Analysis of Presidential Tweets

Sindhya Balasubramanian

Understanding the top 20 most common terms in Donald Trump's tweets.

Steps followed -

- Imported Donald Trump tweets
- Cleaned the text format and converted it to a tidy format for analysis
- Removed re-tweets from data
- Removed tweets without any spaces
- Removed stop words and "&amp" from keywords
- Removed variations on Trump's name
- Removed URLs and twitter usernames

```
library(tokenizers)
```

```
## Warning: package 'tokenizers' was built under R version 4.1.2
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tidytext)
```

```
## Warning: package 'tidytext' was built under R version 4.1.2
```

```
options(scipen=999)
df = read.csv("C:/Users/Shivapriya B/Desktop/NEU Work/Classwork/Semester 1/IDMP/Week 9/twitter/twitter/

df_filter = df %>% filter(isRetweet != "t") %>%
  filter(grepl(" ", text))

tweets = df_filter$text
tweets_tidy = unnest_tokens(df_filter, output="word", token = "tweets", input=text)
```

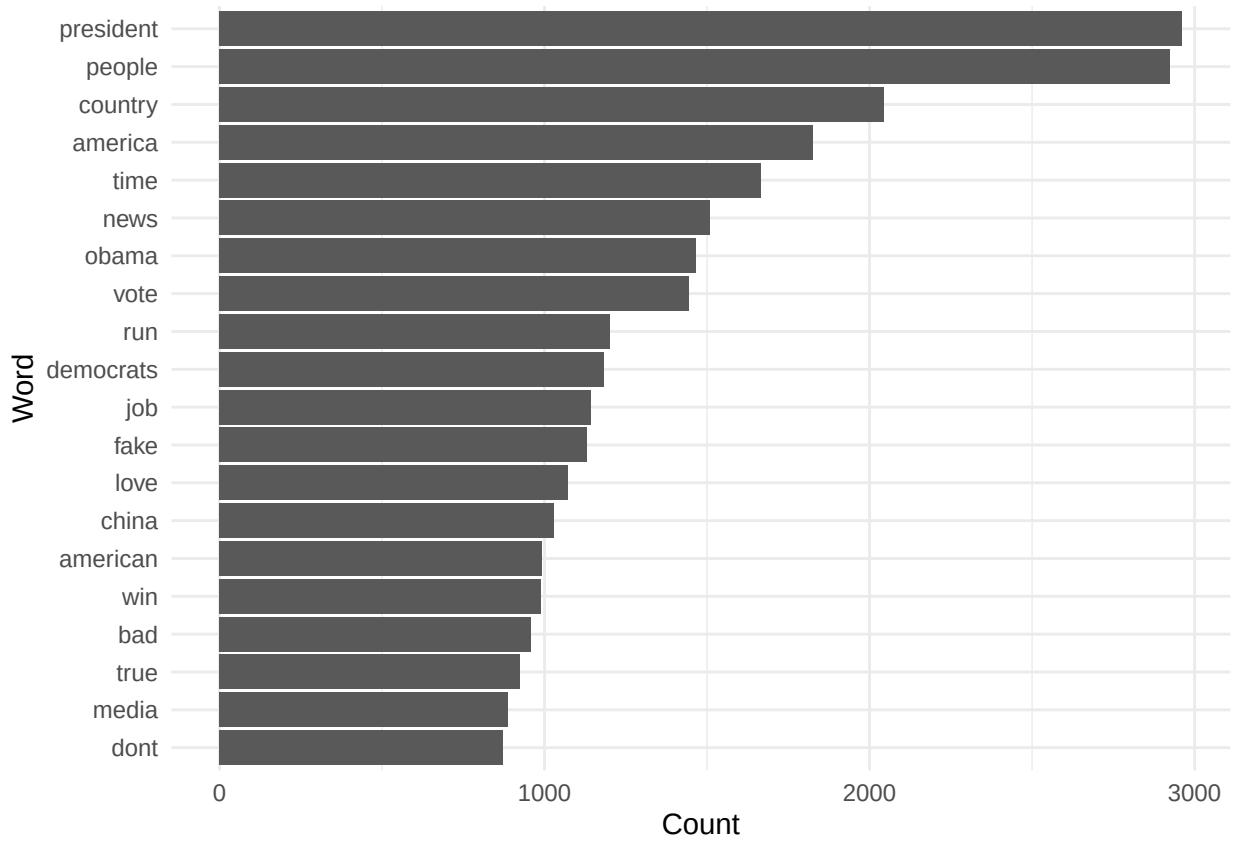
```
## Using `to_lower = TRUE` with `token = 'tweets'` may not preserve URLs.
```

```
tweets_tidy2 <- anti_join(tweets_tidy, stop_words, by="word") %>%
  filter(word != "amp") %>%
  filter(!grepl("donald", word)) %>%
  filter(!grepl("trump", word)) %>%
  filter(!grepl(".com", word)) %>%
```

```
  filter(!grepl("https://", word)) %>%
  filter(!grepl("http://", word)) %>%
  filter(!grepl("@", word))

tweets_tidy2 %>%
  count(word, sort=TRUE) %>%
  top_n(20) %>%
  ggplot(aes(x=reorder(word, n), y=n)) +
  geom_col() +
  coord_flip() +
  labs(x="Word", y="Count") +
  theme_minimal()
```

```
## Selecting by n
```



