Assignment 4: Text Analysis

By Riya Gupta (17BCS026)

Libraries used:

library(rtweet)

- library(ggplot2)
- library(dplyr)
- library(tidytext)
- library(igraph)
- library(ggraph)
- library(widyr)
- library(tidyr)
- library(textdata)

Dataset: Dataset from Kaggle

Tweets from Twitter API with search word = 'demonetization'

```
data <- read.csv("C:\\Users\\Riya\\Desktop\\text analysis\\tweets.csv")
data <- data[,-c(1:2)]
names(data)[5] <- 'created_at'
names(data)[11] <- 'screen_name'
data <- data[, c('text','screen_name','created_at')]
col <- colnames(data)
```

Getting tweets:

```
> dataAPI <- search_tweets(q = "demonetization", n = 10000,
+                          lang = "en",
+                          include_rts = FALSE)
Downloading [=>---------------------------------------]   4%
```

Joining the 2 datasets:

demonetization <- rbind(data,dataAPI)

Total = 16731 tweets (14940 from Kaggle, 1791 from TwitterAPI)

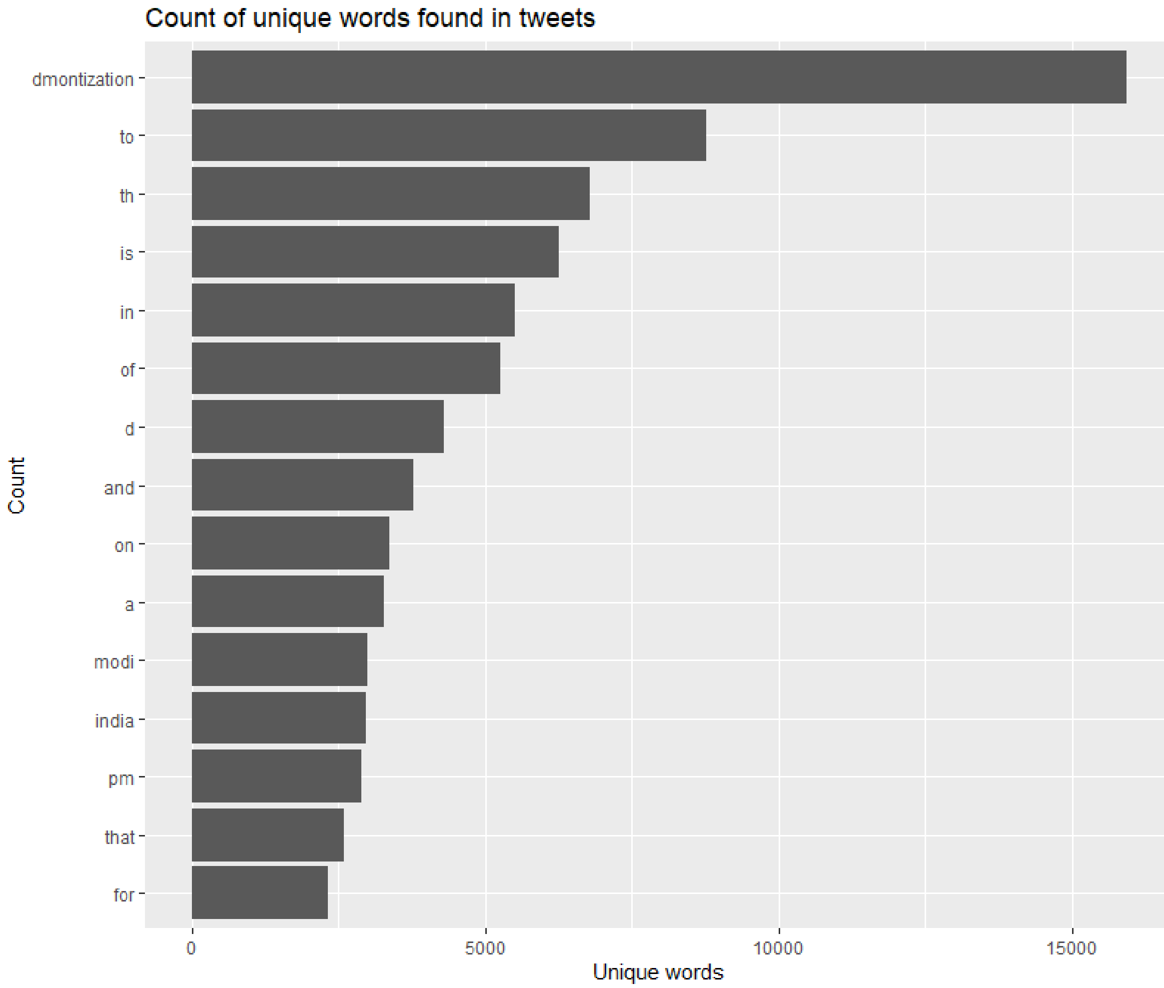
Cleaning dataset:

```
demonetization$stripped_text <- gsub("http.*","",  demonetization$text)
demonetization$stripped_text <- gsub("https.*","", demonetization$stripped_text)
demonetization$stripped_text <- gsub("ed*","", demonetization$stripped_text)
demonetization$stripped_text <- gsub("00B8*","", demonetization$stripped_text)
demonetization$stripped_text <- gsub("00A0*","", demonetization$stripped_text)
demonetization$stripped_text <- gsub("00BD*","", demonetization$stripped_text)
demonetization$stripped_text <- gsub("RT*","", demonetization$stripped_text)
demonetization$stripped_text <- gsub("U","", demonetization$stripped_text)
demonetization$stripped_text <- gsub('[[:digit:]]+',"", demonetization$stripped_text)
```

