TEXT ANALYTICS
BUSINESS INSIGHT REPORT

NAME: CHRISTINA JOHN GEORGE

STUDENT ID: 4800171

COHORT: MSBA5-VALENCIA

Text Analysis on the Retail Market

This study is based on the area that intrests me the most i.e. the retail space. There has been so many changes in the past decade with so many retail stores closing down and study focuses mainly on the reasons behind these major changes. The Retail industry has seen a lot of downs in the past few years and it related to their strategies used to survive in the market .Also, with the increasing competition within the industry, the need to make multiple changes in their model in order to thrive.

The three companies whose failing to survive surprised me the most and drew me to learn more about the reasons behind it are:

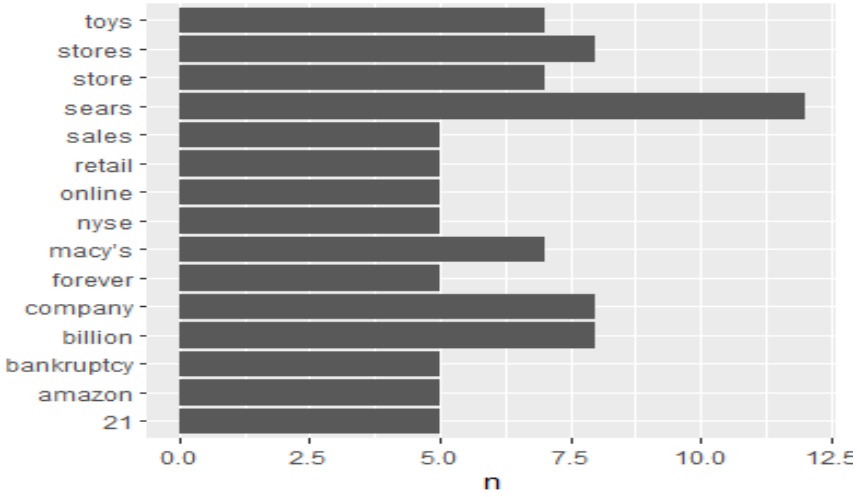
- Macy's
- Sears
- Forever 21
- Toys R Us

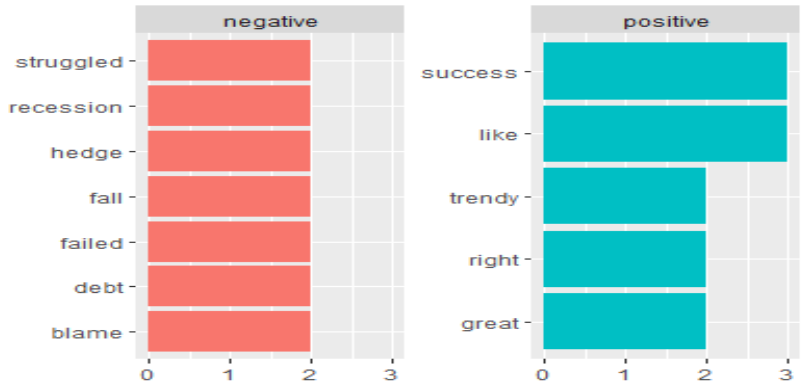
Some of the main frameworks I used to come up with some compelling business insights were:

- Tokenization
- Creating a Data Frame
- Transposing the Data Frame
- Eliminating common stopwords in English
- Analyzing the frequency of words
- Word cloud
- Sentiment Analysis
- Quadrogram

The above frameworks helped me understand the unstructured data and derive business insights in an organized format.

After tokenizing, calculating the frequency of words and segregating them into positive or negative using tidy, the words that were most frequently appearing were:





The above output shows that most of the words are related to bankruptcy, amazon, the stores etc.

