Twitter Sentiment Analysis

Sie Siong Wong - Joe Rovalino - Anil Akyildirim

11/8/2019

Contents

Load R Packages

```
# Load Requried Packages
library(tm)
library(lda)
library(httr)
library(dplyr)
library(tidyr)
library(anytime)
library(stringi)
library(twitteR)
library(syuzhet)
library(tidytext)
library(tidyverse)
```

```
library(SnowballC)
library(wordcloud)
library(topicmodels)
library(BiocManager)

# Package required for running Twitter API authorization and other R packages.
installed.packages('base64enc')
```

```
##      Package LibPath Version Priority Depends Imports LinkingTo Suggests
##      Enhances License License_is_FOSS License_restricts_use OS_type Archs
##      MD5sum NeedsCompilation Built
```

```
# BiocManager::install("Rgraphviz") *** Note that you'll need to install this "Rgraphviz" package in th
```

Introduction

Donald Trump changed the communication platform of politics from burocratic aproaches of scheduled and managed political speeches to direct communication via Twitter. He started using twitter heavily on his 2016 persidential campaign and has not looked back since. His tweets has been analyzed by variety of researchers from frequency of "angry" tweets, his emotional state during the times of his tweets, the type of tweets he sends with specific mobile devices to his tweets impact on financial markets. In this study, the main business question we are trying solve is **"Can we leverage President Trump's trade or interest rate related tweets and predict the market?"** We review the tweets between January 2018 to present, classify his tweets based on their topics and context related to trade wars, interest rate, employment in the US and conusmer spending , create a model and perform sentiment analysis.

Overall the main goal of this study is to see the classified tweets of Donald Trump, discover possible relationship with the stock market and to see how the context of text used on his account impacts the stock market. In order to do this, we will identify and describe common topics and use of text that can change the market in the corpus of the tweets that is sent from the @realDonaldTrump twitter account. We can further compare the stock market data against these tweets to see if there is any correlation and if we can create a topic model and sentiment analysis that can predict the stock market.

Data Collection

Based on the business problem in question, the content of the required data is Tweets and Stock Market Data. They are available via Twitter and Financial news platforms.

Donald Trump's Tweets

Twitter's developer account provides many API procducts including tools to extract tweets and their metadata. We will use this API to extract the wtitter data in a structured format to further wrangling and analysis. In order to use the twitter API we created a twitter account and requested developer API access. Once we received an approval, we have been provided API key and Token access information. We will be using these keys, tokens to access the API and "twitterR" to extract Donald Trump's tweets.

Connect to Twitter through API

```
# Authorization keys.
app_name <- "JAS"
consumer_key <- 'sPwbbZCtf8nfSMxhYTzqI8WHJ'
consumer_secret <- 'KfcOxgElcQ70fi3QNy8LkuDAN18dunXT147MoA8aBOLzpr3Vd3'
access_token <- '600477513-rdd3Fcywq1sfnh5S60egRQxXh0TlDqfrLzyZo4Vk'
access_secret <- 'SdDFCJUOoqAwt671VXeLaD781TdUYdeBSW2gyQMG4P5Zh'

# Extract some tweets from Twitter.
setup_twitter_oauth(consumer_key, consumer_secret, access_token, access_secret)
```

```
## [1] "Using direct authentication"
```

```
tweets <- userTimeline("realDonaldTrump", n=5)
tweets
```

```
## [[1]]
## [1] "realDonaldTrump: Why is the World Bank loaning money to China? Can this be possible? China has p
##
## [[2]]
## [1] "realDonaldTrump: Fake News @CNN is reporting that I am "still using personal cell phone for call
##
## [[3]]
## [1] "realDonaldTrump: Nadler hasn't had a single fact witness testify! Zero substance-Country wants a
##
## [[4]]
## [1] "realDonaldTrump: ...This Bill HELPS students get the student aid they need to go to college! STU
```

Upon extracting tweeter data via Twitter API and converting to dataframe, we notice that there is a limitation on the number of tweets (3200) we can extract using twitter API. This is due to our account being "Free Developer Account" and in order us to increase the tweet account, we are required to upgrade our account. Since this might become problematic and can put a damper on our analysis and future model, we think it will be better to use a service called http://www.trumptwitterarchive.com/archive that archives all Donald Trump's tweets.

Load Data

```
# President Trump tweets from 01/01/2018 to 11/21/2019.
tweets_raw <- read.csv("https://raw.githubusercontent.com/SieSiongWong/Twitter/dev/trumptweets.csv")

# S&P stock price data from year 01/04/2016 o 11/22/2019.
stocks_raw <- read.csv("https://raw.githubusercontent.com/SieSiongWong/Twitter/dev/sandp.csv")

head(tweets_raw)
```

```
##               source
## 1 Twitter for iPhone
## 2 Twitter for iPhone
## 3 Twitter for iPhone
## 4 Twitter for iPhone
## 5 Twitter for iPhone
## 6 Twitter for iPhone
##
## 1
## 2                                                                    Poll: Trump leads top 20
```

```
## 3   RT @realDonaldTrump: Impeachment Witch Hunt is now OVER! Ambassador Sondland asks U.S. President
## 4 RT @realDonaldTrump: ....\x94I WANT NOTHING! I WANT NOTHING! I WANT NO QUID PRO QUO! TELL PRESIDENT
## 5                    \x93All four of Gordon Sondland\x92s lawyers are Democrat Donors.\x94 @TuckerCa
## 6                                                                                  Watch @TuckerCarls
##        created_at retweet_count favorite_count is_retweet       id_str
## 1 11/21/2019 2:47         24221          62863      false 1.197346e+18
## 2 11/21/2019 1:22         14184          52661      false 1.197324e+18
## 3 11/21/2019 1:16         23988              0       true 1.197323e+18
## 4 11/21/2019 1:16         18754              0       true 1.197323e+18
## 5 11/21/2019 1:11         16331          60155      false 1.197322e+18
## 6 11/21/2019 1:03          9837          37564      false 1.197320e+18
```

```
head(stocks_raw)
```

```
##         Date    Open    High     Low   Close Adj.Close     Volume
## 1 2016-01-04 2038.20 2038.20 1989.68 2012.66   2012.66 4304880000
## 2 2016-01-05 2013.78 2021.94 2004.17 2016.71   2016.71 3706620000
## 3 2016-01-06 2011.71 2011.71 1979.05 1990.26   1990.26 4336660000
## 4 2016-01-07 1985.32 1985.32 1938.83 1943.09   1943.09 5076590000
## 5 2016-01-08 1945.97 1960.40 1918.46 1922.03   1922.03 4664940000
## 6 2016-01-11 1926.12 1935.65 1901.10 1923.67   1923.67 4607290000
```

Description of the variables in our Twitter data set is as follows;

- text: Content of the tweet.
- created: Date and time the tweet is created.
- Retweet: The count of retweet of the tweet.
- Favorite: The count of favorited of the tweet.

Description of the variables in our Stock Markget data set is as folows;

- Date: The date of the stock market.
- Open: The stock opening price during the trading date.
- High: The stock highest price during the trading date.
- Low: The stock lowest price during the trading date.
- Close: The stock closing price during the trading date.
- Adj. Close: The adjusted stock closing price during the trading date.
- Volume: The trading volume of stcok during the trading date.

Data Cleaning and Preparation

In this phase of the study, we will construct and clean both Stock Market and Tweets Data Set. The cleaning phase will include, updating the date class, filtering the dataset based on our analysis goal, transforming values such as percentage change in stock value, removing unwanted characters from text and selecting only the columns we need. We will further tokenize the text within tweets data set to see the word frequency and create Document Term Matrix as part of pre-processing.

Stock Data Cleaning

```
# Update Date column into date format.
stocks_raw$Date <- as.Date(stocks_raw$Date)

# Select data from 01/01/2018 to 11/20/2019 and calculate price change percentage between closing and o
stocks.df <- stocks_raw %>%
  filter(between(Date, as.Date("2018-01-01"),as.Date("2019-11-20"))) %>%
  mutate(Pct_Change=(Close-Open)/Open*100)

head(stocks.df)
```

```
##         Date    Open    High     Low   Close Adj.Close     Volume  Pct_Change
## 1 2018-01-02 2683.73 2695.89 2682.36 2695.81   2695.81 3367250000 0.450122743
## 2 2018-01-03 2697.85 2714.37 2697.77 2713.06   2713.06 3538660000 0.563780805
## 3 2018-01-04 2719.31 2729.29 2719.07 2723.99   2723.99 3695260000 0.172099941
## 4 2018-01-05 2731.33 2743.45 2727.92 2743.15   2743.15 3236620000 0.432749747
## 5 2018-01-08 2742.67 2748.51 2737.60 2747.71   2747.71 3242650000 0.183763965
## 6 2018-01-09 2751.15 2759.14 2747.86 2751.29   2751.29 3453480000 0.005093761
```

