Textmining

Team3A - Sky Liu

November 4, 2018

READ TEXT FILE

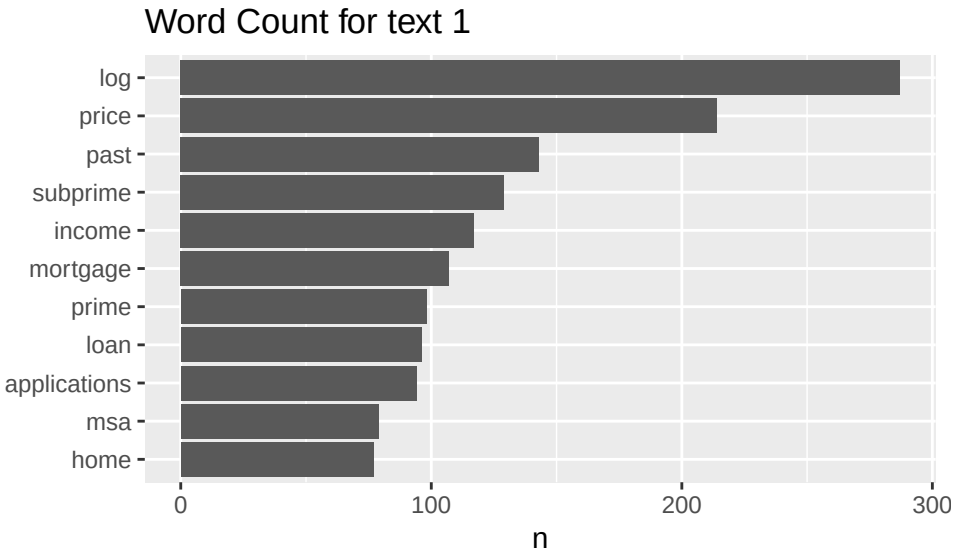
```
#get text from the blog
#read text



text2 <- read.table(file ="2008 Housing crisis.txt",header = FALSE,sep="\n")
text2 <- as.data.frame(text2)
#THE SUBPRIME CRISIS AND HOUSE PRICE APPRECIATION
text1 <- read.table(file ="The Subprime Crisis and House Price Appreciation.txt",header = FALSE,sep="\n
```

GET A TIDY TEXT FORMAT & WORD COUNT

```
#remove empty lines
text1 <- text1 %>% filter(V1 != " ")
text1 <- text1 %>% filter(V1 != "\f")
colnames(text1) <- "text"
line <- c(1:length(text1))
text1 <- cbind(line,text1)
text1$text<-as.character(text1$text)
#a token per row
text1 <-text1 %>%unnest_tokens(word,text)
#get rid of any non-characters
text1 <- text1 %>%mutate(word = str_extract(word,"[a-z']+"))
text1 <-na.omit(text1)
#get rid of stop-words
text1<- text1 %>% anti_join(stop_words)
```

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

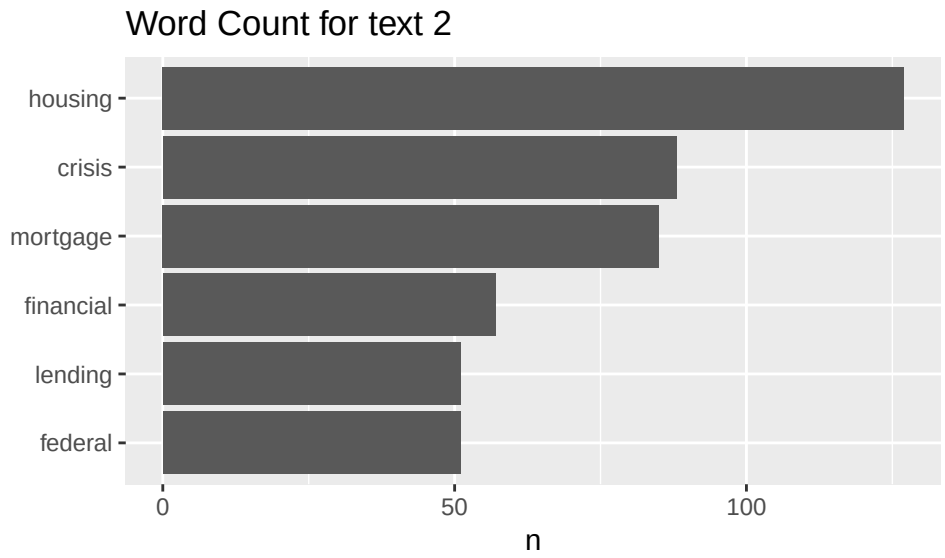
```
s<-stop_words
#word count
text1 %>%
count(word, sort = TRUE)%>%
  filter(n > 75) %>%
  mutate(word = reorder(word, n))%>%
  ggplot(aes(word, n)) +geom_col() +xlab(NULL) +coord_flip() +ggtitle("Word Count for text 1")
```