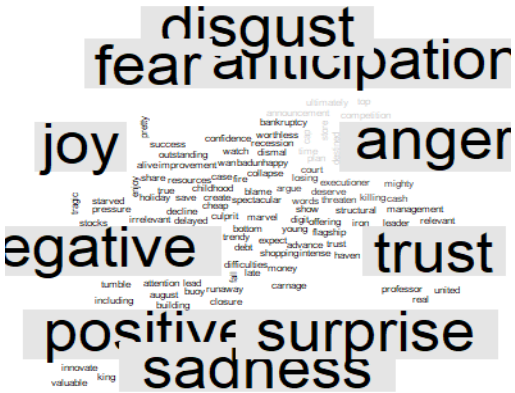
Assuming the reasons being these stores failing could be either because of them going bankrupt or because amazon took over the market.

Also after using "bing" sentiment, it divided the words to positive and negative . For eg: debt, recession etc. being classified as negative which tells some of the reasons behind these stores closing down.

In order to dig deeper into my data, I went ahead and created two types of word cloud:

- A basic word cloud: this word cloud dint give me an accurate idea but just an overview of what is present in the data in a unstructured formart.

- A word cloud comparison between the companies using sentiment analysis("nrc")



As stated above, the first word cloud doesn't produce any specific insight but the second gives a better view of every element.

I also used Quadrogram to understand the data more and come up with more relevant insights.

For example:

	store	quadrogram	n	tf	idf	tf_idf
	<chr>	<chr>	<int>	<dbl>	<dbl>	<dbl>
1	Sears	company.credit united press international	1	0.143	1.39	0.198
2	Sears	consecutive double digit decline	1	0.143	1.39	0.198
3	Sears	hedge fund moneyman edward	1	0.143	1.39	0.198
4	Sears	pocket watch salesman sears	1	0.143	1.39	0.198
5	Sears	salesman sears navigated retailing	1	0.143	1.39	0.198
6	Sears	wall street financial engineering	1	0.143	1.39	0.198
7	Sears	watch salesman sears navigated	1	0.143	1.39	0.198
8	Toys R Us	7.5 billion leveraged buyout	1	0.0833	1.39	0.116
9	Toys R Us	bain capital partners kohlberg	1	0.0833	1.39	0.116
0	Toys R Us	business model incorporate technology	1	0.0833	1.39	0.116

Business Insight

The sentiment analysis emphasizes on the failed struggle against online competitors, and the ever increasing debts private equity firms, once successful and trendy fashion shops are closing. It also brings into the light the key reasons behind the failure of these particular stores.

Toys R Us, done with struggles and finally closed and filed for bankruptcy with the increasing debts and buyout by equity firms. The word bankruptcy is prominent in all the stores and appears quite a few times in all of the 3 stores.

In the analysis done using quadrogram ,it shows that there was a consecutive double digit decline in the sales of these which indirectly tells us that ,the stores were going in a loss since a while and it wasn't a recent phenomenon .In the case of Toys R us, even though it was a store loved by all ,they couldn't survive the debt they had to pay off. Also, the completion being so high with so many technology advances happening in the market –Toys R Us couldn't survive between companies like Amazon.

In my opinion, the success of the stores will be only possible if you can make the customers like their new plans and attractions. Targeting the younger generations with trendy and attractive games corners etc. in the malls and make the customers come back to the stores like before.

We can summarize the above closure of stores due to very high costs to manage retail outlets rising salaries and fall in customer visits due to online giants like amazon for example. Also low investments and weak forward planning has to be worked upon for success .Customers not the stores benefit from recession because stores will have to reduce prices to stay alive.

On an optimistic note, Customers will surely get the urge to visit malls and retail stores sooner or later. Face to face dealings are more exciting and satisfying especially for certain items that people still like to go and purchase from the store. However according to the analysis on the market, I believe that retail stores need to incorporate the latest technology advances in order to survive and surpass their competition in the Market.