Tweets Data Cleaning

```
# Extract columns from trumptweets.csv file that are useful for analysis.
tweets_slc <- tweets_raw %>% select(source, text, created_at)

# Remove source other than iphone.
tweets_slc <- tweets_slc %>% filter(source=="Twitter for iPhone")

# Drop source column.
tweets_slc <- tweets_slc %>% select(text, created_at)

# Separate column "created_at" into "date" and "hour".
tweets_slc <- separate(data = tweets_slc, col = created_at, into  = c('date', 'hour'), sep = ' ') %>% s
```

```
## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 1 rows [8596].
```

```
# Remove minutes in hour column.
tweets_slc$hour <- gsub("\\:+\\w*","", tweets_slc$hour)

# Remove meaningless characters and symbols.
tweets_slc$text <- gsub("&amp","", tweets_slc$text)
tweets_slc$text <- gsub("(RT)((?:\\b\\w*@\\w+)+)","", tweets_slc$text)
tweets_slc$text <- gsub("^RT","", tweets_slc$text)
tweets_slc$text <- gsub("@\\w+","", tweets_slc$text)
tweets_slc$text <- gsub("[[:punct:]]","", tweets_slc$text)
tweets_slc$text <- gsub("[[:digit:]]+\\s","", tweets_slc$text)
tweets_slc$text <- gsub("http\\w+","", tweets_slc$text)
tweets_slc$text <- gsub("[ \t]{2,}"," ", tweets_slc$text)

# Remove all non-ASCII characters
tweets_slc$text <- iconv(tweets_slc$text, "UTF-8", "ASCII", sub="")

# Delete empty text column.
```

```
tweets_slc <- tweets_slc %>% na_if("") %>% na_if(" ") %>% na.omit()

# Tweets that contained less than 20 characters were treated as noise.
tweets_slc <- tweets_slc %>% filter(nchar(text)>20)

# Add id column to consider each text row as a document.
tweets_slc$doc_id <- seq.int(nrow(tweets_slc))

head(tweets_slc)
```

```
##
## 1
## 2                                                                                Impeachment Witch Hu
## 3                                                                                   I WANT NOTHING I
## 4                                                                                           All four
## 5
## 6 Today I opened a major Apple Manufacturing plant in Texas that will bring high paying jobs back to
##         date hour doc_id
## 1 11/21/2019    1      1
## 2 11/21/2019    1      2
## 3 11/21/2019    1      3
## 4 11/21/2019    1      4
## 5 11/20/2019   23      5
## 6 11/20/2019   23      6
```

Tokenizing Text and Word Frequency

```
# Tokenize the text and see frequency of words.
tweets_slc %>%
  unnest_tokens(word, text)%>%
  anti_join(stop_words) %>%
  count(word, sort=TRUE)
```

```
## Joining, by = "word"
```

```
## # A tibble: 12,794 x 2
##    word          n
##    <chr>     <int>
##  1 president  1219
##  2 people     1010
##  3 democrats   898
##  4 trump       853
##  5 country     729
##  6 news        671
##  7 border      648
##  8 fake        589
##  9 time        478
## 10 media       431
## # ... with 12,784 more rows
```

```
# We can see that words such as "president, trump" not pertaining to trade, so we remove them.
tweets_slc <- tweets_slc %>% mutate(text=tolower(text))
tweets_slc$text <- gsub("president?","", tweets_slc$text)
tweets_slc$text <- gsub("trump?","", tweets_slc$text)
```

```
# Retokenize the text and check to see if words being removed.
tweets_slc %>%
  unnest_tokens(word, text)%>%
  anti_join(stop_words) %>%
  count(word, sort=TRUE)
```

```
## Joining, by = "word"
```

```
## # A tibble: 12,744 x 2
##    word          n
##    <chr>     <int>
##  1 people     1010
##  2 democrats   898
##  3 country     729
##  4 news        671
##  5 border      648
##  6 fake        589
##  7 time        479
##  8 media       431
##  9 america     421
## 10 united      414
## # ... with 12,734 more rows
```

```
# Creating tweets frequency dataframe.
top_words <- tweets_slc %>%
  unnest_tokens(word, text)%>%
  anti_join(stop_words) %>%
  count(word, sort=TRUE)
```

```
## Joining, by = "word"
```

```
# Visualizing words which frequency are greater than 300.
top_words <- filter(top_words, n>300)
head(top_words)
```

```
## # A tibble: 6 x 2
##   word          n
##   <chr>     <int>
## 1 people     1010
## 2 democrats   898
## 3 country     729
## 4 news        671
## 5 border      648
## 6 fake        589
```

Creating Document Term Matrix

```
# Select text and id column.
tweetscorpus.df <- tweets_slc %>% select(doc_id, text)

# Create a corpus for document term matrix.
tweetscorpus <- VCorpus(DataframeSource(tweetscorpus.df))

# Remove all punctuation from the corpus.
```

```
tweetscorpus <- tm_map(tweetscorpus, removePunctuation)

# Remove all English stopwords from the corpus.
tweetscorpus <- tm_map(tweetscorpus, removeWords, stopwords("en"))
tweetscorpus <- tm_map(tweetscorpus, removeWords, stopwords("SMART"))

# Remove all number from the corpus.
tweetscorpus <- tm_map(tweetscorpus, removeNumbers)

# Strip extra white spaces in the corpus.
tweetscorpus <- tm_map(tweetscorpus, stripWhitespace)

# Stem words in the corpus.
tweetscorpus <- tm_map(tweetscorpus, stemDocument)

# Build a document term matrix.
tweetsdtm <- DocumentTermMatrix(tweetscorpus)

# Remove sparse terms which don't appear very often. Limit the document term matrix to contain terms ap
tweetsdtm <- removeSparseTerms(tweetsdtm, 0.98)

# Find the sum of words in each document and remove all docs without words.
rowTotals <- apply(tweetsdtm , 1, sum)
tweetsdtm.new   <- tweetsdtm[rowTotals> 0, ]

# Put the document in the format lda package required.
tweetsdtm.matrix <- as.matrix(tweetsdtm.new)

head(tweetsdtm.matrix, n=5)
```

```
##     Terms
## Docs administr america american back bad big billion border call campaign china
##    1         0       0        0    0   0   0       0      0    0        0     0
##    2         0       0        0    0   0   0       0      0    0        0     0
##    3         0       0        0    0   0   0       0      0    0        0     0
##    4         0       0        0    0   0   1       0      0    0        0     0
##    5         0       0        0    0   0   0       0      0    0        0     0
##     Terms
## Docs collus congratul congress continu corrupt countri crime day deal dem
##    1      0         0        0       0       0       0     0   0    0   0
##    2      0         0        0       0       0       0     0   0    0   0
##    3      0         0        0       0       0       0     0   0    0   0
##    4      0         0        0       0       0       0     0   0    0   0
##    5      0         0        0       0       0       0     0   0    0   0
##     Terms
## Docs democrat dollar dont economi elect end fact fake fbi good great happen
##    1        1      0    0       0     0   0    0    0   0    0     0      0
##    2        0      0    0       0     0   0    0    0   0    0     0      0
##    3        0      0    0       0     0   0    0    0   0    0     0      0
##    4        1      0    0       0     0   0    0    0   0    0     0      0
##    5        0      0    0       0     0   0    0    0   0    0     0      0
##     Terms
## Docs hard high hillari histori honor hous hunt illeg immigr impeach import
##    1    0    0       0       0     0    0    0     0      0       0      0
```

