EDA analysis - Classification of news headlines with impact on the probability of stock prices changes

Max Franke

March 30, 2020

Course: CIS-627 CAPSTONE

Term: Spring/T2

Last modified date: 03/27/2020

Data: news headlines and descriptions ~5k news articles pertaining to several publicly traded companies

Approach of the analysis

This analysis aims to visualize the collected data in order to include subsequent results and relationships in the further process of the project. The following procedure is applied:

First, a data set is examined that evaluates the mood of headlines and their descriptions with regard to different stocks in a labeled form.

Based on this, the most frequently used words from headline and description are linked to google trends. The background to this approach is that the assumption is made that these words are representative of a word family that has an influence on stocks, so that the trend of these words can be examined more closely.

Workspace preparation

In this section the workspace will be prepared. This means that first the global environment will be cleaned, and all required packages will be loaded.

```
# knitr::opts_chunk$set(echo = FALSE)
```

Clean Workspace

Install packages

```
#install.packages("tidyverse")
#install.packages("knitr")
#install.packages("rmarkdown")
#install.packages("tinytex")
#install.packages("dplyr")
#install.packages("kableExtra")
```

Load libraries

```
library(tidyverse)
library(knitr)
library(rmarkdown)
#tinytex::install_tinytex()
library(tinytex)
library(dplyr)
library(kableExtra)
options(knitr.table.format = "latex")
options(tinytex.verbose = TRUE)
```

Load data of stocks with headlines and description

```
setwd("/Users/MaxFranke/Desktop/05_Big Data Analytics/04_Classes/04 SP:Term2/CIS-627 CAPSTONE/News-Clas
stock_text <- read.csv("Stock_Text_Symbol_new.csv")
```

First look at the data

```