APPENDIX

R CODE

```
install.packages('pdftools')
install.packages('shapeR')
install.packages('tidytext')
install.packages('tidyverse')
install.packages("textreadr")
install.packages("textshape")
install.packages("dplyr")
install.packages("textdata")
install.packages("wordcloud")
install.packages("RColorBrewer")
install.packages("wordcloud2")
install.packages("tm")
install.packages("packages")
rm(list=ls())
#Instaling libraries
library(pdftools)
library(shapeR)
library(tidyverse)
library(textshape)
library(textreadr)
library(tidytext)
library(dplyr)
library(tidyr)
library(scales)
library(dplyr)
```

```
library(stringr)
library(ggplot2)
library(tidytext)
library(textdata)
library(wordcloud)
library(wordcloud2)
library(RColorBrewer)
library(tm)
library(NLP)
library(reshape2)
#file for assignment
setwd("C:/Users/Christina/Desktop/MSBA SPRING/TEXT ANALTICS/pdf")
nm <- list.files("C:/Users/Christina/Desktop/MSBA SPRING/TEXT ANALTICS/pdf")

#Read read and save as a data frame
my_pdf_text <- as.data.frame(do.call(rbind, lapply(nm,function(x) pdf_text(x))))
View(my_pdf_text)
#Transposing and divinding company wise
store_names <- c('Macy','Sears','Forever 21','Toys R Us')
my_pdf_text <data_frame(line=1:ncol(my_pdf_text),text=t(as.matrix(my_pdf_text)),store=store_names)
view(my_pdf_text)



token_list <- my_pdf_text %>%
  unnest_tokens(word, text)
print(token_list)
#frequencies
frequencies_tokens <- my_pdf_text %>%
  unnest_tokens(word, text) %>%
  count(word, sort=TRUE)
```

```
print(frequencies_tokens)
#frequency
data(stop_words)
frequency_tokens_nostop <- my_pdf_text %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  count(word, sort=TRUE)
print(frequency_tokens_nostop)
#WORDCLOUD
wordcloud2(data=frequency_tokens_nostop, size=0.5, color='random-dark')
#plot histogram
freq_hist <- my_pdf_text %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  count(word, sort=FALSE) %>%
  filter(n >4) %>%
  ggplot(aes(word, n))+
  geom_col()+
  xlab(NULL)+
  coord_flip()
print(freq_hist)

#sentiment analysis
print(sentiments)
afinn <-get_sentiments("afinn")
nrc <- get_sentiments("nrc")
bing <- get_sentiments("bing")
table(nrc$sentiment)
table(bing$sentiment)
sentiments <- bind_rows(
```

```
  (mutate(afinn,lexicon="afinn")),
  (mutate(nrc,lexicon="nrc")),
  (mutate(bing,lexicon="bing"))
)
my_sentiment <-my_pdf_text %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  inner_join(get_sentiments("afinn")) %>%
  #summary mean value
  summarise(mean(value))
View(my_sentiment)

tidy_store <- my_pdf_text %>%
  unnest_tokens(word, text) %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort=T) %>%
  ungroup()
tidy_store

tidy_store <- my_pdf_text %>%
  unnest_tokens(word, text) %>%
  inner_join(get_sentiments("nrc")) %>%
  count(word, sentiment, sort=T) %>%
  ungroup()
tidy_store

#plot the graph
tidy_store %>%
  group_by(sentiment) %>%
  top_n(5) %>%
```

```
  ungroup() %>%
  mutate(word=reorder(word, n)) %>%
  ggplot(aes(word, n, fill=sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y")+
  labs(y="different types of sentiment", x=NULL)+
  coord_flip()

#wordcloud
install.packages("wordcloud")
library(wordcloud)
data("stop_words")

tidy_store %>%
  inner_join(get_sentiments("nrc")) %>%
  count(word, sentiment, sort=TRUE) %>%
  acast(word ~sentiment, value.var="n", fill=0) %>%
  comparison.cloud(colors = c("grey20", "gray80"),
                   max.words=100,
                   scale<-c(0.5,0.5),
                   fixed.asp=TRUE,
                   title.size =2)
#quadrogram
quadrogram <- my_pdf_text %>%
  unnest_tokens(quadrogram, text, token = "ngrams", n=4) %>%
  separate(quadrogram, c("word1", "word2", "word3", "word4"), sep=" ") %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word) %>%
  filter(!word3 %in% stop_words$word) %>%
  filter(!word4 %in% stop_words$word)
```

```
quadrogram

quadrogram_united <- quadrogram %>%

 unite(quadrogram, word1, word2, word3, word4, sep=" ")

quadrogram_tf_idf <- quadrogram_united %>%

 count(store, quadrogram) %>%

 bind_tf_idf(quadrogram, store, n) %>%

 arrange(desc(tf_idf))

quadrogram_tf_idf
```

