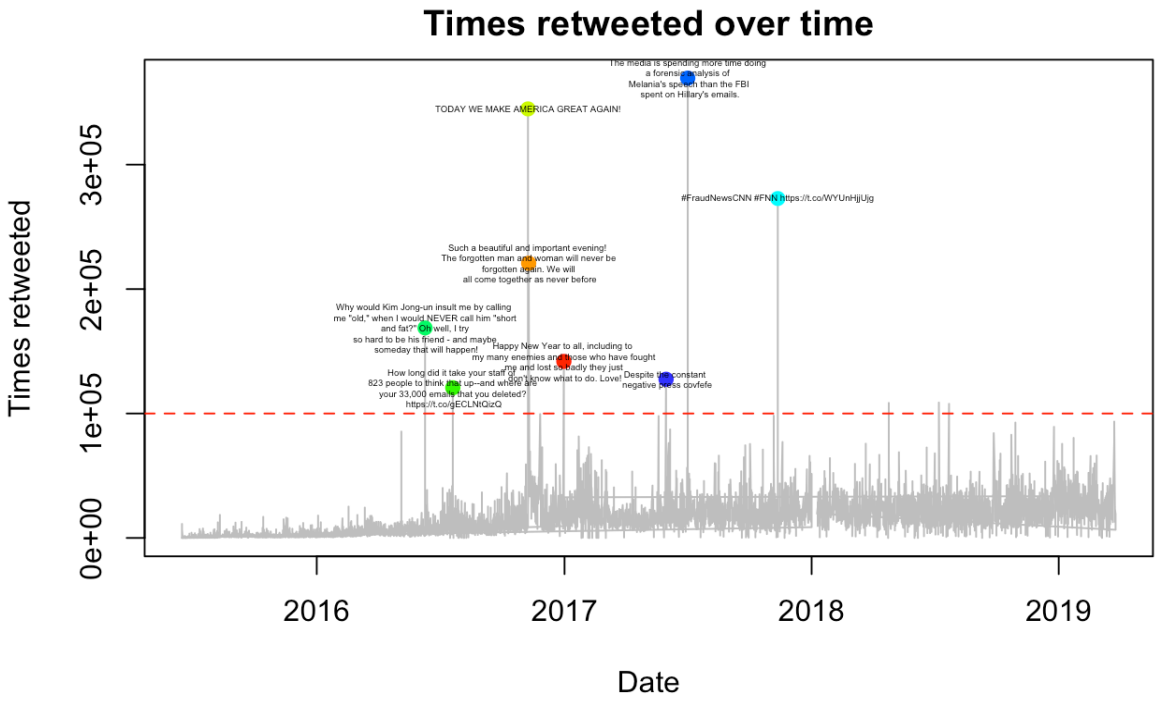

Figure 6: Times retweeted over time

building the topic model which also produce the same result.

2. Hierarchical Clustering

Another method was used to determine topics of Trump's tweets. Hierarchical clustering method (Kodali 2016) clustered words into several clusters which formed a few topics. A cluster dendrogram is shown in Figure 4. According to the dendrogram, the word "great" itself formed a single topic which is not understandable. Also, Trump's name and his Twitter ID formed as another topic. And the rest of the clusters seem like they formed in some random fashion. As a conclusion, the method of clustering did not produce comprehensible topics. This result is consistent with the result of the LDA model.

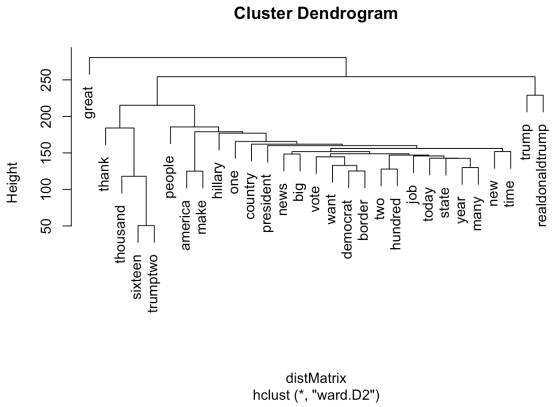

Figure 4: Dendrogram

IV. Social Network Analysis

In this section, the study switched from text mining to the social network analysis on Trump's Twitter account. This section is mainly focusing on who and how many followers are following @RealDonaldTrump as well as followers locations.