knitr::kable(head(stock_text), "latex",
             booktabs = T,
             caption = "Dataset stocks with headlines and description") %>%
  kable_styling(full_width = TRUE) %>%
  column_spec(c(1,2), width = "5cm") %>%
  column_spec(c(3,4), width = "2cm") %>%
  row_spec(c(1,3,5), background = "gray")
```

Table 1: Dataset stocks with headlines and description

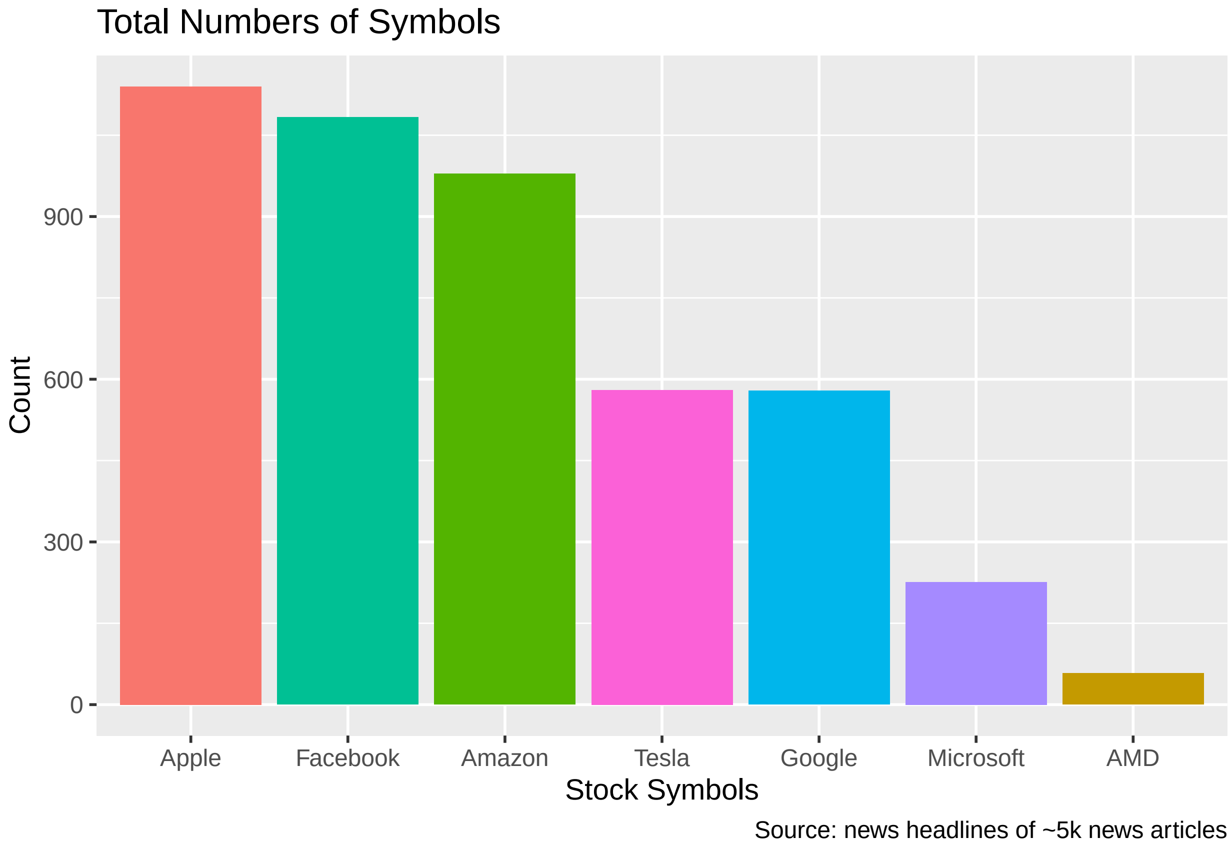
title	description	symbol	sentiment
Apple updates iMac with Intel processors	(Reuters) - Apple Inc said it updated the iMac with fourth generation Intel Corp processors, better graphics, next generation Wi-Fi and faster flash storage options.	AAPL	1
Google pays $55 million tax in Britain on 2012 sales of $5 billion	LONDON (Reuters) - Google, which has been grilled twice in the past year by a UK parliamentary committee over its tax practices, had a UK tax bill of 35 million pounds ($55 million) in 2012, on sales	GOOGL	-1
Microsoft plans to cut 1,000 jobs in Finland -newspaper	HELSINKI, July 16 (Reuters) - Microsoft Corp is planning to cut 1,000 jobs in Finland from its mobile phone unit, a Finnish daily said on Wednesday, quoting anonymous sources.	MSFT	0
Microsoft plans to cut 1,000 jobs in Finland: newspaper	HELSINKI (Reuters) - Microsoft Corp is planning to cut 1,000 jobs in Finland from its mobile phone unit, a Finnish daily said on Wednesday, quoting anonymous sources.	MSFT	0
Smartphone suit against Google plays into rivals' hands	SAN FRANCISCO (Reuters) - A U.S. consumer lawsuit accusing Google of monopolizing prime real estate on Android smartphones will help mobile rivals like Microsoft make their antitrust case with Europea	GOOGL	-1
Apple should do more to tackle in-app purchases problem: EU	BRUSSELS (Reuters) - Apple has provided no concrete and immediate solutions to tackle the problem of adults and children racking up credit card bills by making "in-app" purchases on tablets and mobile	AAPL	-1