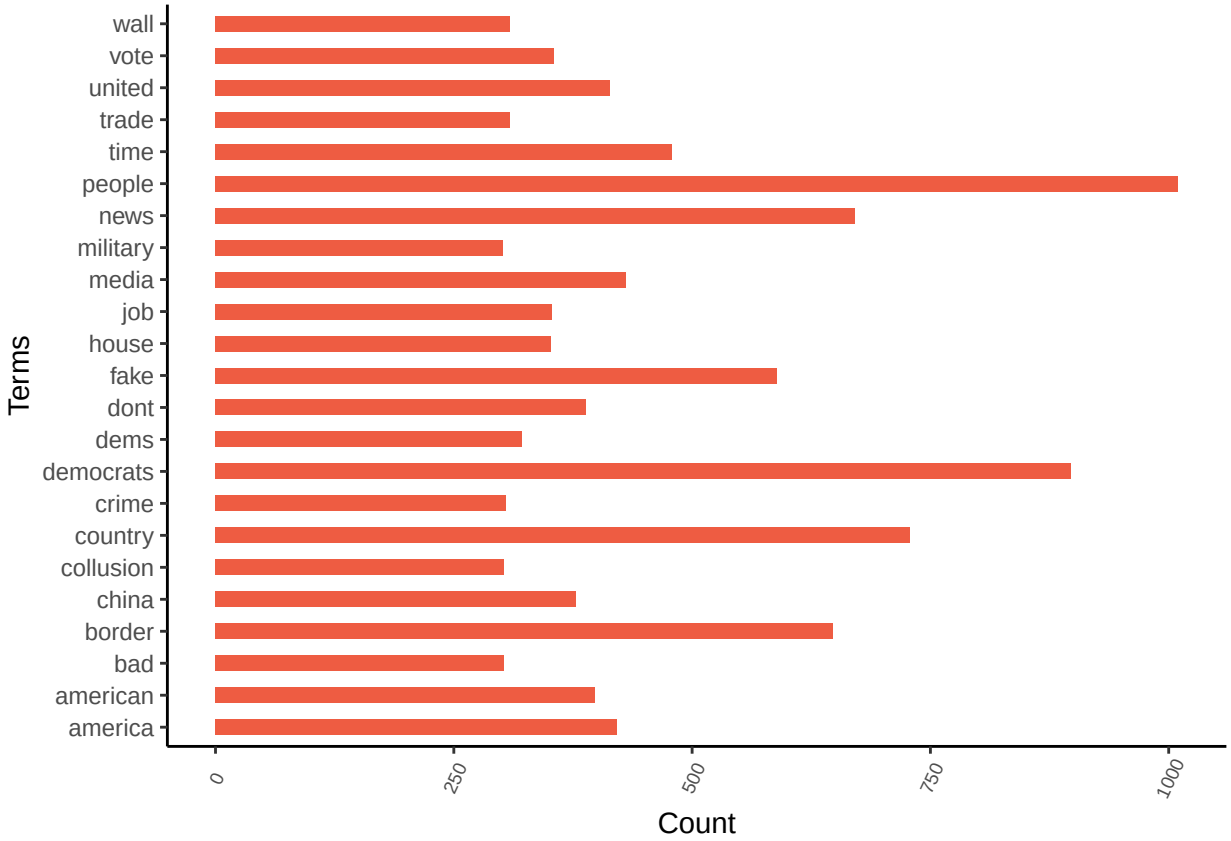
```
##    2    0    0       0       0     0    0    1     0      0       1      0
##    3    0    0       0       0     0    0    0     0      0       0      0
##    4    0    0       0       0     0    0    0     0      0       0      0
##    5    0    0       0       0     0    0    0     0      0       0      0
##     Terms
## Docs includ job law long love made make media meet militari mueller nation news
##    1      0   0   0    0    0    0    0     0    0        0       0      0    0
##    2      0   0   0    0    0    0    0     0    0        0       0      0    0
##    3      0   0   0    0    0    0    0     0    0        0       0      0    0
##    4      0   0   0    0    0    0    0     0    0        0       0      0    0
##    5      0   0   0    0    0    0    0     0    0        0       0      0    0
##     Terms
## Docs north number obama parti peopl rate record report republican russia schiff
##    1     0      0     0     0     0    0      0      0          0      0      0
##    2     0      0     0     0     0    0      0      0          0      0      0
##    3     0      0     0     0     0    0      0      0          0      0      0
##    4     0      0     0     0     0    0      0      0          1      0      0
##    5     0      0     0     0     0    0      0      0          0      0      0
##     Terms
## Docs secur senat show start state stop stori strong support talk tax thing time
##    1     0     0    0     0     0    0     0      0       0    0   0     0    0
##    2     0     0    0     0     0    0     0      0       0    0   0     0    0
##    3     0     0    0     0     0    0     0      0       0    0   0     1    0
##    4     0     0    0     0     0    0     0      0       0    0   0     0    0
##    5     0     0    0     0     0    1     0      0       0    0   0     0    0
##     Terms
## Docs today total trade unit usa vote wall watch win witch work world year
##    1     0     0     0    0   0    0    0     0   0     0    0     0    0
##    2     0     0     0    0   0    0    0     0   0     1    0     0    0
##    3     0     0     0    0   0    0    0     0   0     0    0     0    0
##    4     1     0     0    0   0    0    0     0   1     0    0     0    0
##    5     0     0     0    0   0    0    0     0   0     0    0     0    0
```

Data Exploration

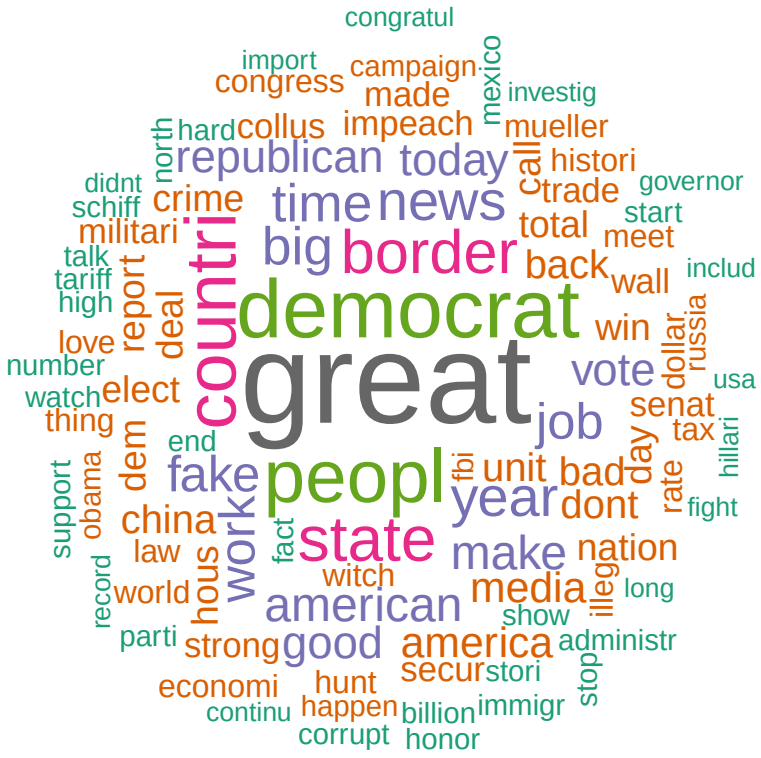
In order to define our analytical approach we would like to understand the data gained, review initial insights about our data and make sure we do not require additional data in order to find the answer of our problem in question.

We can initially take a look at the top words within the tweets.

```
# Visualization of top words within the complete tweets data.
theme_set(theme_classic())

ggplot(top_words, aes(x=word, y=n))+
  geom_bar(stat="identity", width = 0.5, fill="tomato2")+
  xlab("Terms") + ylab("Count") + coord_flip() +
  theme(axis.text.x = element_text(angle=65, vjust=0.6, size=7))
```




```
# Visualizing wordcloud.
wordcloud(tweetscorpus, max.words = 100, random.order = FALSE, rot.per = 0.15, min.freq = 5, colors = bi
```



There are some interesting finds here such as the top two words used within the tweets are "people" and "democrats". Great is another word that is commonly used. However none of this top words analysis is very helpful to reach our business objective as they are not related to "Trade". To be more specific, we can take a look at words individually and review their relationship between them.

```
# Which words are associated with 'trade'?
findAssocs(tweetsdtm.new, "trade", 0.05)
```

```
## $trade
##    deal billion   china countri  dollar    year    unit    talk    good     usa
##    0.25    0.20    0.19    0.13    0.13    0.12    0.10    0.08    0.06    0.06
##    long    meet
##    0.05    0.05
```

```
# Which words are associated with 'china'?
findAssocs(tweetsdtm.new, "china", 0.05)
```

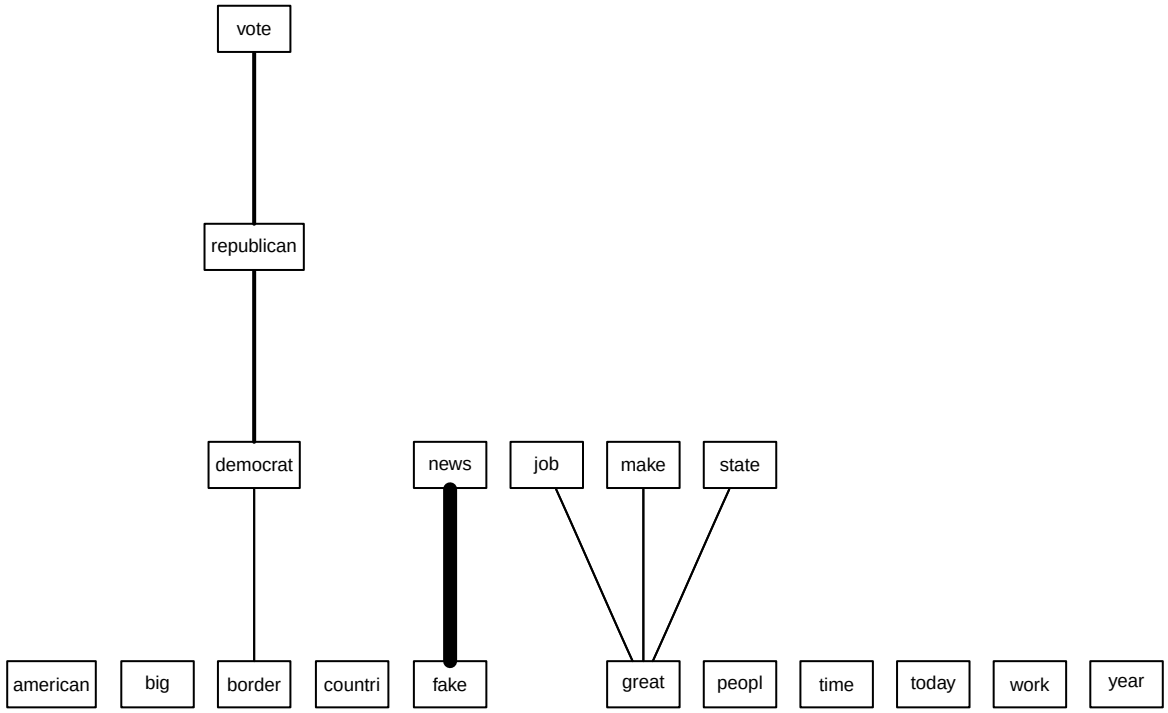
```
## $china
## billion    deal   trade  dollar continu     usa    good    meet    make    year
##    0.19    0.19    0.19    0.16    0.13    0.11    0.08    0.08    0.07    0.07
##    unit   start
##    0.06    0.05
```

```
# Which words are associated with 'job'?
findAssocs(tweetsdtm.new, "job", 0.05)
```

```
## $job
##    great militari  economi   record   number      tax
##     0.13     0.09     0.07     0.07     0.06     0.06
```

We can see “trade” has associations with multiple words such as deal, billion and china and text “job” has associations with great, militari and economi.

