Realestate Sentiment Analysis

Intro

For this analysis we will be using Yahoo Financal News to see which Real Estate companies have had the most positive sentiments recently (January 16, 2018). We will be using the 'tm.plugin.webmining' package to search up articles from Yahoo Finance.

```
library(tm.plugin.webmining)
library(magrittr)
library(dplyr)
library(purrr)
library(tidyr)
library(tidytext)
library(ggplot2)
```

```
download_articles <- function(Symbol) {
  WebCorpus(YahooFinanceSource(Symbol))
}
real_state <- read.csv("~/desktop/realestate_landlords.csv", header = TRUE)

stock_articles <- real_state[real_state$Symbol != "",c("Company", "Symbol")] %>%
  mutate(corpus = map(Symbol, download_articles))
```

```
## Warning in strptime(val, format = "%a, %d %b %Y %H:%M:%S", tz = "GMT"):
## unknown timezone 'zone/tz/2017c.1.0/zoneinfo/America/New_York'
```

Ngram Collection

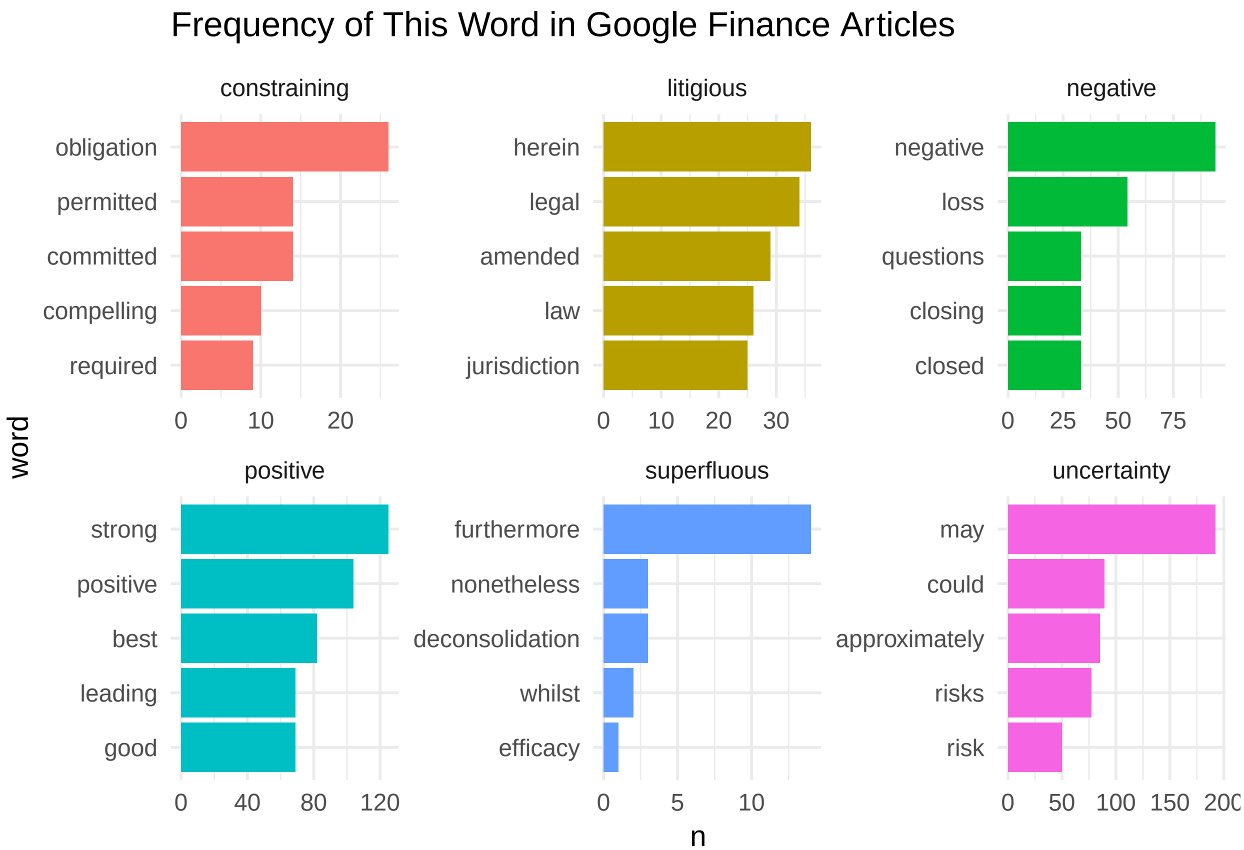
Now that we have our corpus set up we will put each individual document into an observation in a dataframe. We could keep it as a character vector but I choose to convert it to a data frame because I feel that I work with data frames better and that functions can work through them faster.

```
ngram_tokens <-unnest(stock_articles, map(corpus, tidy))
ngram_tokens <- unnest_tokens(ngram_tokens, word, text)
ngram_tokens <- select(ngram_tokens,Company, datetimestamp, word, id, heading)
```