Analyze distribution

Graph: Total numbers of symbols

The first graph shows the total numbers of symbols so we get a feeling about the distribution.

```
p1 <- ggplot(data = stock_text, mapping = aes(x = forcats::fct_infreq(symbol))) +
  geom_bar(mapping = aes(y = ..count.., fill = symbol)) +
  guides(fill = FALSE) +
  scale_x_discrete(labels = c("AAPL" = "Apple", "AMD" = "AMD", "AMZN" = "Amazon", "FB" =
              "Facebook", "GOOGL" = "Google", "MSFT" = "Microsoft", "TSLA" = "Tesla")) +
  labs(title = "Total Numbers of Symbols",
       x = "Stock Symbols",
       y = "Count",
       caption = "Source: news headlines of ~5k news articles")
p1
```



Description of the graph:

Apple, facebook and amazon shows the highest hits. The sum of these three stocks are 46.9% of the total dataset.

In the next step, the relationship between the stocks per sentiment will be analyzed.

Graph: Number of stock symbols per sentiment

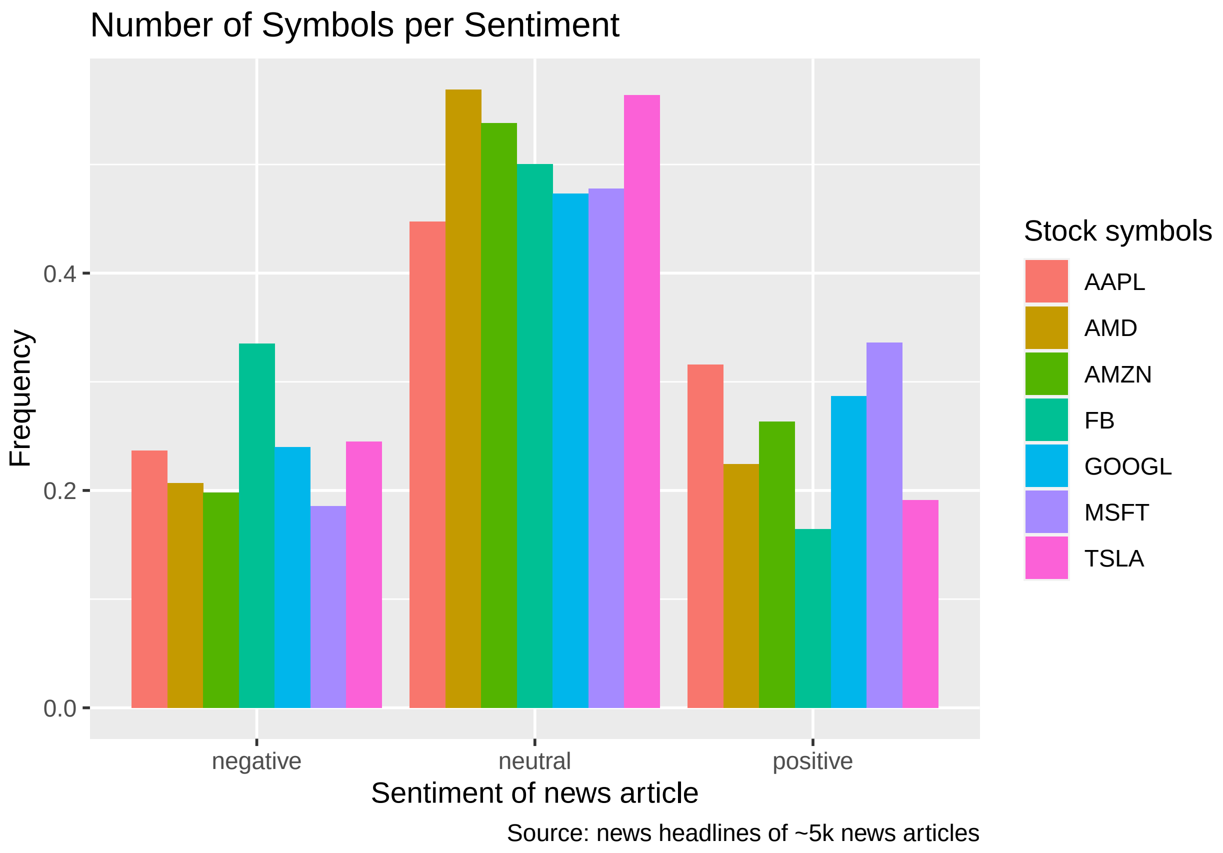
Description of the column sentiment

1 if the news is positive for the company and may encourage people to buy shares.

0 if the news is neural or not possible to identify it as positive or negative

-1 if the news is negative for the company, bad publicity, or would discourage people from owning shares.

```
# First, change the class of column sentiment to change the labels in ggplot
graph2 <- stock_text
graph2$sentiment <- as.character(graph2$sentiment)

# Plot: Barplot to compare number of stock symbol for the 3 different sentiments
p2 <- ggplot(data = graph2, mapping = aes(x = sentiment, fill = symbol)) +
  geom_bar(mapping = aes(y = ..prop..,group = symbol), position = "dodge") +
  scale_x_discrete(labels = c("1" = "positive", "0" = "neutral", "-1" = "negative")) +
  labs(fill = "Stock symbols",
       title = "Number of Symbols per Sentiment",
       x = "Sentiment of news article",
       y = "Frequency",
       caption = "Source: news headlines of ~5k news articles")
p2
```



Description of the graph:

The frequency of symbols per sentiment is described above. It can be seen that facebook has a high number of bad publicity.

In general, the frequency in the neutral area is the strongest. Apple and Microsoft show a high number with positive news.

Create word clouds

Preparation

Install packages

```r
#install.packages("tm")  # for text mining
#install.packages("SnowballC") # for text stemming
#install.packages("wordcloud") # word-cloud generator
#install.packages("RColorBrewer") # color palettes
```

Load packages

```r
library("tm")
library("SnowballC")
library("wordcloud")
library("RColorBrewer")
```

Titles

Extract and then Load the titles from dataset

```r
# Write column title to txt. file
write.table(stock_text$title, "title.txt", row.names = FALSE, col.names = FALSE)
text <- readLines("title.txt")

# Head of titles
kable(data.frame(Number = c(1:6), Titles = head(text)), "latex",
      longtable = T, booktabs = T,
      caption = "First 10 titles") %>%
  kable_styling(latex_options = c("repeat_header"))
```

Table 2: First 10 titles

Number	Titles
1	"Apple updates iMac with Intel processors"
2	"Google pays $55 million tax in Britain on 2012 sales of $5 billion"
3	"Microsoft plans to cut 1,000 jobs in Finland -newspaper"