Removing punctuation, converting to lowercase and adding id for all tweets

```
demonetization_clean <- demonetization %>%
```

```
dplyr::select(stripped_text) %>%
unnest_tokens(word, stripped_text)
```



Stop words:

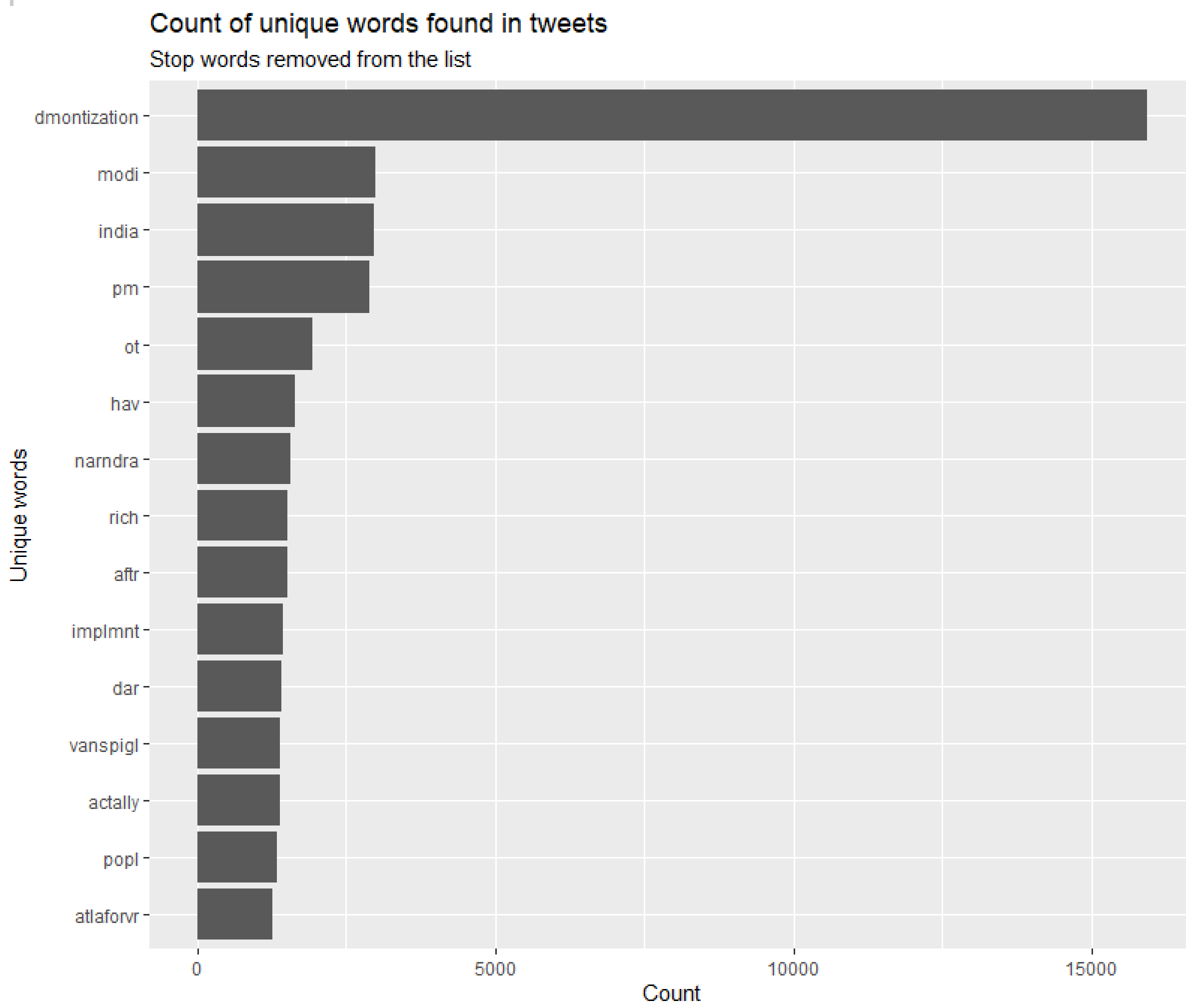
data("stop_words")