OUTPUT

```
>setwd("C:/Users/Christina/Desktop/MSBA SPRING/TEXT ANALTICS/pdf")
> nm <- list.files("C:/Users/Christina/Desktop/MSBA SPRING/TEXT ANALTICS/pdf")
> my_pdf_text <- as.data.frame(do.call(rbind, lapply(nm,function(x) pdf_text(x))))
> View(my_pdf_text)
> store_names <- c('Macy','Sears','Forever 21','Toys R Us')
>
> my_pdf_text <- data_frame(line=1:ncol(my_pdf_text),text=t(as.matrix(my_pdf_text)),store=store_n
Warning message:
`data_frame()` is deprecated, use `tibble()`.
This warning is displayed once per session.
> view(my_pdf_text)
> token_list <- my_pdf_text %>%
+   unnest_tokens(word, text)
> print(token_list)
# A tibble: 1,429 x 3
    line store word
   <int> <chr> <chr>
 1     1 Macy  macy
 2     1 Macy  of
 3     1 Macy  all
 4     1 Macy  the
 5     1 Macy  american
 6     1 Macy  marquee
 7     1 Macy  department
 8     1 Macy  store
 9     1 Macy  chains
10     1 Macy  that
# ... with 1,419 more rows
> frequencies_tokens <- my_pdf_text %>%
+   unnest_tokens(word, text) %>%
+   count(word, sort=TRUE)
> print(frequencies_tokens)
# A tibble: 688 x 2
   word      n
   <chr> <int>
 1 the      95
 2 of       46
 3 in       39
 4 to       37
 5 a        30
 6 and      29
```

```
 7 that     24
 8 its      19
 9 it       15
10 has      12
# ... with 678 more rows
> data(stop_words)
> frequency_tokens_nostop <- my_pdf_text %>%
+   unnest_tokens(word, text) %>%
+   anti_join(stop_words) %>%
+   count(word, sort=TRUE)
Joining, by = "word"
> print(frequency_tokens_nostop)
# A tibble: 513 x 2
   word           n
   <chr>      <int>
 1 sears         12
 2 billion        8
 3 company        8
 4 stores         8
 5 macy's         7
 6 store          7
 7 toys           7
 8 21             5
 9 amazon         5
10 bankruptcy     5
# ... with 503 more rows
> freq_hist <- my_pdf_text %>%
+   unnest_tokens(word, text) %>%
+   anti_join(stop_words) %>%
+   count(word, sort=FALSE) %>%
+   filter(n >4) %>%
+   ggplot(aes(word, n))+
+   geom_col()+
+   xlab(NULL)+
+   coord_flip()
Joining, by = "word"
> print(freq_hist)
> print(sentiments)
# A tibble: 6,786 x 2
   word        sentiment
   <chr>       <chr>
 1 2-faces     negative
 2 abnormal    negative
 3 abolish     negative
 4 abominable  negative
 5 abominably  negative
 6 abominate   negative
 7 abomination negative
 8 abort       negative
 9 aborted     negative
10 aborts      negative
# ... with 6,776 more rows
>
> afinn <-get_sentiments("afinn")
> nrc <- get_sentiments("nrc")
> bing <- get_sentiments("bing")
> table(nrc$sentiment)

       anger anticipation      disgust         fear          joy     negative     positive
        1247          839         1058         1476          689         3324         2312
    surprise        trust
         534         1231
> table(bing$sentiment)

negative positive
    4781     2005
```

```
>
> sentiments <- bind_rows(
+   (mutate(afinn,lexicon="afinn")),
+   (mutate(nrc,lexicon="nrc")),
+   (mutate(bing,lexicon="bing"))
+ )
> my_sentiment <-my_pdf_text %>%
+   unnest_tokens(word, text) %>%
+   anti_join(stop_words) %>%
+   inner_join(get_sentiments("afinn")) %>%
+   #summary mean value
+   summarise(mean(value))
Joining, by = "word"
Joining, by = "word"