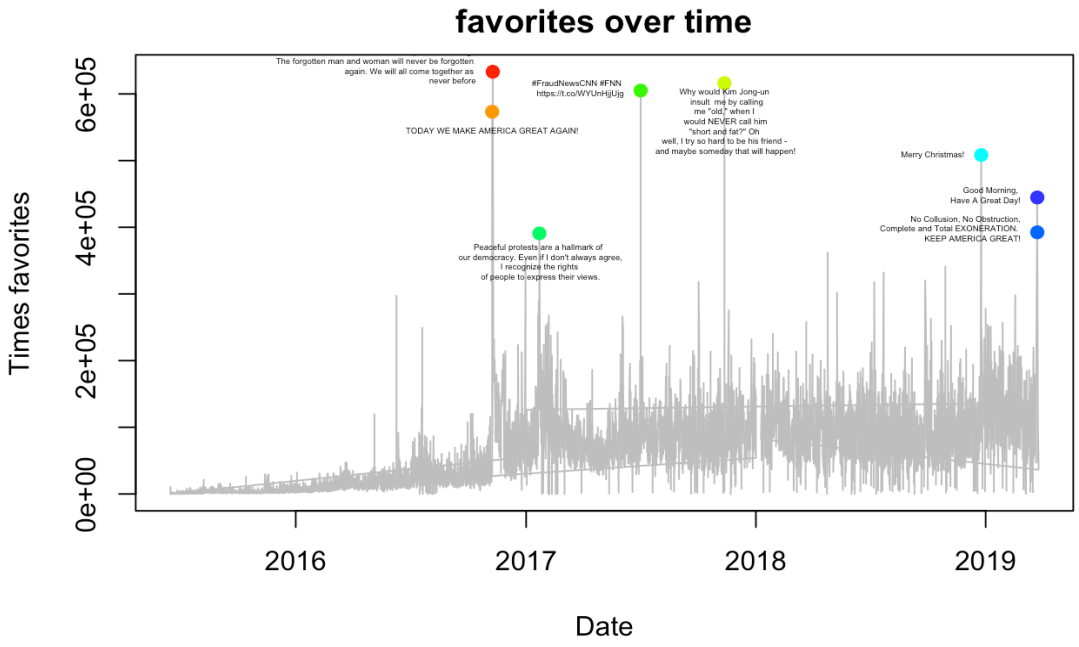

Figure 8: Favorites over time

Also, which tweets Trump tweeted have large counts of favorites and large counts of retweeting.

A. Followers' Locations Analysis

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.

Since obtaining all followers is a heavy task for our computers, 10,000 followers were sampled from the complete list of followers. Geocode function (Puente) in R was used to transform locations of followers to latitudes and longitudes. However, most of Trump's followers did not specify their locations or their locations cannot be recoded to latitudes and longitudes. Therefore, only 1254 out of 10000 followers are graphed. A Twitter Followers Map presented as figure 5 by using *ggplot* function built in R. Trump's followers is cosmopolitan. Most of the followers are from the United States. However, there are only a few followers from Russia. Table 1 shows the top 3 frequent of each country. In figure 6, most of followers are from east coast and only few of them are from west coast. In table 2, top 1 frequent following state is California, but it takes the majority of followers from west coast. Followers from east coast are fully diverse. Each state from east only takes a portion of followers, but all followers from east coast contribute a large number

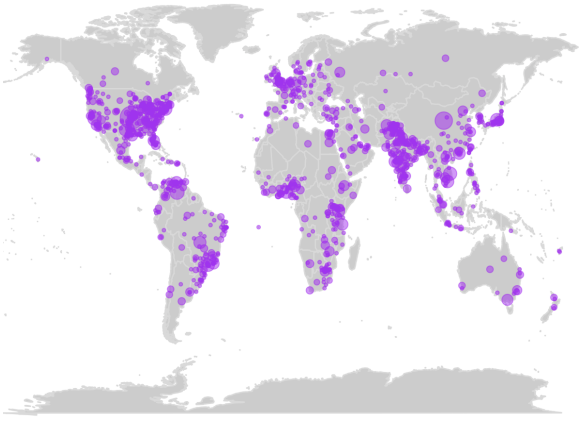

Figure 5: Twitter Followers Map

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

Table 1: Top 5 frequent countries

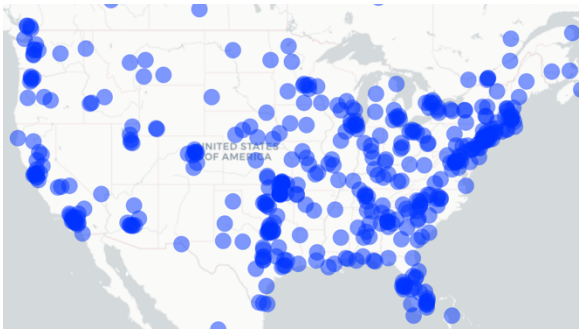

Figure 6: Twitter Followers Map (USA)