1149 stop words found

	word	lexicon
737	nor	snowball
738	not	snowball
739	only	snowball
740	own	snowball
741	same	snowball
742	so	snowball
743	than	snowball
744	too	snowball
745	very	snowball
746	a	onix
747	about	onix
748	above	onix
749	across	onix
750	after	onix
751	again	onix
752	against	onix
753	all	onix

Removing stop words from our word list:

```
> nrow(demonetization)
[1] 16731
> nrow(demonetization_clean)
[1] 352674
> cleaned_tweet_words <- demonetization_clean %>%
+   anti_join(stop_words)
Joining, by = "word"
> nrow(cleaned_tweet_words)
[1] 202241
```



Exploring Network of words:

```
> demo_tweets_paired_words %>%
+   count(paired_words, sort = TRUE)
# A tibble: 75,986 x 2
   paired_words        n
   <chr>           <int>
 1 d d              2131
 2 india is         1671
 3 dmontization to  1608
 4 narndra modi     1573
 5 had to           1450
 6 is so            1449
 7 so rich          1434
 8 that pm          1426
 9 pm narndra       1424
10 modi had         1423
# ... with 75,976 more rows
```

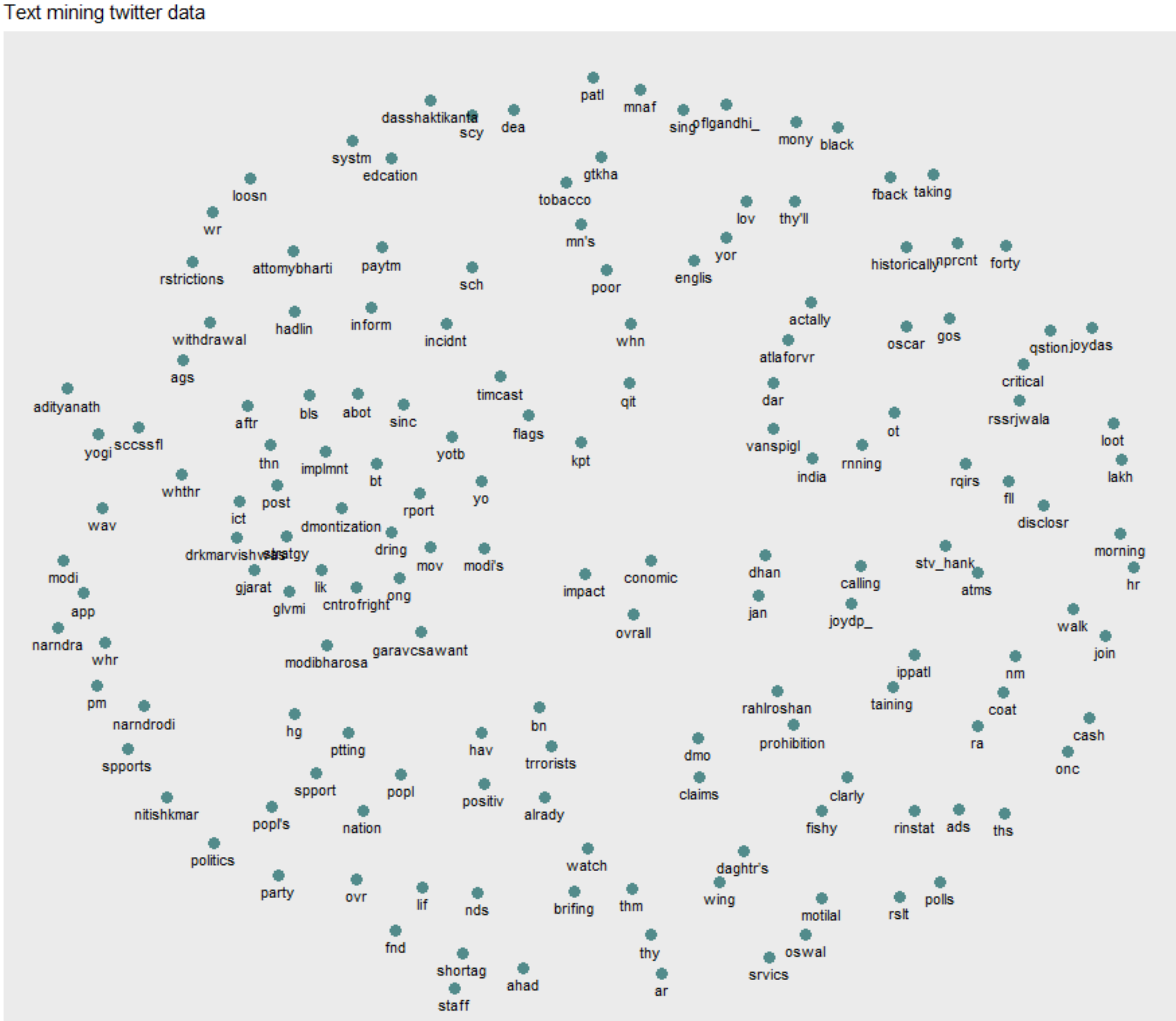
```
> demo_tweets_filtered <- demo_tweets_separated_words %>%
+   filter(!word1 %in% stop_words$word) %>%
+   filter(!word2 %in% stop_words$word)
> demo_words_counts <- demo_tweets_filtered %>%
+   count(word1, word2, sort = TRUE)
> head(demo_words_counts)
# A tibble: 6 x 3
  word1     word2            n
  <chr>     <chr>        <int>
1 narndra   modi          1573
2 pm        narndra       1424
3 implmnt   dmontization  1409
4 vanspigl  india         1370
5 dar       vanspigl      1366
6 atlaforvr dar           1273
```

Below is the graph of words found in the tweets talking about demonetization. These are the topics related to demonetization being talked about.

The words clubbed nearby in the graph are more related to each other.