Now that we have our unigrams we can match it with a sentiment lexicon. For finance articles in particular we use a "loughran" sentiment because it is trained to specifically avoid giving negative and positive sentiments to words that hold a neutral value in finance.

```
sentiments <- ngram_tokens %>%
  count(word) %>%
  inner_join(get_sentiments("loughran"), by = "word") %>%
  group_by(sentiment) %>%
  top_n(5, n) %>%
  ungroup() %>%
  mutate(word = reorder(word, n))

ggplot(data = sentiments, aes(word, n, fill = sentiment)) +
  geom_col() +
  coord_flip() +
  facet_wrap(~ sentiment, scales = "free") +
  ggtitle("Frequency of This Word in Google Finance Articles") + theme_minimal() +
  theme(legend.position = "none")
```



In the plot below shows us what are the most common words in each category of the "loughran" sentiment. It may not always be important for us to know which one of these categories the words are from, but as an example if a company is in legal trouble we may see that it has more words in the litigous category. For our purposes though we just want to know whether the company is generally being talked about more positively or negatively. Below I am going to call 'inner_join' from the dplyr package and pipe it straight into count so we can see how many times each sentiment occurs for each company.

```
company_sentiment_freq <- ngram_tokens %>%
  inner_join(get_sentiments("loughran"), by = "word") %>%
  count(sentiment, Company) %>%
  spread(sentiment, n, fill = 0)
company_sentiment_freq
```

```
## # A tibble: 16 x 7
##                             Company constraining litigious negative positive
##   *                          <fctr>        <dbl>     <dbl>    <dbl>    <dbl>
##   1                      Blackstone           11        52       60       64
##   2          Brixmor Property Group            7        22       89       87
##   3           Brookfield Properties            9         8       35       56
##   4     CBL & Associates Properties           10        19      112       88
##   5           DCT Industrial Trust             9        28      104      151
##   6 Developers Diversitfied Realty           19        40       98      108
##   7                    Duke Realty           14        28       73      101
##   8  First Industrial Realty Trust           12        37       70       84
##   9       General Growth Properties          20        33      118       82
##  10                    Kimco Realty            7        27       93       72
##  11          Liberty Property Trust            9        27       80      140
##  12                        Macerich            3        14       70      121
```

```
## 13                         Prologis            6         20        56        84
## 14               Simon Property Group           2          7        29        40
## 15               Vornado Realty Trust          13         36        81       111
## 16        Weingarten Realty Managment           4         35        92       111
## # ... with 2 more variables: superfluous <dbl>, uncertainty <dbl>
```

Now we can plot this data seperating out only the columns labeled 'negative' and 'positive' and then follow by creating a score for each company that is based on the amount of positive and negative words in each article. To do this we will first subtract the number of positive words from the negative words to get our numerator, the sentiment score is completley dependent on this value. After we get the numerator we divide it by the total amount of words. After plotting we get our company sentiments.

```
company_sentiment_freq %>%
  mutate(score = (positive - negative) / (positive + negative)) %>%
  mutate(company = reorder(Company, score)) %>%
  ggplot(aes(company, score, fill = score > 0)) +
  geom_col(show.legend = FALSE) + coord_flip() +
  theme_minimal() + ggtitle('Positive or Negative Scores for the Top Real Estate Companies')
```