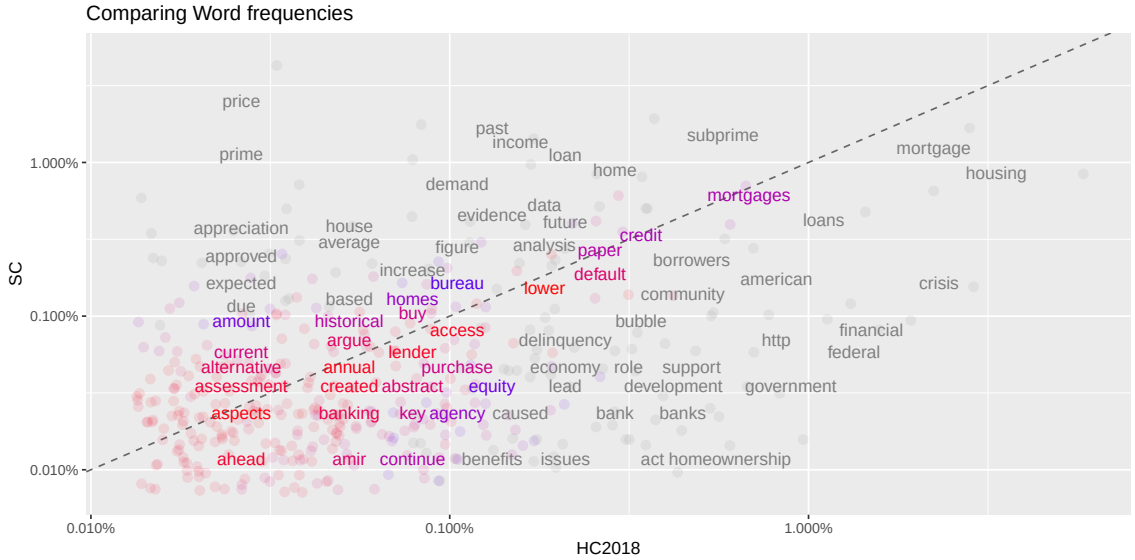



```
#remove empty lines
text2 <- text2 %>% filter(text2 != " ")
text2 <- text2 %>% filter(text2 != "\f")
colnames(text2) <- "text"
line <- c(1:length(text2))
text2 <- cbind(line,text2)
text2$text<-as.character(text2$text)
#a token per row
text2 <-text2 %>%unnest_tokens(word,text)
#get rid of any non-characters
text2 <- text2 %>%mutate(word = str_extract(word,"[a-z']+"))
text2 <-na.omit(text2)
#get rid of stop-words
text2<- text2 %>% anti_join(stop_words)
```

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

```
s<-stop_words
#word count
text2 %>%
count(word, sort = TRUE)%>%
  filter(n > 50) %>%
  mutate(word = reorder(word, n))%>%
  ggplot(aes(word, n)) +geom_col() +xlab(NULL) +coord_flip() +ggtitle("Word Count for text 2")
```





From this plot we can see that "mortgages" is frequently used in both papers. The differences is: "The Subprime Crisis and House Price Appreciation" focuses more on housing price appreciation while "The 2008 Housing Crisis" focuses more on federal/government policy.

Word Cloud

Word Cloud of "Subprime Crisis"

```
text1 %>%
  anti_join(stop_words) %>%
  count(word) %>%
  with(wordcloud(word, n, max.words = 100))
```