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

Table 2: Top 5 frequent states

B. Tweets: Most Favorites and Most Retweeted

After studying the aspect of Trump's followers, this section will start to analyze the most popular and influential tweets from Donald J. Trump via the retweeting and favorite features from Twitter.

Figure 6 shows 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".

Figure 7 shows tweets posted that have been clicked "favorite". Tweets with counts of favorites over 370000 are selected in the graph (forms top 8 followers' favorite tweets). The results are similar to figure 6. Most high retweeted tweets are also high in "favorite".

V. Sentiment Analysis

Trump is an extremely emotional president. By using functions in *syuzhet* (Jockers 2017) package in R, a sentiment score for each tweet can be computed. NRC is a word lexicon that is used to compute sentiment scores. The advantage of NRC is that it will return 8 different types of emotions instead of generally positive and negative emotions. Scores of each type of sentiment are greater than 0 is counted into the corresponding type of sentiment. A 0 sentiment score in a certain type means the tweet does not contain this type of sentiment. Therefore, zero is used as a bar in this section.

In figure 8, 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). There are some minor differences between the two time periods. Generally, after Trump became president, he tended to post more positive tweets.

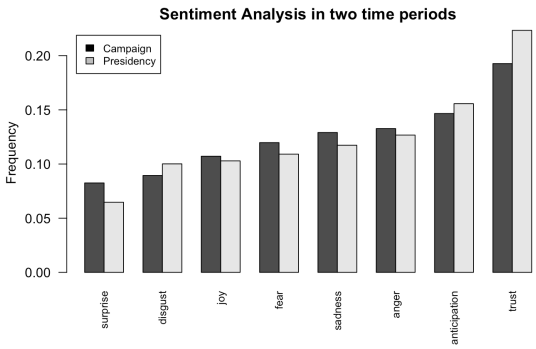

Figure 9: Sentiment scores

In addition, A study on online showed the author's speculation about Trump only use Andriod cellphone to post tweets. 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.

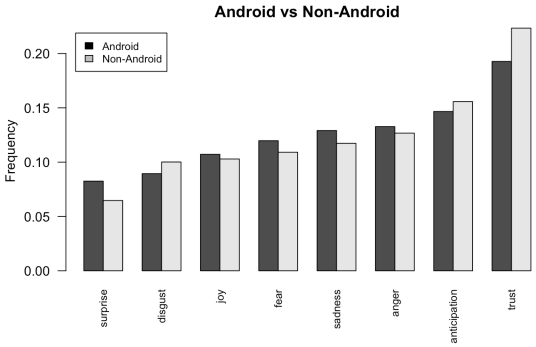

Figure 10: Android vs Non-Android

In figure 9, 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.

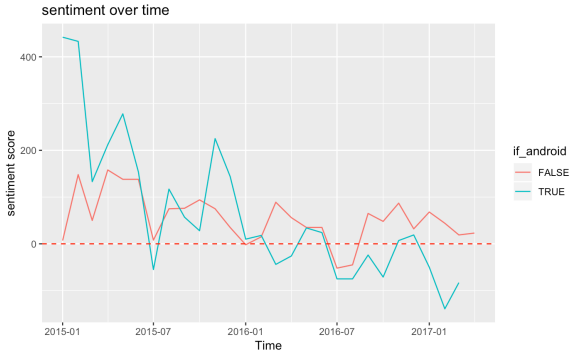

Figure 11: sentiment over time

In figure 10, 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.

VI. Conclusions and Future Work

This research is focusing on the desired interests in Trump's tweets. The study included term network about his tweets, topic distinction, followers study, and sentiment analysis. Some future work can be done to enhance the result.

This study only applied relatively simple methods for obtaining findings. Some more efficient methods can be used to gain some further insights into Trump's tweets. Also, R is a public-free programming software as well as datasets from Twitter. Some new data can be added to the study to investigate if there are any new findings.

Another future work can be done is not only to use tweets from Trump but also use tweets that talked about him from his followers. This approach can be done by using the Twitter API.

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

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