Table 2: First 10 titles *(continued)*

Number	Titles
4	"Microsoft plans to cut 1,000 jobs in Finland: newspaper"
5	"Smartphone suit against Google plays into rivals' hands"
6	"Apple should do more to tackle in-app purchases problem: EU"

Load the data as a corpus

```r
# Load the data as a corpus
docs <- Corpus(VectorSource(text))
```

Inspect the content of the document

```r
inspect(docs[1:10])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 10
##
##  [1] "Apple updates iMac with Intel processors"
##  [2] "Google pays $55 million tax in Britain on 2012 sales of $5 billion"
##  [3] "Microsoft plans to cut 1,000 jobs in Finland -newspaper"
##  [4] "Microsoft plans to cut 1,000 jobs in Finland: newspaper"
##  [5] "Smartphone suit against Google plays into rivals' hands"
##  [6] "Apple should do more to tackle in-app purchases problem: EU"
##  [7] "Federal appeals court set to hear Microsoft 'cloud' case"
##  [8] "Mystery of 'Gold Artifact' That Stumped Archaeologists Solved by FB User"
##  [9] "How Mark Zuckerberg could prevent gun violence"
## [10] "Banksy's Steve Jobs mural spotlights refugee crisis"
```

Text transformation

```r
toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))
docs <- tm_map(docs, toSpace, "/")
docs <- tm_map(docs, toSpace, "@")
docs <- tm_map(docs, toSpace, "\\|")
```

Cleaning the text

```r
# Convert the text to lower case
docs <- tm_map(docs, content_transformer(tolower))
# Remove punctuations
docs <- tm_map(docs, removePunctuation)
```

```r
# Eliminate extra white spaces
docs <- tm_map(docs, stripWhitespace)
# Remove english common stopwords
docs <- tm_map(docs, removeWords, stopwords("english"))
# Remove the name of the stocks
docs <- tm_map(docs, removeWords,
               c("apple", "amazon", "facebook", "google", "tesla", "microsoft"))
```

Building a term-document matrix

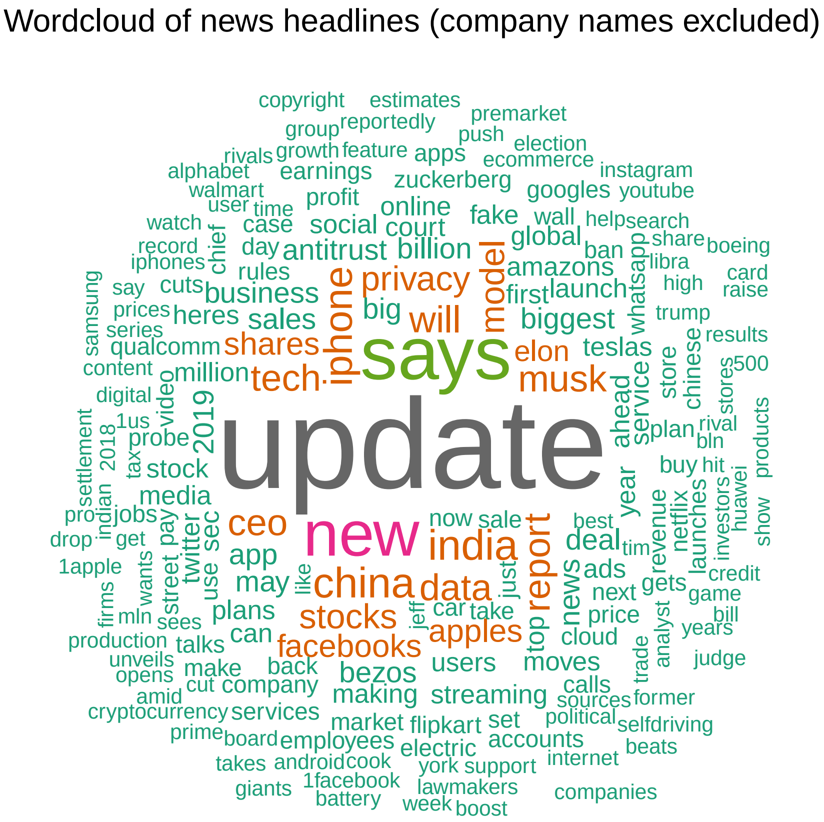
```r
dtm <- TermDocumentMatrix(docs)
dtm_matrix <- as.matrix(dtm)
dtm_vector <- sort(rowSums(dtm_matrix),decreasing=TRUE)
dtm_dataframe <- data.frame(word = names(dtm_vector),freq=dtm_vector)

kable(data.frame(head(dtm_dataframe, 10), row.names = NULL), "latex",
      longtable = T, booktabs = T,
      caption = "Top 10 words in title by frequency") %>%
  kable_styling(latex_options = c("repeat_header"))
```