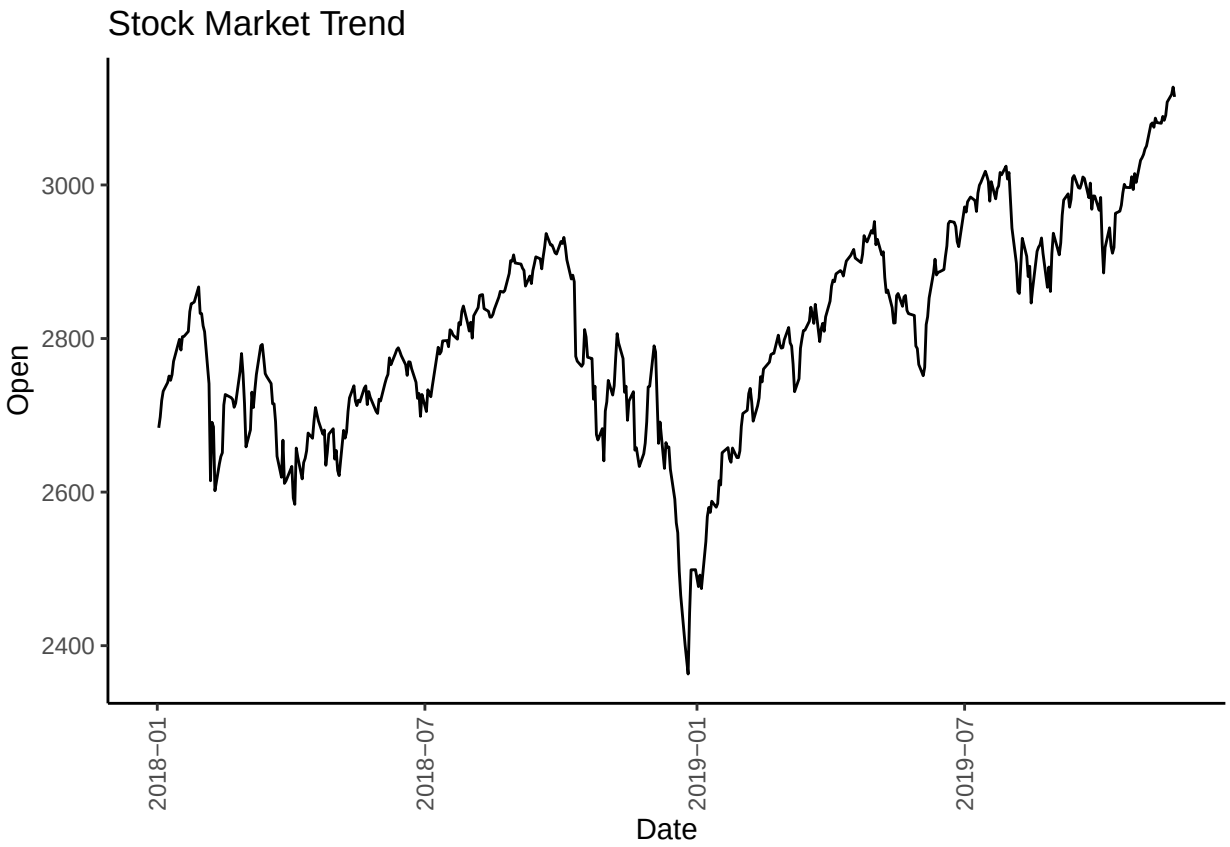
```
freq_terms <- findFreqTerms(tweetsdtm.new, lowfreq = 500)

# Visualizing the association.
plot(tweetsdtm.new, term = freq_terms, corThreshold = 0.10, weighting = T)
```



We can also see the association between words such as “news” and “fake”, “great”, “jobs” and “state” are commonly used together.

We should also look at how stock market has been trending within our target date range.

```
ggplot(stocks.df, aes(x=Date))+
  geom_line(aes(y=Open))+
  labs(title = "Stock Market Trend")+
  theme(axis.text.x = element_text(angle=90, vjust=0.5),
        panel.grid.minor = element_blank())
```



We can see that starting from 2019-01, the stock market is trending upwards.

Model Development

Based on our business objective and the data we have prepared, we decided to proceed with topic modeling as our analytical approach for model development. The idea is for us to identify topics as set of documents, select the right topic and create a final stock market dataframe for prediction. In terms of topic modeling, we have selected Latent Dirichlet Allocation (LDA).

LDA Model

LDA is an unsupervised learning that views the documents as bag of words. In each topic that is generated, picks a set of words against it. Below outlines the each step the LDA does;

- Assume there are k topics across all the documents.
- Distribute these topics across a dopcument by assigning each word a topic.
- For each word in the document, assume its topic is wrong but every other word is assigned the topic is correct.
- Assign a word for each topic based on what topics are in the document and how many times a word has been assigned to a particular topic accross all of the documents.
- Repeat this process a number of times for each document.