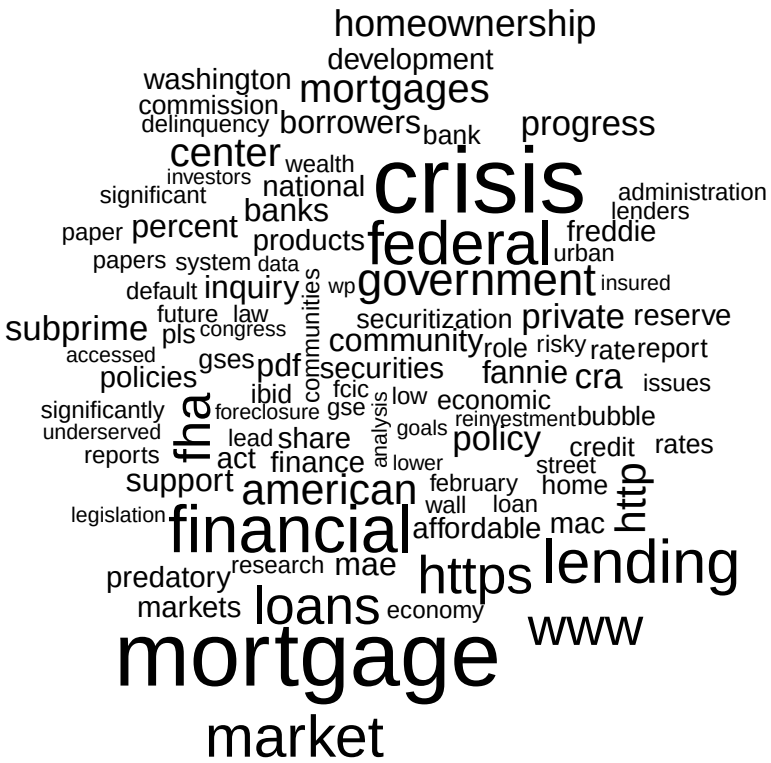
```
## Joining, by = "word"
```

Word Cloud of "Housing Crisis 2018"

```
text2 %>%
  anti_join(stop_words) %>%
  count(word) %>%
  with(wordcloud(word, n, max.words = 100))
```

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

```
## Warning in wordcloud(word, n, max.words = 100): housing could not be fit on
## page. It will not be plotted.
```



Sentiment Analysis

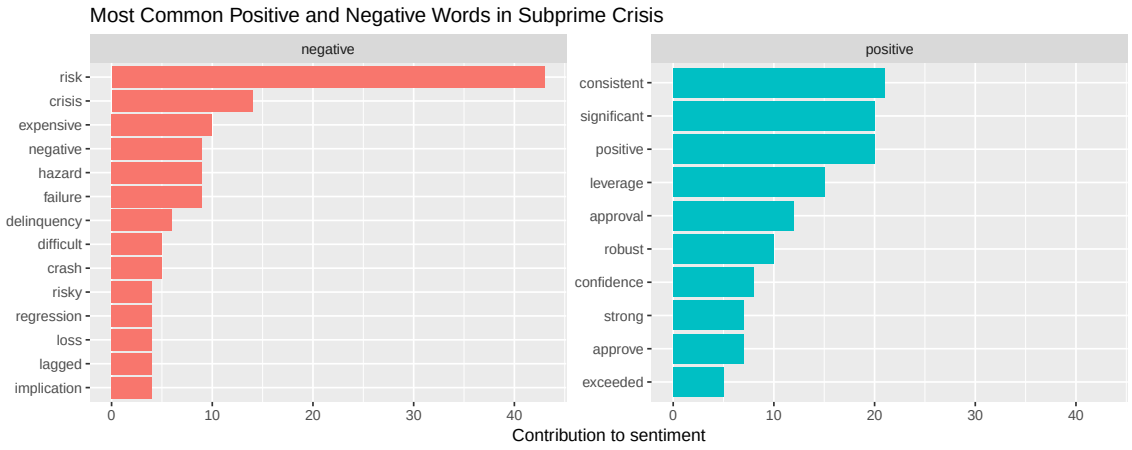
Most common positive and negative words

```
bing_word_counts <- text1 %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort = TRUE) %>%
  ungroup()
```

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

```
bing_word_counts %>%
  group_by(sentiment) %>%
  top_n(10) %>%
  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 = "Contribution to sentiment",
       x = NULL) +
  coord_flip() + ggtitle("Most Common Positive and Negative Words in Subprime Crisis")
```