Table 3: Top 10 words in title by frequency

word	freq
update	720
says	372
new	315
india	175
china	172
iphone	152
data	139
will	138
report	137
tech	136

Generate the Word cloud

```r
set.seed(1234)
layout(matrix(c(1, 2), nrow = 2), heights = c(1, 10))
par(mar=rep(0, 4))
plot.new()
text(x=0.5, y=0.5, "Wordcloud of news headlines (company names excluded)")
wordcloud(words = dtm_dataframe$word, freq = dtm_dataframe$freq, min.freq = 20,
          max.words = 200, random.order = FALSE, rot.per = 0.2,
          colors = brewer.pal(8, "Dark2"),
          main = "Title")
```


Wordcloud of news headlines (company names excluded)

Frequent terms in the term-document matrix

```
kable(data.frame("Words 1-10" = findFreqTerms(dtm, lowfreq = 30,)[1:10],
           "Words 11-20" = findFreqTerms(dtm, lowfreq = 30,)[11:20],
           "Words 21-30" = findFreqTerms(dtm, lowfreq = 30,)[21:30],
           "Words 31-40" = findFreqTerms(dtm, lowfreq = 30,)[31:40],
           "Words 41-50" = findFreqTerms(dtm, lowfreq = 30,)[41:50],
           "Words 51-60" = findFreqTerms(dtm, lowfreq = 30,)[51:60],
           check.names = FALSE), "latex",
      longtable = T, booktabs = T,
      caption = "Frequent words in title") %>%
  kable_styling(latex_options = c("repeat_header"))
```

Table 4: Frequent words in title

Words 1-10	Words 11-20	Words 21-30	Words 31-40	Words 41-50	Words 51-60
billion	set	support	report	talks	video
million	user	elon	shares	deal	trump
sales	zuckerberg	musk	twitter	internet	top
tax	back	calls	streaming	1apple	wants
cut	take	street	data	bln	ads
jobs	iphone	tech	update	car	political

Table 4: Frequent words in title *(continued)*

Words 1-10	Words 11-20	Words 21-30	Words 31-40	Words 41-50	Words 51-60
plans	moves	hit	fake	ceo	walmart
case	cook	store	stock	firms	board
cloud	tim	make	settlement	antitrust	facebooks
court	says	plan	sources	record	1us

Frequency of top 10 words

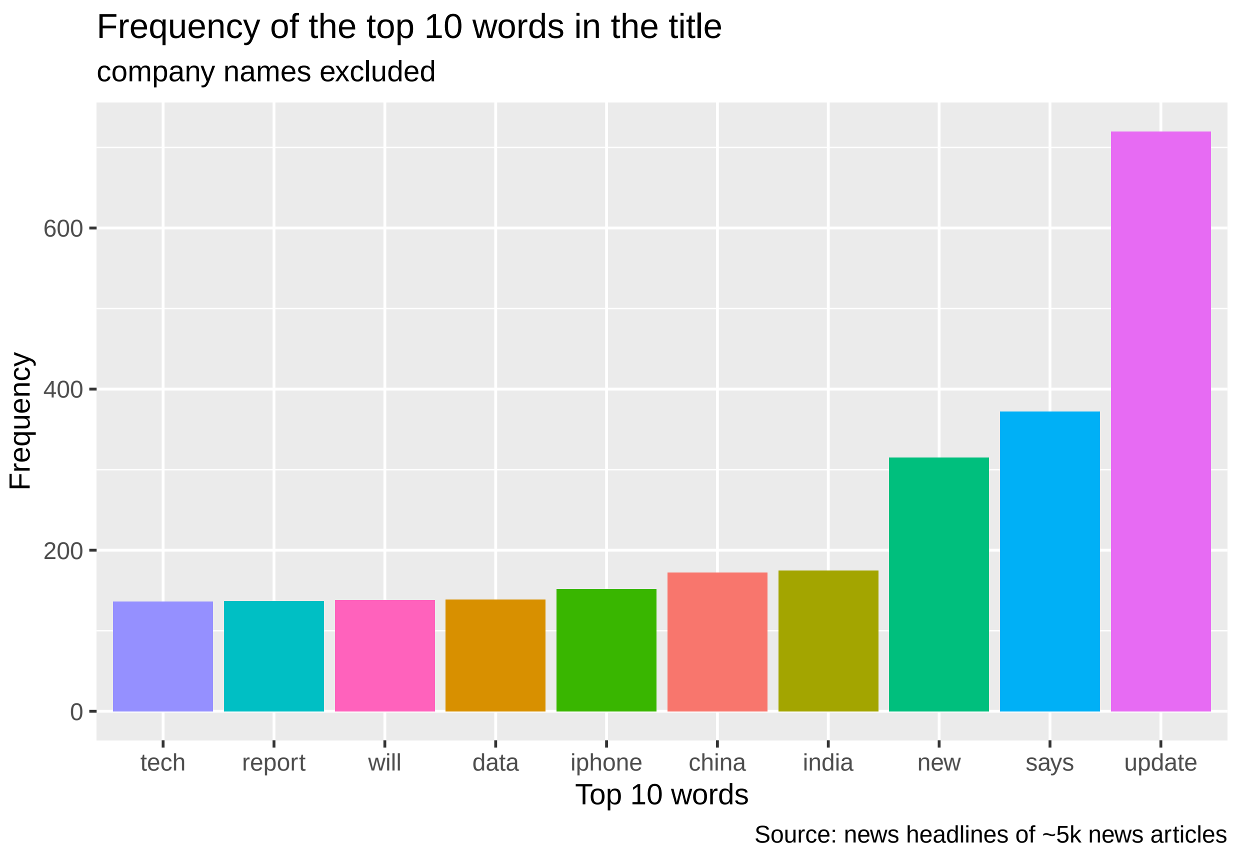
```r
top_10 <- head(dtm_dataframe, 10)

kable(data.frame(top_10, row.names = NULL), "latex",
      longtable = T, booktabs = T,
      caption = "Frequency of top 10 words in title") %>%
  kable_styling(latex_options = c("repeat_header"))
```