Topics Modeling

After running the LDA model few times, we found that using 30 topics will produce better result of topic classifying.

```
# Create a LDA model with Gibbs method for 30 topics.
tweetsLDA <- LDA(tweetsdtm.matrix, 30, method="Gibbs", control = list(seed = 123))

# Top 30 words per topic.
terms(tweetsLDA, 30)
```

```
##       Topic 1     Topic 2    Topic 3     Topic 4     Topic 5      Topic 6
##  [1,] "work"      "democrat" "russia"    "big"       "today"      "back"
##  [2,] "hard"      "dont"     "campaign"  "win"       "nation"     "great"
##  [3,] "great"     "bad"      "fbi"       "congratul" "honor"      "day"
##  [4,] "peopl"     "fact"     "hillari"   "great"     "great"      "bad"
##  [5,] "continu"   "happen"   "collus"    "includ"    "record"     "includ"
##  [6,] "number"    "total"    "elect"     "dem"       "law"        "happen"
##  [7,] "dem"       "watch"    "report"    "dollar"    "big"        "hous"
##  [8,] "start"     "media"    "total"     "american"  "congress"   "number"
##  [9,] "parti"     "militari" "work"      "world"     "dont"       "collus"
## [10,] "high"      "today"    "watch"     "campaign"  "talk"       "america"
## [11,] "countri"   "world"    "stop"      "hard"      "republican" "long"
## [12,] "today"     "hunt"     "end"       "strong"    "border"     "fact"
## [13,] "fbi"       "corrupt"  "show"      "show"      "corrupt"    "histori"
## [14,] "call"      "fbi"      "law"       "vote"      "import"     "job"
## [15,] "histori"   "hillari"  "state"     "year"      "american"   "deal"
## [16,] "stop"      "time"     "big"       "collus"    "china"      "elect"
## [17,] "thing"     "china"    "hunt"      "happen"    "dem"        "fbi"
## [18,] "administr" "strong"   "made"      "honor"     "impeach"    "impeach"
## [19,] "back"      "talk"     "rate"      "meet"      "militari"   "make"
## [20,] "day"       "vote"     "unit"      "russia"    "obama"      "obama"
## [21,] "fake"      "american" "administr" "state"     "senat"      "schiff"
## [22,] "rate"      "collus"   "america"   "thing"     "stop"       "strong"
## [23,] "schiff"    "economi"  "american"  "today"     "strong"     "wall"
## [24,] "america"   "end"      "back"      "administr" "unit"       "watch"
## [25,] "american"  "number"   "bad"       "america"   "administr"  "administr"
## [26,] "collus"    "obama"    "billion"   "back"      "america"    "american"
## [27,] "deal"      "peopl"    "border"    "bad"       "back"       "big"
## [28,] "dont"      "senat"    "call"      "billion"   "bad"        "billion"
## [29,] "hillari"   "stop"     "china"     "border"    "billion"    "border"
## [30,] "illeg"     "unit"     "congratul" "call"      "call"       "call"
##       Topic 7    Topic 8     Topic 9     Topic 10    Topic 11    Topic 12
##  [1,] "economi"  "hous"      "year"      "meet"      "state"     "crime"
##  [2,] "tax"      "senat"     "obama"     "north"     "unit"      "strong"
##  [3,] "high"     "great"     "administr" "talk"      "dont"      "militari"
##  [4,] "number"   "includ"    "american"  "import"    "continu"   "border"
##  [5,] "record"   "record"    "crime"     "continu"   "talk"      "love"
##  [6,] "countri"  "big"       "support"   "end"       "watch"     "peopl"
##  [7,] "crime"    "end"       "militari"  "long"      "news"      "tax"
##  [8,] "parti"    "impeach"   "day"       "show"      "great"     "long"
##  [9,] "big"      "stop"      "corrupt"   "state"     "nation"    "total"
## [10,] "great"    "illeg"     "vote"      "good"      "congress"  "world"
## [11,] "dem"      "congress"  "senat"     "call"      "obama"     "start"
## [12,] "america"  "dont"      "economi"   "russia"    "strong"    "wall"
```

```
## [13,] "militari" "hillari"    "made"      "stop"       "american"  "countri"
## [14,] "talk"     "bad"        "north"     "fbi"        "day"       "elect"
## [15,] "vote"     "countri"    "fake"      "happen"     "long"      "good"
## [16,] "watch"    "hard"       "long"      "hard"       "mueller"   "win"
## [17,] "collus"   "love"       "stori"     "usa"        "start"     "american"
## [18,] "day"      "win"        "big"       "love"       "support"   "call"
## [19,] "long"     "parti"      "elect"     "back"       "administr" "dollar"
## [20,] "obama"    "republican" "hous"      "border"     "america"   "economi"
## [21,] "peopl"    "support"    "america"   "dont"       "back"      "fact"
## [22,] "state"    "year"       "back"      "hunt"       "bad"       "fake"
## [23,] "hard"     "administr"  "bad"       "republican" "big"       "high"
## [24,] "histori"  "america"    "billion"   "support"    "billion"   "honor"
## [25,] "illeg"    "american"   "border"    "administr"  "border"    "hunt"
## [26,] "news"     "back"       "call"      "america"    "call"      "made"
## [27,] "report"   "billion"    "campaign"  "american"   "campaign"  "administr"
## [28,] "today"    "border"     "china"     "bad"        "china"     "america"
## [29,] "usa"      "call"       "collus"    "big"        "collus"    "back"
## [30,] "wall"     "campaign"   "congratul" "billion"    "congratul" "bad"
##       Topic 13    Topic 14    Topic 15    Topic 16    Topic 17    Topic 18
##  [1,] "american"  "time"      "stori"     "great"     "good"      "rate"
##  [2,] "peopl"     "long"      "media"     "total"     "thing"     "elect"
##  [3,] "histori"   "start"     "news"      "support"   "happen"    "bad"
##  [4,] "made"      "countri"   "corrupt"   "love"      "dem"       "show"
##  [5,] "fact"      "year"      "fake"      "show"      "usa"       "watch"
##  [6,] "great"     "bad"       "fact"      "strong"    "north"     "great"
##  [7,] "day"       "deal"      "bad"       "big"       "time"      "call"
##  [8,] "dem"       "work"      "state"     "fbi"       "world"     "dont"
##  [9,] "report"    "make"      "report"    "win"       "great"     "end"
## [10,] "stop"      "record"    "total"     "corrupt"   "work"      "make"
## [11,] "includ"    "trade"     "time"      "histori"   "big"       "militari"
## [12,] "happen"    "includ"    "job"       "work"      "obama"     "time"
## [13,] "work"      "secur"     "rate"      "meet"      "administr" "year"
## [14,] "china"     "end"       "big"       "happen"    "includ"    "collus"
## [15,] "dont"      "campaign"  "dont"      "make"      "meet"      "republican"
## [16,] "good"      "stop"      "american"  "today"     "talk"      "talk"
## [17,] "hous"      "high"      "continu"   "american"  "fact"      "long"
## [18,] "news"      "total"     "countri"   "border"    "china"     "nation"
## [19,] "republican" "russia"   "end"       "call"      "corrupt"   "usa"
## [20,] "schiff"    "congress"  "fbi"       "hunt"      "elect"     "crime"
## [21,] "show"      "histori"   "law"       "media"     "state"     "start"
## [22,] "strong"    "import"    "made"      "news"      "strong"    "administr"
## [23,] "today"     "media"     "mueller"   "number"    "democrat"  "america"
## [24,] "unit"      "rate"      "record"    "administr" "economi"   "american"
## [25,] "administr" "usa"       "start"     "america"   "fake"      "back"
## [26,] "america"   "congratul" "stop"      "back"      "news"      "big"
## [27,] "back"      "crime"     "administr" "bad"       "report"    "billion"
## [28,] "bad"       "day"       "america"   "billion"   "america"   "border"
## [29,] "big"       "meet"      "back"      "campaign"  "american"  "campaign"
## [30,] "billion"   "north"     "billion"   "china"     "back"      "china"
##       Topic 19    Topic 20    Topic 21    Topic 22    Topic 23    Topic 24
##  [1,] "job"       "democrat"  "call"      "make"      "border"    "vote"
##  [2,] "great"     "impeach"   "made"      "america"   "wall"      "republican"
##  [3,] "world"     "dem"       "congress"  "deal"      "secur"     "parti"
##  [4,] "back"      "schiff"    "schiff"    "back"      "vote"      "big"
```

```
##  [5,] "deal"      "fact"      "end"       "good"      "schiff"     "elect"
##  [6,] "campaign"  "report"    "deal"      "includ"    "hard"       "countri"
##  [7,] "thing"     "world"     "happen"    "continu"   "world"      "crime"
##  [8,] "high"      "start"     "corrupt"   "high"      "deal"       "impeach"
##  [9,] "senat"     "end"       "year"      "start"     "high"       "bad"
## [10,] "dollar"    "crime"     "today"     "bad"       "make"       "import"
## [11,] "happen"    "hard"      "import"    "dont"      "nation"     "media"
## [12,] "obama"     "administr" "high"      "job"       "republican" "russia"
## [13,] "america"   "dollar"    "support"   "year"      "today"      "trade"
## [14,] "big"       "love"      "campaign"  "great"     "administr"  "end"
## [15,] "call"      "meet"      "collus"    "meet"      "america"    "stori"
## [16,] "economi"   "stop"      "countri"   "show"      "american"   "talk"
## [17,] "includ"    "congress"  "economi"   "trade"     "back"       "back"
## [18,] "long"      "militari"  "illeg"     "unit"      "bad"        "democrat"
## [19,] "parti"     "number"    "immigr"    "administr" "big"        "happen"
## [20,] "state"     "witch"     "senat"     "american"  "billion"    "hous"
## [21,] "strong"    "america"   "trade"     "big"       "call"       "illeg"
## [22,] "today"     "american"  "watch"     "billion"   "campaign"   "meet"
## [23,] "unit"      "back"      "administr" "border"    "china"      "administr"
## [24,] "administr" "bad"       "america"   "call"      "collus"     "america"
## [25,] "american"  "big"       "american"  "campaign"  "congratul"  "american"
## [26,] "bad"       "billion"   "back"      "china"     "congress"   "billion"
## [27,] "billion"   "border"    "bad"       "collus"    "continu"    "border"
## [28,] "border"    "call"      "big"       "congratul" "corrupt"    "call"
## [29,] "china"     "campaign"  "billion"   "congress"  "countri"    "campaign"
## [30,] "collus"    "china"     "border"    "corrupt"   "crime"      "china"
##       Topic 25     Topic 26    Topic 27     Topic 28     Topic 29
##  [1,] "countri"    "china"     "law"        "peopl"      "witch"
##  [2,] "histori"    "trade"     "illeg"      "countri"    "hunt"
##  [3,] "world"      "dollar"    "immigr"     "great"      "mueller"
##  [4,] "usa"        "billion"   "stop"       "start"      "report"
##  [5,] "trade"      "deal"      "democrat"   "day"        "collus"
##  [6,] "make"       "continu"   "end"        "american"   "media"
##  [7,] "democrat"   "peopl"     "work"       "dem"        "today"
##  [8,] "includ"     "usa"       "continu"    "end"        "corrupt"
##  [9,] "republican" "job"       "deal"       "made"       "total"
## [10,] "end"        "fact"      "peopl"      "import"     "china"
## [11,] "back"       "good"      "total"      "thing"      "support"
## [12,] "day"        "happen"    "dollar"     "total"      "work"
## [13,] "peopl"      "senat"     "campaign"   "watch"      "administr"
## [14,] "thing"      "stop"      "republican" "stop"       "call"
## [15,] "long"       "today"     "corrupt"    "talk"       "crime"
## [16,] "total"      "administr" "import"     "administr"  "happen"
## [17,] "news"       "america"   "media"      "fact"       "import"
## [18,] "parti"      "american"  "support"    "fbi"        "obama"
## [19,] "state"      "back"      "american"   "good"       "secur"
## [20,] "american"   "bad"       "call"       "republican" "america"
## [21,] "border"     "big"       "china"      "russia"     "american"
## [22,] "congratul"  "border"    "hillari"    "unit"       "back"
## [23,] "hous"       "call"      "report"     "border"     "bad"
## [24,] "import"     "campaign"  "secur"      "fake"       "big"
## [25,] "meet"       "collus"    "strong"     "news"       "billion"
## [26,] "report"     "congratul" "trade"      "number"     "border"
## [27,] "administr"  "congress"  "witch"      "tax"        "campaign"
```

```
## [28,] "america"    "corrupt"   "administr"  "witch"      "congratul"
## [29,] "bad"        "countri"   "america"    "work"       "congress"
## [30,] "big"        "crime"     "back"       "america"    "continu"
##       Topic 30
##  [1,] "news"
##  [2,] "fake"
##  [3,] "media"
##  [4,] "fact"
##  [5,] "militari"
##  [6,] "report"
##  [7,] "nation"
##  [8,] "world"
##  [9,] "strong"
## [10,] "wall"
## [11,] "border"
## [12,] "call"
## [13,] "dollar"
## [14,] "hous"
## [15,] "talk"
## [16,] "trade"
## [17,] "administr"
## [18,] "america"
## [19,] "american"
## [20,] "back"
## [21,] "bad"
## [22,] "big"
## [23,] "billion"
## [24,] "campaign"
## [25,] "china"
## [26,] "collus"
## [27,] "congratul"
## [28,] "congress"
## [29,] "continu"
## [30,] "corrupt"
```