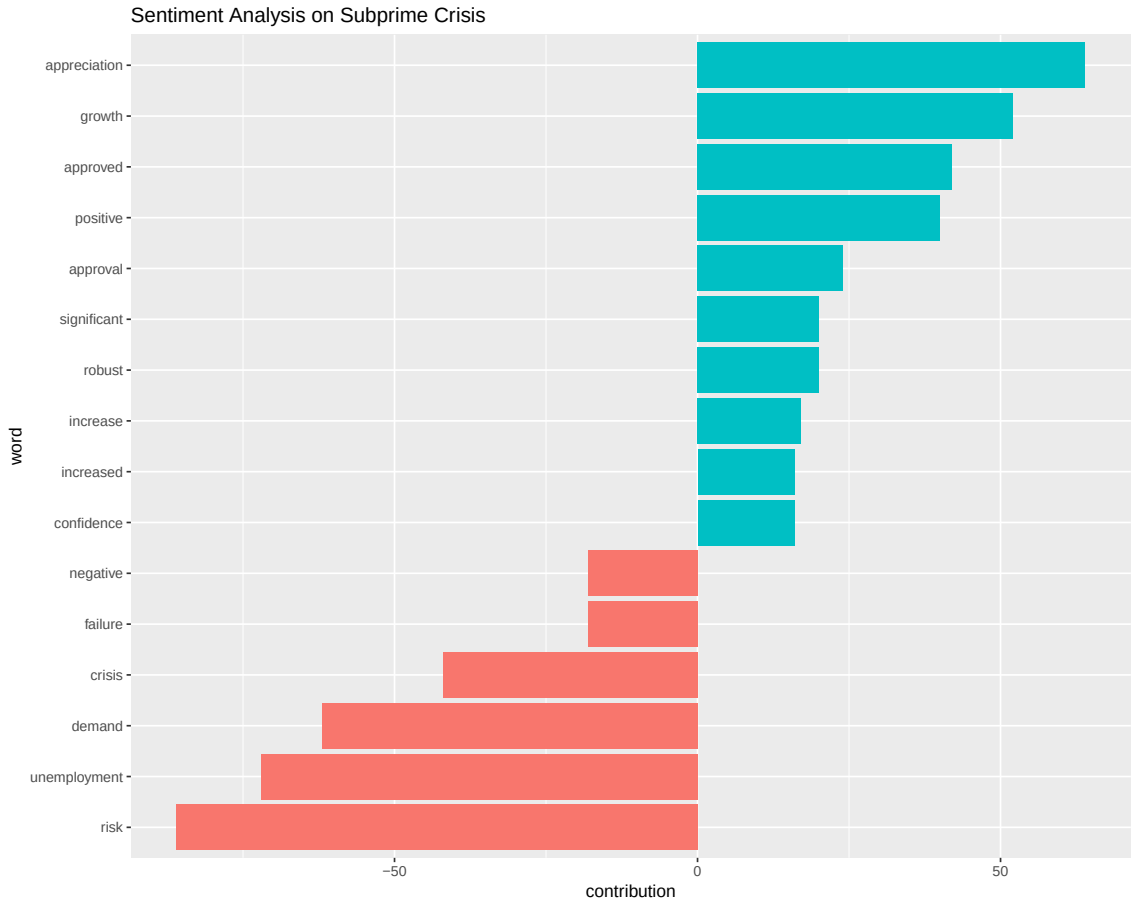
```
## Selecting by n
```



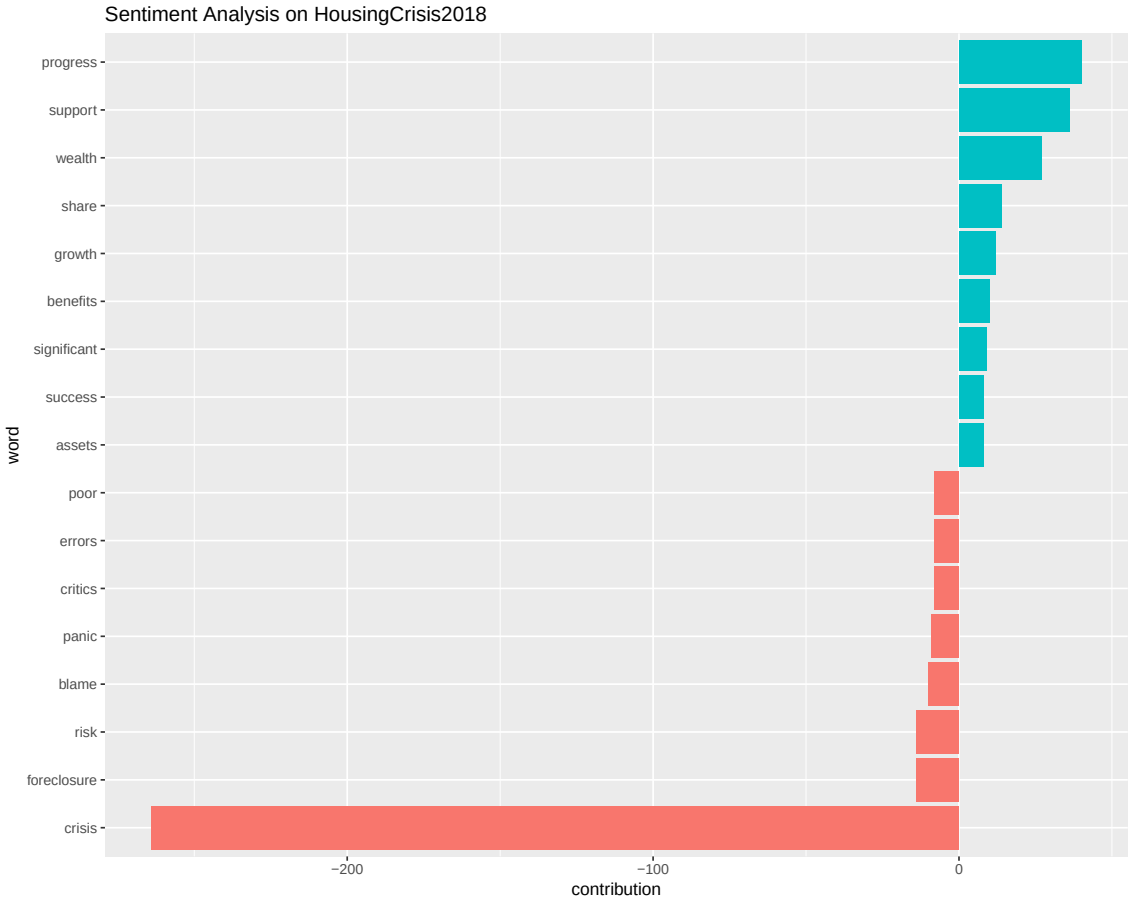
Sentiment Contribution

```
contributions1 <- text1 %>%
  inner_join(get_sentiments("afinn"), by = "word") %>%
  group_by(word) %>%
  summarize(occurences = n(),
            contribution = sum(score))

contributions1 %>%
  top_n(15, abs(contribution)) %>%
  mutate(word = reorder(word, contribution)) %>%
  ggplot(aes(word, contribution, fill = contribution > 0)) +
```

```
  geom_col(show.legend = FALSE) +
  coord_flip() + ggtitle("Sentiment Analysis on Subprime Crisis ")
```




```
contributions2 <- text2 %>%
  inner_join(get_sentiments("afinn"), by = "word") %>%
  group_by(word) %>%
  summarize(occurences = n(),
            contribution = sum(score))

contributions2 %>%
  top_n(15, abs(contribution)) %>%
  mutate(word = reorder(word, contribution)) %>%
  ggplot(aes(word, contribution, fill = contribution > 0)) +
  geom_col(show.legend = FALSE) +
  coord_flip() + ggtitle("Sentiment Analysis on HousingCrisis2018 ")
```



Because "crisis" appears much more often in "HousingCrisis2018" than in "Subprime Crisis", it takes the most nagetive sentiment contribution in the paper, while in "Subprime Crisis", the effects of "risk" and "unemployment" are more dominant.

Bigram sentiment Analysis

```