> View(my_sentiment)
>
> tidy_store <- my_pdf_text %>%
+   unnest_tokens(word, text) %>%
+   inner_join(get_sentiments("bing")) %>%
+   count(word, sentiment, sort=T) %>%
+   ungroup()
Joining, by = "word"
> tidy_store
# A tibble: 60 x 3
   word       sentiment     n
   <chr>      <chr>     <int>
 1 like       positive      3
 2 success    positive      3
 3 blame      negative      2
 4 debt       negative      2
 5 failed     negative      2
 6 fall       negative      2
 7 great      positive      2
 8 hedge      negative      2
 9 recession  negative      2
10 right      positive      2
# ... with 50 more rows
> tidy_store %>%
+   group_by(sentiment) %>%
+   top_n(5) %>%
+   ungroup() %>%
+   mutate(word=reorder(word, n)) %>%
+   ggplot(aes(word, n, fill=sentiment)) +
+   geom_col(show.legend = FALSE) +
+   facet_wrap(~sentiment, scales = "free_y")+
+   labs(y="different types of sentiment", x=NULL)+
+   coord_flip()
Selecting by n
> tidy_store %>%
+   group_by(sentiment) %>%
+   top_n(5) %>%
+
+   ungroup() %>%
+   mutate(word=reorder(word, n)) %>%
+   ggplot(aes(word, n, fill=sentiment)) +
+   geom_col(show.legend = FALSE) +
+   facet_wrap(~sentiment, scales = "free_y")+
+   labs(y="different types of sentiment", x=NULL)+
+   coord_flip()
Selecting by n
> library(wordcloud)
> data("stop_words")
>
> tidy_store %>%
+   inner_join(get_sentiments("nrc")) %>%
+   count(word, sentiment, sort=TRUE) %>%
```

```
+   acast(word ~sentiment, value.var="n", fill=0) %>%
+   comparison.cloud(colors = c("grey20", "gray80"),
+                    max.words=100,
+                    scale<-c(0.5,0.5),
+                    fixed.asp=TRUE,#fixthe aspect
+                    title.size =2)
Joining, by = c("word", "sentiment")
Error in acast(., word ~ sentiment, value.var = "n", fill = 0) :
  could not find function "acast"
> tidy_store %>%
+   inner_join(get_sentiments("bing")) %>%
+   count(word, sentiment, sort=TRUE) %>%
+   acast(word ~sentiment, value.var="n", fill=0) %>%
+   comparison.cloud(colors = c("grey20", "black"),
+                    max.words=50,
+                    scale<-c(0.5,0.5),
+                    fixed.asp=TRUE,#fixthe aspect
+                    title.size = 2 )
Joining, by = c("word", "sentiment")
Error in acast(., word ~ sentiment, value.var = "n", fill = 0) :
  could not find function "acast"
> library(reshape2)

Attaching package: ‘reshape2’

The following object is masked from ‘package:tidyr’:

    smiths

> tidy_store %>%
+   inner_join(get_sentiments("nrc")) %>%
+   count(word, sentiment, sort=TRUE) %>%
+   acast(word ~sentiment, value.var="n", fill=0) %>%
+   comparison.cloud(colors = c("grey20", "gray80"),
+                    max.words=100,
+                    scale<-c(0.5,0.5),
+                    fixed.asp=TRUE,#fixthe aspect
+                    title.size =2)
Joining, by = c("word", "sentiment")
There were 50 or more warnings (use warnings() to see the first 50)
> tidy_store %>%
+   inner_join(get_sentiments("nrc")) %>%
+   count(word, sentiment, sort=TRUE) %>%
+   acast(word ~sentiment, value.var="n", fill=0) %>%
+   comparison.cloud(colors = c("grey20", "gray80"),
+                    max.words=100,