Table 5: Frequency of top 10 words in title

word	freq
update	720
says	372
new	315
india	175
china	172
iphone	152
data	139
will	138
report	137
tech	136

Plot 4: Word frequency for the top 10

```r
p4 <- ggplot(data = top_10, mapping = aes(x = reorder(word, freq), y = freq, fill = word)) +
  geom_col() +
  guides(fill = FALSE) +
  labs(title = "Frequency of the top 10 words in the title",
       subtitle = "company names excluded",
       x = "Top 10 words",
       y = "Frequency",
       caption = "Source: news headlines of ~5k news articles")
p4
```



Description

Extract and then Load the description from dataset

```
# Write column title to txt. file
write.table(stock_text$description, "description.txt", row.names = FALSE,
            col.names = FALSE)
text <- readLines("description.txt")

# Head of titles
kable(data.frame(Number = c(1:6),Description = head(text)), "latex",
      longtable = T, booktabs = T,
      caption = "First 10 descriptions") %>%
  kable_styling(latex_options = c("repeat_header"), full_width = TRUE) %>%
  column_spec(column = 1, width = "1cm")
```

Table 6: First 10 descriptions

Number	Description
1	"(Reuters) - Apple Inc said it updated the iMac with fourth generation Intel Corp processors, better graphics, next generation Wi-Fi and faster flash storage options."

Table 6: First 10 descriptions *(continued)*

Number	Description
2	"LONDON (Reuters) - Google, which has been grilled twice in the past year by a UK parliamentary committee over its tax practices, had a UK tax bill of 35 million pounds ($55 million) in 2012, on sales"
3	"HELSINKI, July 16 (Reuters) - Microsoft Corp is planning to cut 1,000 jobs in Finland from its mobile phone unit, a Finnish daily said on Wednesday, quoting anonymous sources."
4	"HELSINKI (Reuters) - Microsoft Corp is planning to cut 1,000 jobs in Finland from its mobile phone unit, a Finnish daily said on Wednesday, quoting anonymous sources."
5	"SAN FRANCISCO (Reuters) - A U.S. consumer lawsuit accusing Google of monopolizing prime real estate on Android smartphones will help mobile rivals like Microsoft make their antitrust case with Europea"
6	"BRUSSELS (Reuters) - Apple has provided no concrete and immediate solutions to tackle the problem of adults and children racking up credit card bills by making \"in-app\" purchases on tablets and mobile"

Load the data as a corpus

```r
# Load the data as a corpus
docs <- Corpus(VectorSource(text))
```

Inspect the content of the document

```r
inspect(docs[1:10])
```

Text transformation

```r
toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))
docs <- tm_map(docs, toSpace, "/")
docs <- tm_map(docs, toSpace, "@")
docs <- tm_map(docs, toSpace, "\\|")
```

Cleaning the text

```r
# Convert the text to lower case
docs <- tm_map(docs, content_transformer(tolower))
# Remove punctuations
docs <- tm_map(docs, removePunctuation)
# Eliminate extra white spaces
docs <- tm_map(docs, stripWhitespace)
# Remove english common stopwords
docs <- tm_map(docs, removeWords, stopwords("english"))
# Remove the name of the stocks
docs <- tm_map(docs, removeWords,
               c("apple", "amazon", "facebook", "google", "tesla", "microsoft"))
```