Per-Document Classification

```
# Per-topic-per-word probabilities.
tweetsLDA.topicword.prob <- tidy(tweetsLDA, matrix="beta")
head(tweetsLDA.topicword.prob)
```

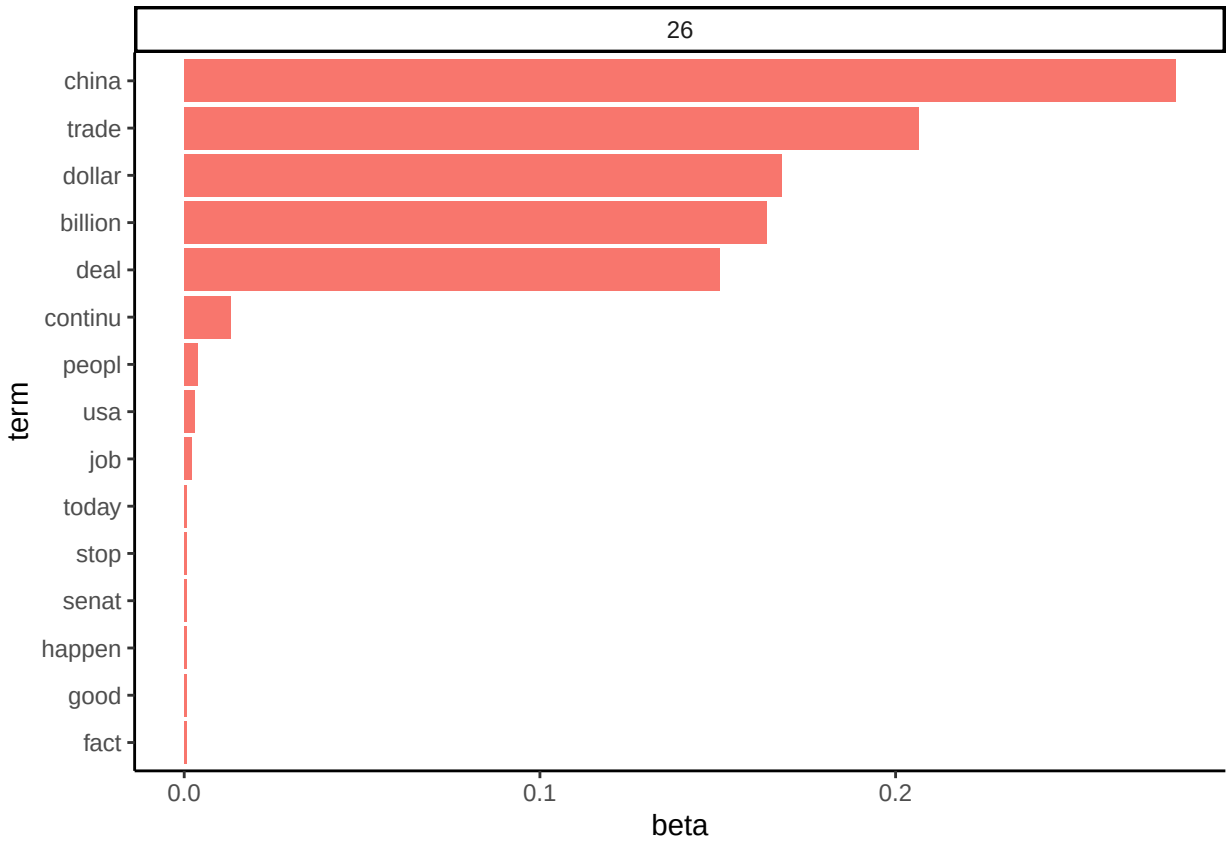
```
## # A tibble: 6 x 3
##   topic term           beta
##   <int> <chr>         <dbl>
## 1     1 administr 0.00184
## 2     2 administr 0.0000909
## 3     3 administr 0.0000798
## 4     4 administr 0.0000814
## 5     5 administr 0.0000876
## 6     6 administr 0.0000938
```

```
# Find the 10 terms that are most common within each topic.
tweetsLDA.topterms <- tweetsLDA.topicword.prob %>%
  group_by(topic) %>%
```

```
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

head(tweetsLDA.topterms)
```

```
## # A tibble: 6 x 3
##   topic term      beta
##   <int> <chr>    <dbl>
## 1     1 work    0.447
## 2     1 hard    0.182
## 3     1 great   0.0780
## 4     1 peopl   0.0622
## 5     1 continu 0.0509
## 6     1 number  0.0465
```

```
# Plot per-topic-per-word probabilities for topic #26.
tweetsLDA.topterms %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  filter(topic==26) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  scale_x_reordered()
```



```
# Classify the selected topic #26 per document.
tweetsLDA.class <- data.frame(topics(tweetsLDA))
tweetsLDA.class <- cbind(tweetsLDA.class, 1:nrow(tweetsLDA.class))
colnames(tweetsLDA.class)[ncol(tweetsLDA.class)] <-'doc_id'
tweetsLDA.class <- tweetsLDA.class %>% filter(topics.tweetsLDA.==26)

head(tweetsLDA.class)
```

```
##   topics.tweetsLDA. doc_id
## 1                26     61
## 2                26     62
## 3                26    512
## 4                26    719
## 5                26    786
## 6                26    865
```

```
# Inner join selected classified topic with original dataframe.
tweets.final <- inner_join(tweetsLDA.class, tweets_slc)
```

```
## Joining, by = "doc_id"
```

```
head(tweets.final)
```

```
##   topics.tweetsLDA. doc_id
## 1                26     61
## 2                26     62
## 3                26    512
## 4                26    719
## 5                26    786
## 6                26    865
##
## 1
## 2
## 3
## 4 general michael flynns attorney is demanding that charges be immediately dropped after they found
## 5
## 6                   doral in miami would have been the best place to hold the gand free but too much
##         date hour
## 1 11/19/2019   18
## 2 11/19/2019   18
## 3  11/2/2019   21
## 4 10/26/2019   11
## 5 10/23/2019   17
## 6 10/21/2019   13
```

Based on the probability per topic, per word , we can see that “china”, “trade”, “dollar”, “billion” and “deal” has the highest probability in the topic #26 we chose. These words we consider have highly relevant to the trade topic we’re focusing on. Therefore, we’re able to reduce the cleaned original 9,171 tweets to 253 tweets. We’ll use these 253 identified trade-related tweets for sentiment analysis.

Sentiment Analysis

In the sentiment analysis, each tweets will get an emotion score. The ‘Syuzhet’ package breaks the emotion into 10 different emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, trust, negative and positive.

Each tweet will be evaluated by these 10 emotions and then assigned a sum score.

```
# Turn tweets text into vector.
tweets.df <- as.vector(tweets.final$text)

# Getting emotion score for each tweet.
tweets.emotion <- get_nrc_sentiment(tweets.df)
tweets.emotion <- cbind(tweets.final, tweets.emotion)
head(tweets.emotion)
```

```
##   topics.tweetsLDA. doc_id
## 1                26     61
## 2                26     62
## 3                26    512
## 4                26    719
## 5                26    786
## 6                26    865
##
## 1
## 2
## 3
## 4 general michael flynns attorney is demanding that charges be immediately dropped after they found
## 5
## 6                  doral in miami would have been the best place to hold the gand free but too much
##         date hour anger anticipation disgust fear joy sadness surprise trust
## 1 11/19/2019   18     1            1       0    0   0       0        0     0
## 2 11/19/2019   18     0            0       0    0   0       0        0     1
## 3  11/2/2019   21     0            0       0    0   0       0        0     0
## 4 10/26/2019   11     2            1       0    3   1       2        1     4
## 5 10/23/2019   17     1            0       1    0   0       0        0     2
## 6 10/21/2019   13     0            1       0    0   1       0        1     1
##   negative positive
## 1        1        0
## 2        0        0
## 3        0        0
## 4        4        5
## 5        2        0
## 6        1        2
```

```
# Getting sentiment score for each tweet.
tweets.score <- get_sentiment(tweets.df)
tweets.score <- cbind(tweets.final,tweets.score )
head(tweets.score)
```

```
##   topics.tweetsLDA. doc_id
## 1                26     61
## 2                26     62
## 3                26    512
## 4                26    719
## 5                26    786
## 6                26    865
##
## 1
## 2
## 3
## 4 general michael flynns attorney is demanding that charges be immediately dropped after they found
```

```
## 5
## 6                  doral in miami would have been the best place to hold the gand free but too much l
##         date hour tweets.score
## 1 11/19/2019   18        -1.00
## 2 11/19/2019   18         0.00
## 3  11/2/2019   21         0.00
## 4 10/26/2019   11        -1.50
## 5 10/23/2019   17        -1.55
## 6 10/21/2019   13         2.10
```

We have defined the topics in sets of documents using LDA topics modeling, we have also assigned a tweet score with our sentiment analysis. Our next step is to map the sentiment scores against the stock price change.