+                    scale<-c(1.5,1.5),
+                    fixed.asp=TRUE,#fixthe aspect
+                    title.size =2)
Joining, by = c("word", "sentiment")
There were 50 or more warnings (use warnings() to see the first 50)
> tidy_store %>%
+   inner_join(get_sentiments("nrc")) %>%
+   count(word, sentiment, sort=TRUE) %>%
+   acast(word ~sentiment, value.var="n", fill=0) %>%
+   comparison.cloud(colors = c("grey20", "gray80"),
+                    max.words=100,
+                    scale<-c(1.0,1.0),
+                    fixed.asp=TRUE,#fixthe aspect
+                    title.size =1)
Joining, by = c("word", "sentiment")
There were 50 or more warnings (use warnings() to see the first 50)
> tidy_store %>%
+   inner_join(get_sentiments("nrc")) %>%
+   count(word, sentiment, sort=TRUE) %>%
+   acast(word ~sentiment, value.var="n", fill=0) %>%
```

```
+   comparison.cloud(colors = c("grey20", "gray80"),
+                    max.words=100,
+                    scale<-c(0.4,1),
+                    fixed.asp=TRUE,#fixthe aspect
+                    title.size =1)
Joining, by = c("word", "sentiment")
There were 50 or more warnings (use warnings() to see the first 50)
> tidy_store %>%
+   inner_join(get_sentiments("nrc")) %>%
+   count(word, sentiment, sort=TRUE) %>%
+   acast(word ~sentiment, value.var="n", fill=0) %>%
+   comparison.cloud(colors = c("grey20", "gray80"),
+                    max.words=100,
+                    scale<-c(0.5,0.5),
+                    fixed.asp=TRUE,#fixthe aspect
+                    title.size =1)
Joining, by = c("word", "sentiment")
There were 50 or more warnings (use warnings() to see the first 50)
> tidy_store %>%
+   inner_join(get_sentiments("nrc")) %>%
+   count(word, sentiment, sort=TRUE) %>%
+   acast(word ~sentiment, value.var="n", fill=0) %>%
+   comparison.cloud(colors = c("grey20", "gray80"),
+                    max.words=100,
+                    scale<-c(0.5,0.5),
+                    fixed.asp=TRUE,#fixthe aspect
+                    title.size =2)
Joining, by = c("word", "sentiment")
There were 50 or more warnings (use warnings() to see the first 50)
> quadrogram <- my_pdf_text %>%
+   unnest_tokens(quadrogram, text, token = "ngrams", n=4) %>%
+   separate(quadrogram, c("word1", "word2", "word3", "word4"), sep=" ") %>%
+   filter(!word1 %in% stop_words$word) %>%
+   filter(!word2 %in% stop_words$word) %>%
+   filter(!word3 %in% stop_words$word) %>%
+   filter(!word4 %in% stop_words$word)
There were 50 or more warnings (use warnings() to see the first 50)
>
> quadrogram
# A tibble: 59 x 6
    line store word1       word2       word3      word4
   <int> <chr> <chr>       <chr>       <chr>      <chr>
 1     1 Macy  american    marquee     department store
 2     1 Macy  marquee     department  store      chains
 3     1 Macy  american    consumerism it's       true
 4     1 Macy  cousins     jcpenney    nyse       jcp
 5     1 Macy  jcpenney    nyse        jcp        sears
 6     1 Macy  nyse        jcp         sears      kohl's
 7     1 Macy  jcp         sears       kohl's     nyse
 8     1 Macy  sears       kohl's      nyse       kss
 9     1 Macy  digital     marketplace amazon     nyse