Building a term-document matrix

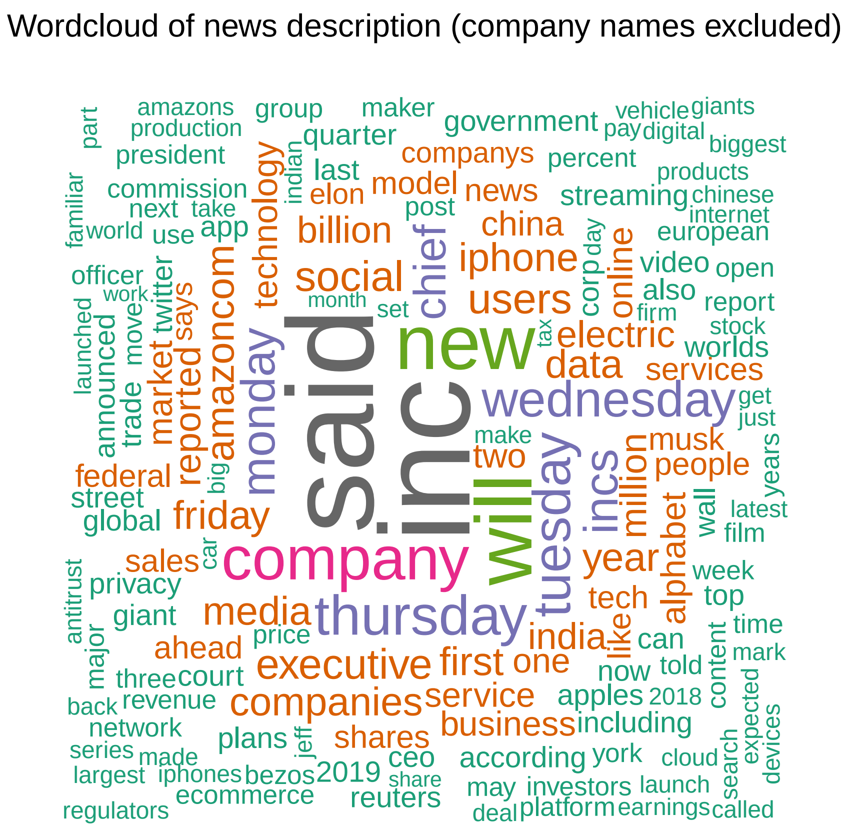
```
dtm <- TermDocumentMatrix(docs)
dtm_matrix <- as.matrix(dtm)
dtm_vector <- sort(rowSums(dtm_matrix),decreasing=TRUE)
dtm_dataframe <- data.frame(word = names(dtm_vector),freq=dtm_vector)

kable(data.frame(head(dtm_dataframe, 10), row.names = NULL), "latex",
      longtable = T, booktabs = T,
      caption = "Top 10 words in description by frequency") %>%
  kable_styling(latex_options = c("repeat_header"))
```

Table 7: Top 10 words in description by frequency

word	freq
inc	1118
said	1066
will	627
new	611
company	445
thursday	390
tuesday	381
wednesday	355
monday	321
incs	318

Generate the Word cloud

```
set.seed(1234)
layout(matrix(c(1, 2), nrow = 2), heights = c(1, 10))
par(mar=rep(0, 4))
plot.new()
text(x=0.5, y=0.5, "Wordcloud of news description (company names excluded)")
wordcloud(words = dtm_dataframe$word, freq = dtm_dataframe$freq, min.freq = 20,
          max.words = 200, random.order = FALSE, rot.per = 0.2,
          colors = brewer.pal(8, "Dark2"),
          main = "Title")
```


Wordcloud of news description (company names excluded)

Frequent terms in the term-document matrix

```
kable(data.frame("Words 1-10" = findFreqTerms(dtm, lowfreq = 30,)[1:10],
           "Words 11-20" = findFreqTerms(dtm, lowfreq = 30,)[11:20],
           "Words 21-30" = findFreqTerms(dtm, lowfreq = 30,)[21:30],
           "Words 31-40" = findFreqTerms(dtm, lowfreq = 30,)[31:40],
           "Words 41-50" = findFreqTerms(dtm, lowfreq = 30,)[41:50],
           "Words 51-60" = findFreqTerms(dtm, lowfreq = 30,)[51:60],
           check.names = FALSE), "latex",
      longtable = T, booktabs = T,
      caption = "Frequent words in description") %>%
  kable_styling(latex_options = c("repeat_header"))
```