#taking bigram and filtering out non character and OAs
text11<-read.table(file ="The Subprime Crisis and House Price Appreciation.txt",header = FALSE,sep="\n"

text11 <- text11 %>% filter(V1 != " ")
text11 <- text11 %>% filter(V1 != "\f")
colnames(text11) <- "text"
line <- c(1:length(text11))
text11 <- cbind(line,text11)
text11$text<-as.character(text11$text)

text11<- text11%>%unnest_tokens(bigram, text, token = "ngrams", n = 2)%>%
  separate(bigram, c("word1", "word2"), sep = " ") %>%
  mutate(word1 = str_extract(word1,"[a-z']+"))%>%
  na.omit(word1)%>%
  mutate(word2 = str_extract(word2,"[a-z']+"))%>%
  na.omit(word2)

negation_words <- c("not", "no", "never", "without","don't")
```

```r
#filter out bigrams starts with negation words
negation_words <- text11 %>%
  filter(word1 %in% negation_words) %>%
  inner_join(get_sentiments("afinn"), by = c(word2 = "word")) %>%
  count(word1, word2, score, sort = TRUE) %>%
  ungroup()


#plot negation words
 negation_words %>%
  mutate(contribution = n * score) %>%
  arrange(desc(abs(contribution))) %>%
  head(20) %>%
  mutate(word2 = reorder(word2, contribution)) %>%
  ggplot(aes(word2, n * score, fill = n * score > 0)) +
  geom_col(show.legend = FALSE) +
  xlab("Words preceded by \"not\"") +
  ylab("Sentiment score * number of occurrences") +
  coord_flip()
```