Sentiment Scores vs Stock Price Change

In order to map the sentiment scores, we first group the date and sum the sentiment scores into single day and then merge with stocks dataframe.

```
# Update column name.
colnames(tweets.score)[4]<-"Date"

# Aggregate scores into single day.
tweets.score.sum <- tweets.score %>%
  select(Date, tweets.score) %>%
  group_by(Date) %>%
  summarise(scores=sum(tweets.score))

# Update date column into date format.
tweets.score.sum$Date <- anydate(tweets.score.sum$Date)

# Merge stocks dataframe and scores dataframe.
stocks.df.new <-  stocks.df %>% select(Date, Pct_Change)
stocks.scores <- merge(stocks.df.new,tweets.score.sum, by='Date')

head(stocks.scores)
```

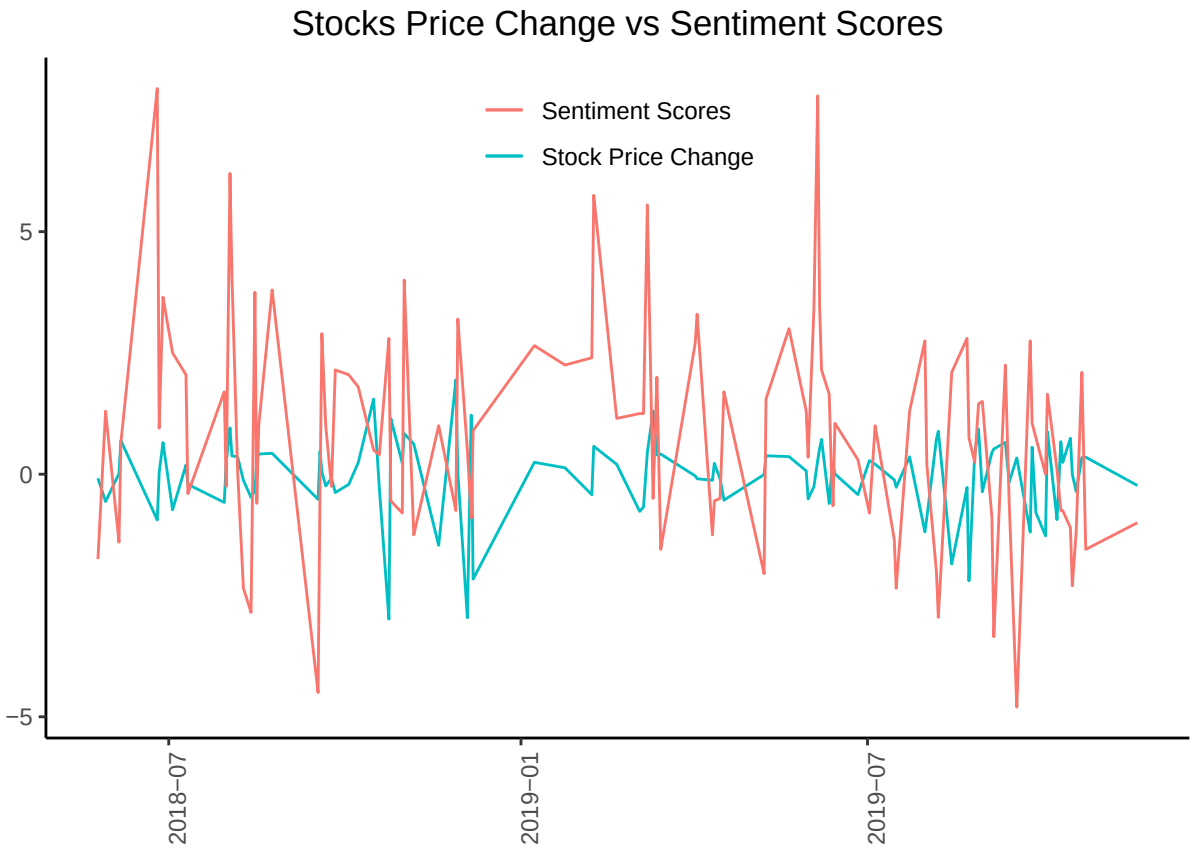
```
##         Date  Pct_Change scores
## 1 2018-05-25 -0.08334630  -1.75
## 2 2018-05-29 -0.56374785   1.30
## 3 2018-06-05  0.01237377  -1.40
## 4 2018-06-06  0.69372916   0.65
## 5 2018-06-25 -0.94314398   7.95
## 6 2018-06-26  0.03452978   0.95
```

When we look at our combined stocks and scores dataframe, we are able to see the percentage change of stock market for a given date and its sentiment score.

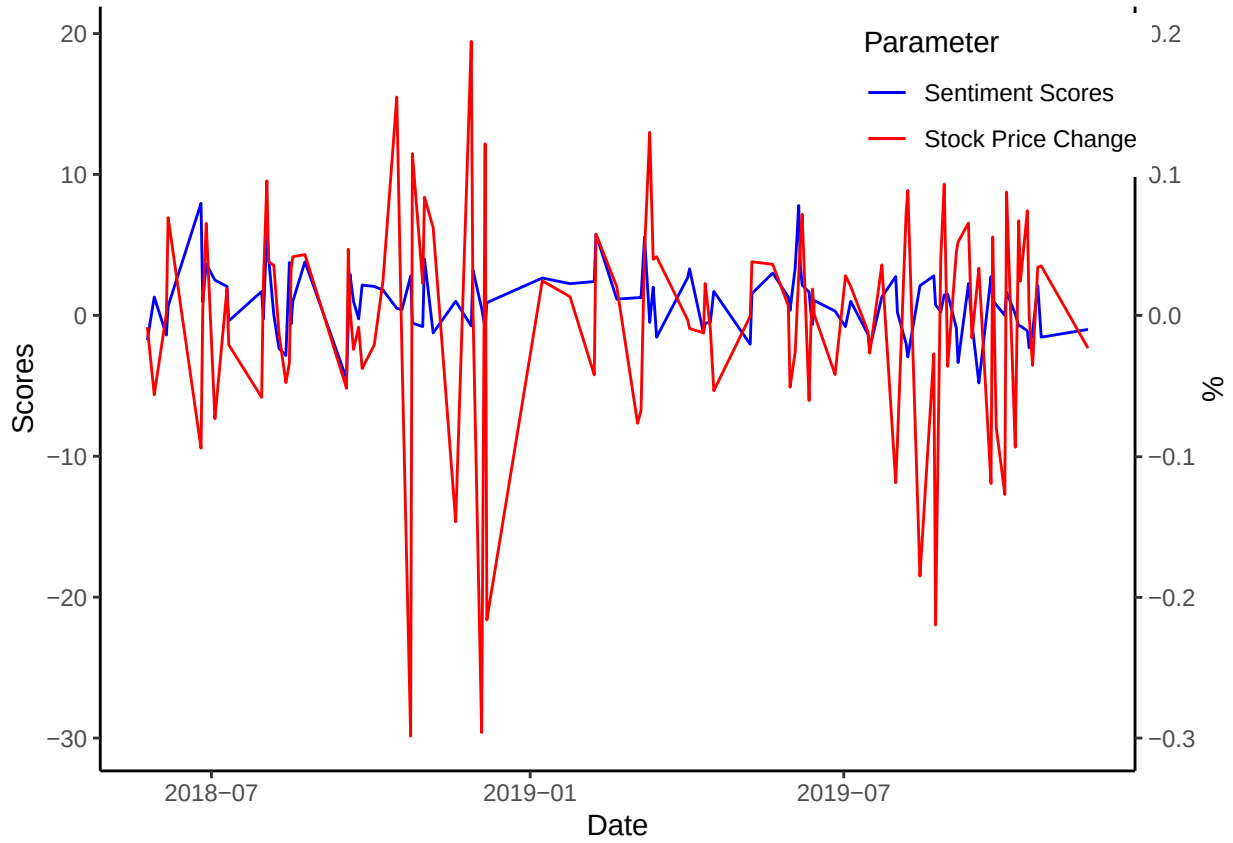
Visualization

```
## Compare stocks price percentage change with sentiment score.
```

```
# Two variables on same y-axis.
ggplot(stocks.scores, aes(Date)) + ggtitle("Stocks Price Change vs Sentiment Scores") + ylab("") +  geo
```