Table 8: Frequent words in description

Words 1-10	Words 11-20	Words 21-30	Words 31-40	Words 41-50	Words 51-60
better	sales	android	san	states	says
corp	tax	antitrust	smartphones	united	use
inc	year	case	will	week	appeared
next	cut	consumer	card	york	ceo
reuters	jobs	help	credit	according	late
said	mobile	lawsuit	making	user	latest

Table 8: Frequent words in description *(continued)*

Words 1-10	Words 11-20	Words 21-30	Words 31-40	Words 41-50	Words 51-60
bill	phone	like	court	big	one
million	sources	make	customers	business	allow
past	unit	prime	federal	data	comments
practices	wednesday	rivals	new	founder	group

Frequency of top 10 words

```r
top_10 <- head(dtm_dataframe, 10)

kable(data.frame(top_10, row.names = NULL), "latex",
      longtable = T, booktabs = T,
      caption = "Frequency of top 10 words in title") %>%
  kable_styling(latex_options = c("repeat_header"))
```

Table 9: Frequency of top 10 words in title

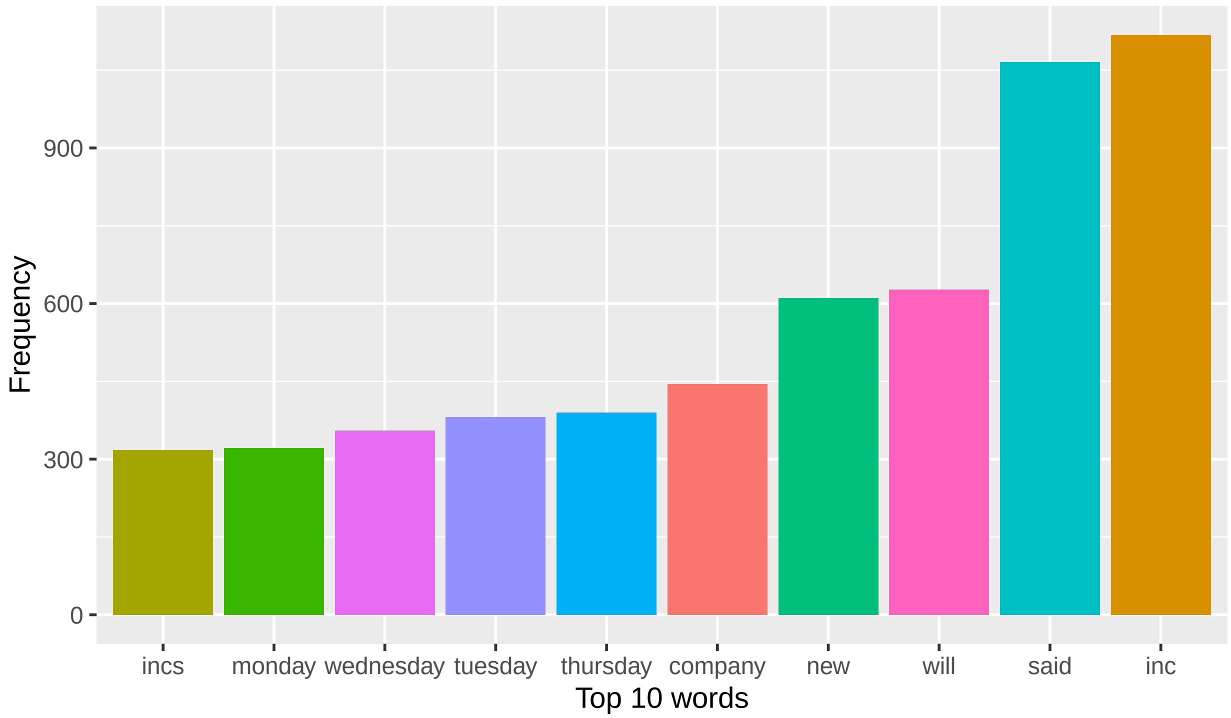
word	freq
inc	1118
said	1066
will	627
new	611
company	445
thursday	390
tuesday	381
wednesday	355
monday	321
incs	318

Plot 6: Word frequency for the top 10

```r
p6 <- ggplot(data = top_10, mapping = aes(x = reorder(word, freq), y = freq, fill = word)) +
  geom_col() +
  guides(fill = FALSE) +
  labs(title = "Frequency of the top 10 words in the description",
       subtitle = "company names excluded",
       x = "Top 10 words",
       y = "Frequency",
       caption = "Source: news headlines of ~5k news articles")
p6
```

Frequency of the top 10 words in the description

company names excluded



Source: news headlines of ~5k news articles