10     1 Macy  marketplace amazon      nyse       amzn
# ... with 49 more rows
> tidy_store <- my_pdf_text %>%
+   unnest_tokens(word, text) %>%
+   inner_join(get_sentiments("nrc")) %>%
+   count(word, sentiment, sort=T) %>%
+   ungroup()
Joining, by = "word"
> tidy_store
# A tibble: 256 x 3
   word        sentiment        n
   <chr>       <chr>        <int>
 1 store       anticipation     7
 2 store       positive         7
 3 bankruptcy  anger            5
```

```
 4 bankruptcy  disgust          5
 5 bankruptcy  fear             5
 6 bankruptcy  negative         5
 7 bankruptcy  sadness          5
 8 competition anticipation     3
 9 competition negative         3
10 shopping    anticipation     3
# ... with 246 more rows
> tidy_store <- my_pdf_text %>%
+   unnest_tokens(word, text) %>%
+   inner_join(get_sentiments("bing")) %>%
+   count(word, sentiment, sort=T) %>%
+   ungroup()
Joining, by = "word"
> tidy_store
# A tibble: 60 x 3
   word       sentiment     n
   <chr>      <chr>     <int>
 1 like       positive      3
 2 success    positive      3
 3 blame      negative      2
 4 debt       negative      2
 5 failed     negative      2
 6 fall       negative      2
 7 great      positive      2
 8 hedge      negative      2
 9 recession  negative      2
10 right      positive      2
# ... with 50 more rows
> my_sentiment <-my_pdf_text %>%
+   unnest_tokens(word, text) %>%
+   anti_join(stop_words) %>%
+   inner_join(get_sentiments("afinn")) %>%
+   #summary mean value
+   summarise(mean(value))
Joining, by = "word"
Joining, by = "word"
> View(my_sentiment)
> print(sentiments)
# A tibble: 23,164 x 4
   word       value lexicon sentiment
   <chr>      <dbl> <chr>   <chr>
 1 abandon       -2 afinn   NA
 2 abandoned     -2 afinn   NA
 3 abandons      -2 afinn   NA
 4 abducted      -2 afinn   NA
 5 abduction     -2 afinn   NA
 6 abductions    -2 afinn   NA
 7 abhor         -3 afinn   NA
 8 abhorred      -3 afinn   NA
 9 abhorrent     -3 afinn   NA
10 abhors        -3 afinn   NA
# ... with 23,154 more rows
>
> afinn <-get_sentiments("afinn")
> nrc <- get_sentiments("nrc")
> bing <- get_sentiments("bing")
> table(nrc$sentiment)

     anger anticipation      disgust         fear          joy     negative     positive      sad
      1247          839         1058         1476          689         3324         2312
  surprise        trust
       534         1231
> table(bing$sentiment)

negative positive
```

```
    4781     2005
>
> sentiments <- bind_rows(
+   (mutate(afinn,lexicon="afinn")),
+   (mutate(nrc,lexicon="nrc")),
+   (mutate(bing,lexicon="bing"))
+ )
>
> my_sentiment <-my_pdf_text %>%
+   unnest_tokens(word, text) %>%
+   anti_join(stop_words) %>%
+   inner_join(get_sentiments("afinn")) %>%
+   #summary mean value
+   summarise(mean(value))
Joining, by = "word"
Joining, by = "word"
> View(my_sentiment)
> tidy_store %>%
+   group_by(sentiment) %>%
+   top_n(5) %>%
+   ungroup() %>%
+   mutate(word=reorder(word, n)) %>%
+   ggplot(aes(word, n, fill=sentiment)) +
+   geom_col(show.legend = FALSE) +
+   facet_wrap(~sentiment, scales = "free_y")+
+   labs(y="different types of sentiment", x=NULL)+
+   coord_flip()
Selecting by n
> tidy_store %>%