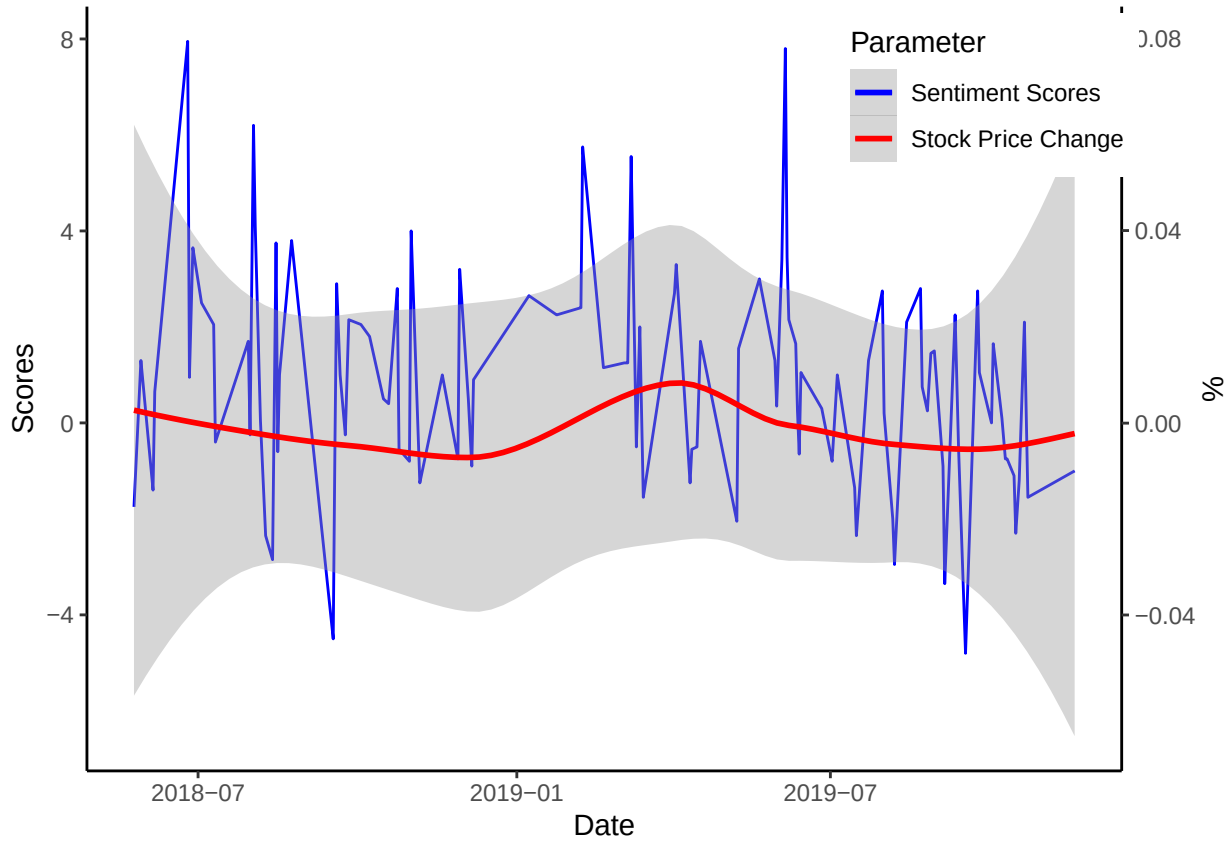



```
# Each variable on different y-axis with geom_line.
ggplot(stocks.scores,aes(x=Date)) +  geom_line(aes(y=scores, colour="Sentiment Scores")) + geom_line(aes
```




```
# Each variable on different y-axis with geom_line and geom_smooth.
ggplot(stocks.scores,aes(x=Date)) +  geom_line(aes(y=scores, colour="Sentiment Scores")) + geom_smooth(a
```

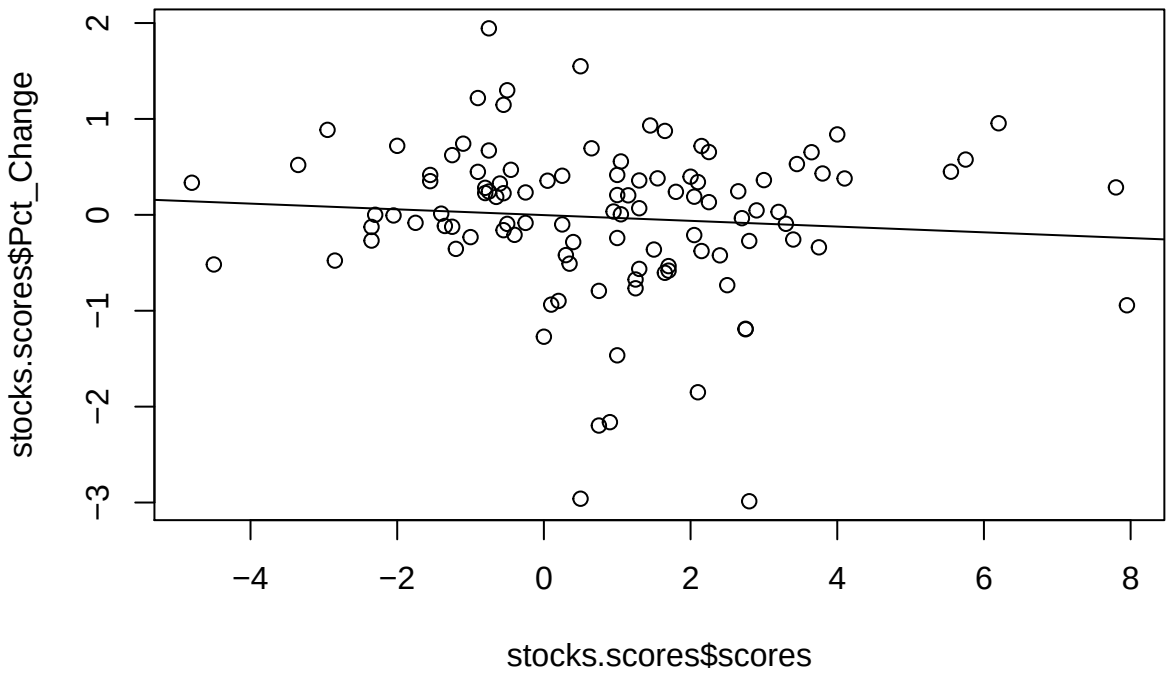
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
## Linear Regression

# Checking to see if there is meaningful linear relationship between sentiment scores and stock price c
stocks.scores.lm <- lm(Pct_Change~scores, data=stocks.scores)
summary(stocks.scores.lm)
```

```
##
## Call:
## lm(formula = Pct_Change ~ scores, data = stocks.scores)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9426 -0.3251  0.1289  0.4568  1.9250
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.002823   0.083193  -0.034    0.973
## scores      -0.029931   0.034463  -0.869    0.387
##
## Residual standard error: 0.8041 on 105 degrees of freedom
## Multiple R-squared:  0.007133,   Adjusted R-squared:  -0.002323
## F-statistic: 0.7543 on 1 and 105 DF,  p-value: 0.3871
```

```
plot(x = stocks.scores$scores, y = stocks.scores$Pct_Change)
abline(stocks.scores.lm)
```



Conclusion

- Top 5 words that are used with the topic that has the most impact on stock market price change are, “china”, “trade”, “dollar”, “billion” and “deal”
- When “trade” word is used in a tweet, it is common that words “such as deal”, “billion” , “china”, “countri”, “dollar”, “year”, “unit”, “talk”, “good”, “usa”, “long” and “meet” are used as well.
- Even though the linear regression result where p-value is greater than the significant level of 0.05 and R-squared value is approximately zero suggests that there is no meaningful relationship between stock price change and sentiment scores, but we do see there are patterns of stock price change and sentiment scores moving in the same direction in visualization section.
- Overall, we have achieved what are trying to do be able to clean up the raw tweets, classify tweets into topics, sentimentalize tweets, and finally correlate the sentiment scores with stock price change to see if both have a strong relationship. Certainly, there are something we can do better to improve the relationship between sentiment scores and stock price change such as considering tweets after 4pm when stock market close into next day sentiment analysis. This way sentiment scores trend will match better to the stock price change.

Appendix

In this section, we included additional approaches we have executed along the way. You might consider these as different iterations of the project/output.

ITERATION 2

ITERATION 1

References

- Sagar, C. (2018, March 22). Twitter Sentiment Analysis Using R. Dataaspirant. Retrieved from https://dataaspirant.com/2018/03/22/twitter-sentiment-analysis-using-r/
- Silge, J., & Robinson, D. (2019, November 24). Topic Modeling. Text Mining with R. Retrieved from https://www.tidytextmining.com/topicmodeling.html
- Cochrane, N. (2019, September 5). Trump, Tweets, and Trade. Medium. Retrieved from https://towardsdatascience.com/trump-tweets-and-trade-96ac157ef082
- Doll, T. (2018, June 24). LDA Topic Modeling: An Explanation. Medium. Retrieved from https://towardsdatascience.com/lda-topic-modeling-an-explanation-e184c90aadcd
- Brown, B. (n.d.). Trump Twitter Archive. Retrieved from http://www.trumptwitterarchive.com/archive