Word Network: Tweets using the search word - demonetization

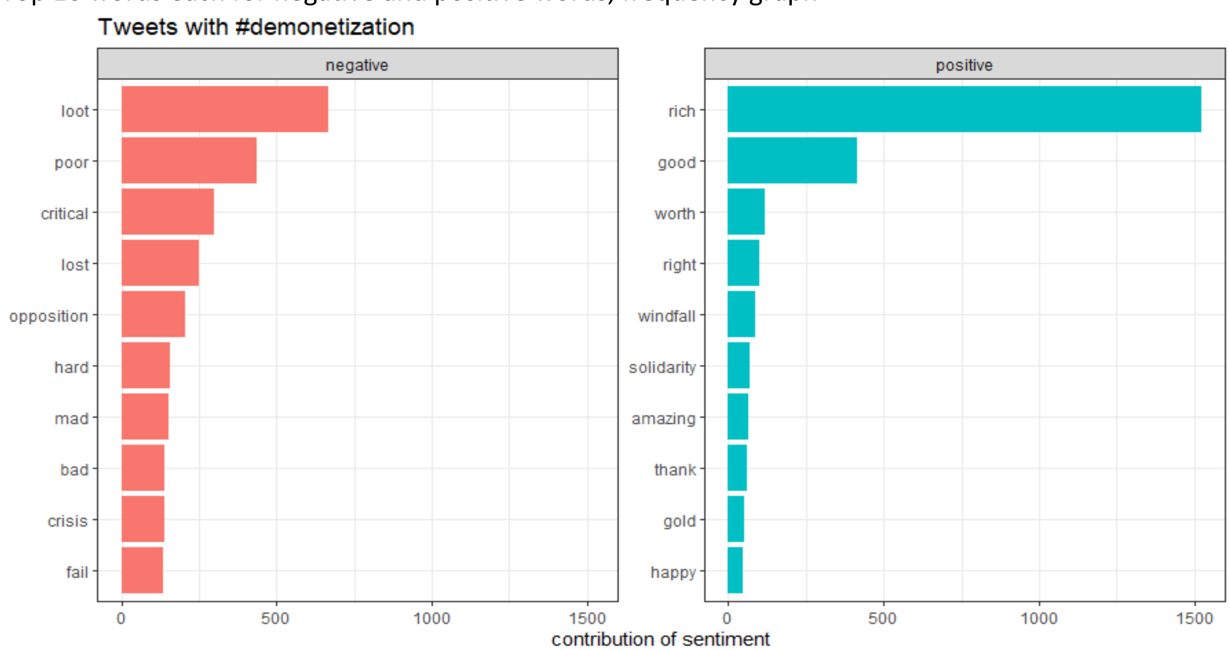
Text mining twitter data



Sentiment Analysis

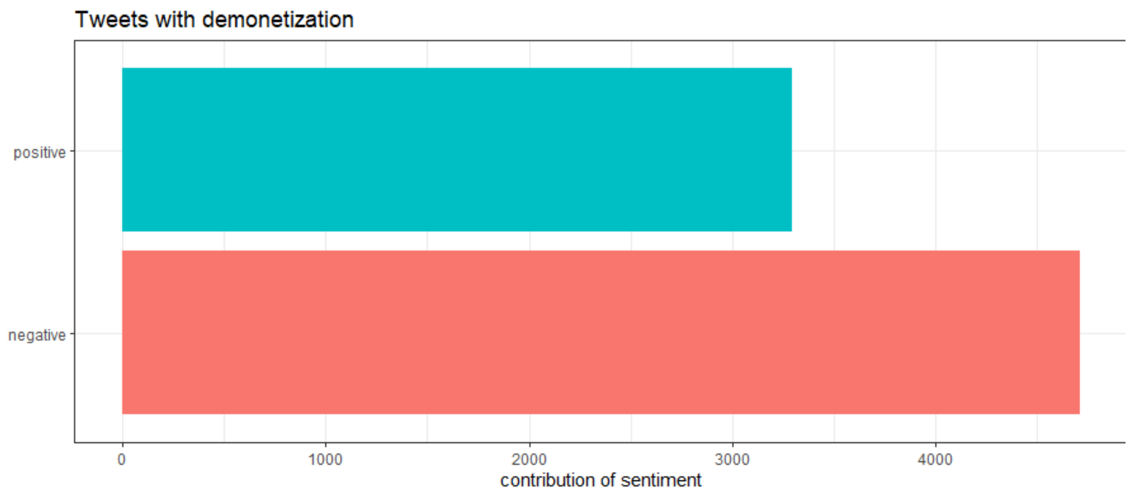
```
# A tibble: 441 x 3
   word       sentiment     n
   <chr>      <chr>     <int>
 1 rich       positive   1524
 2 loot       negative    667
 3 poor       negative    435
 4 good       positive    413
 5 critical   negative    298
 6 lost       negative    251
 7 opposition negative    206
 8 hard       negative    155
 9 mad        negative    151
10 bad        negative    139
# ... with 431 more rows
```

Top 10 words each for negative and positive words, frequency graph




```
Selecting by n
> positive <- demonetization_clean %>%
+   inner_join(get_sentiments("bing") %>% filter(sentiment=='positive')) %>%
+   count(word, sentiment, sort = TRUE) %>%
+   ungroup()
Joining, by = "word"
> negative <- demonetization_clean %>%
+   inner_join(get_sentiments("bing") %>% filter(sentiment=='negative')) %>%
+   count(word, sentiment, sort = TRUE) %>%
+   ungroup()
Joining, by = "word"
> print(sum(positive[3]))
[1] 3292
> print(sum(negative[3]))
[1] 4714
```

The total frequency of negative word sentiments is more than positive ones. Therefore, we can say that a greater number of people are talking negative about demonetization.



Top 50 words each