+   inner_join(get_sentiments("nrc")) %>%
+   count(word, sentiment, sort=TRUE) %>%
+   acast(word ~sentiment, value.var="n", fill=0) %>%
+   comparison.cloud(colors = c("grey20", "gray80"),
+                    max.words=100,
+                    scale<-c(0.5,0.5),
+                    fixed.asp=TRUE,
+                    title.size =2)
Joining, by = c("word", "sentiment")
There were 50 or more warnings (use warnings() to see the first 50)
> #plot the graph
> tidy_store %>%
+   group_by(sentiment) %>%
+   top_n(5) %>%
+   ungroup() %>%
+   mutate(word=reorder(word, n)) %>%
+   ggplot(aes(word, n, fill=sentiment)) +
+   geom_col(show.legend = FALSE) +
+   facet_wrap(~sentiment, scales = "free_y")+
+   labs(y="different types of sentiment", x=NULL)+
+   coord_flip()
Selecting by n
There were 50 or more warnings (use warnings() to see the first 50)
> quadrogram <- my_pdf_text %>%
+   unnest_tokens(quadrogram, text, token = "ngrams", n=4) %>%
+   separate(quadrogram, c("word1", "word2", "word3", "word4"), sep=" ") %>%
+   filter(!word1 %in% stop_words$word) %>%
+   filter(!word2 %in% stop_words$word) %>%
+   filter(!word3 %in% stop_words$word) %>%
+   filter(!word4 %in% stop_words$word)
>
> quadrogram
# A tibble: 59 x 6
    line store word1       word2       word3      word4
   <int> <chr> <chr>       <chr>       <chr>      <chr>
 1     1 Macy  american    marquee     department store
 2     1 Macy  marquee     department  store      chains
```

```
 3     1 Macy  american    consumerism it's       true
 4     1 Macy  cousins     jcpenney    nyse       jcp
 5     1 Macy  jcpenney    nyse        jcp        sears
 6     1 Macy  nyse        jcp         sears      kohl's
 7     1 Macy  jcp         sears       kohl's     nyse
 8     1 Macy  sears       kohl's      nyse       kss
 9     1 Macy  digital     marketplace amazon     nyse
10     1 Macy  marketplace amazon      nyse       amzn
# ... with 49 more rows
> quadrogram_united <- quadrogram %>%
+   unite(quadrogram, word1, word2, word3, word4, sep=" ")
> quadrogram_tf_idf <- quadrogram_united %>%
+   count(store, quadrogram) %>%
+   bind_tf_idf(quadrogram, store, n) %>%
+   arrange(desc(tf_idf))
>
> quadrogram_tf_idf
# A tibble: 59 x 6
   store     quadrogram                                      n     tf   idf tf_idf
   <chr>     <chr>                                       <int>  <dbl> <dbl>  <dbl>
 1 Sears     company.credit united press international     1 0.143   1.39  0.198
 2 Sears     consecutive double digit decline             1 0.143   1.39  0.198
 3 Sears     hedge fund moneyman edward                   1 0.143   1.39  0.198
 4 Sears     pocket watch salesman sears                  1 0.143   1.39  0.198
 5 Sears     salesman sears navigated retailing           1 0.143   1.39  0.198
 6 Sears     wall street financial engineering            1 0.143   1.39  0.198
 7 Sears     watch salesman sears navigated               1 0.143   1.39  0.198
 8 Toys R Us 7.5 billion leveraged buyout                 1 0.0833  1.39  0.116
 9 Toys R Us bain capital partners kohlberg               1 0.0833  1.39  0.116
10 Toys R Us business model incorporate technology        1 0.0833  1.39  0.116
# ... with 49 more rows



>
```