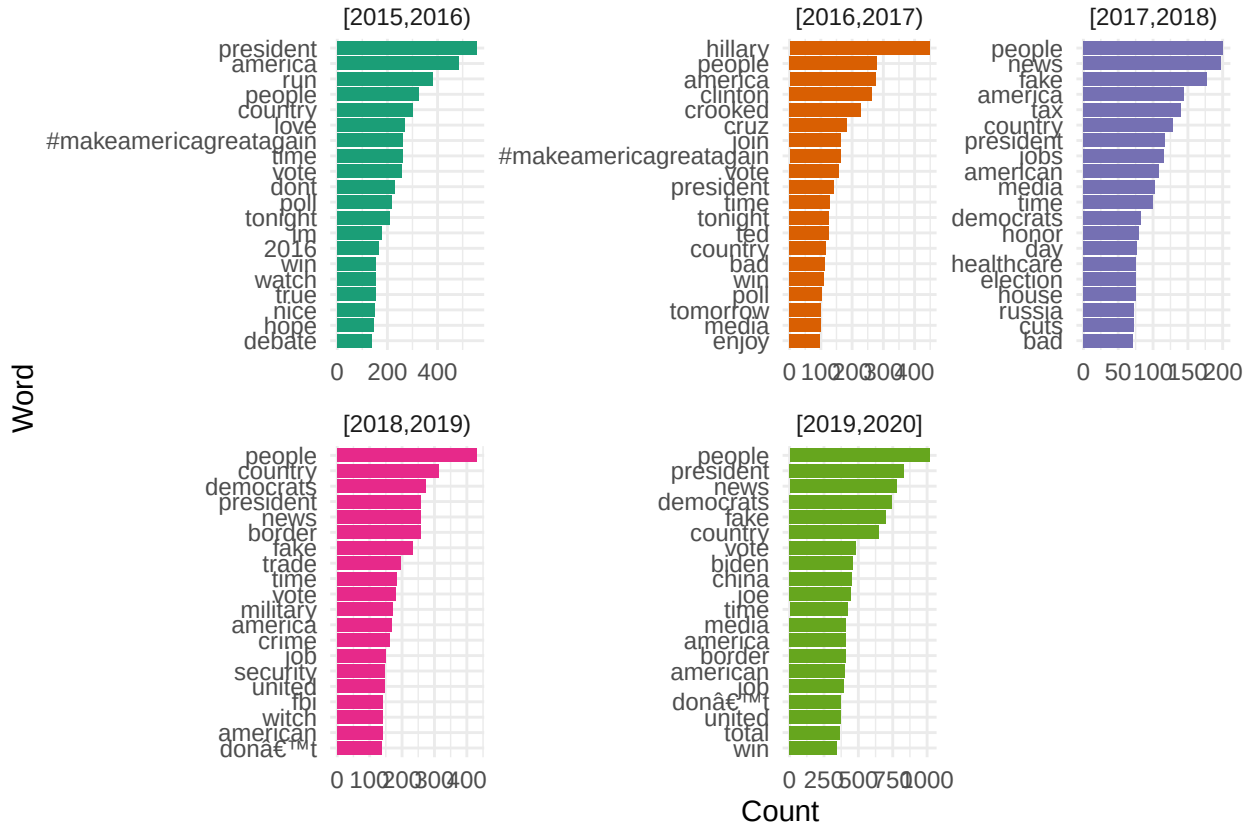
Visualizing the top 20 most common terms in Donald Trump's tweets for each year from 2015-2020.

```
library(stringr)
tweets_tidy3 = mutate(tweets_tidy2,
                      year = substr(date, 1, 4),
                      period = cut_width(year, 1,
                                         boundary = 0,
                                         closed = "left",
                                         dig.lab = -1))

tweets_tidy3 %>%
```

```
  filter(year %in% c("2015", "2016", "2017", "2018", "2019", "2020")) %>%
  count(word, period, sort=TRUE) %>%
  group_by(period) %>%
  top_n(20) %>%
  ggplot(aes(x=reorder_within(word, n, period), y=n, fill=period)) +
  geom_col(show.legend=FALSE) +
  facet_wrap(~period, scales="free") +
  coord_flip() +
  labs(x="Word", y="Count",
       fill="Time Period") +
  scale_fill_brewer(palette="Dark2") +
  scale_x_reordered() +
  theme_minimal()
```

```
## Selecting by n
```



Observations -

1. The term people slowly started taking importance from the year to 2015 to 2017 post which it is the most used term in the tweets till 2020
2. The hashtag Make America great again features in the top 10 terms in 2015 and 2016 but does not appear in the tweets post 2016
3. Some common terms we see throughout the years are as given below -
   people, president, america, time
4. In the 2016, his tweets heavily consisted of Hillary Clinton. Similarly in 2019-20, his tweets contain alot about Joe Biden

Visualizing the top 20 most characteristic terms in Donald Trump's tweets for each year from 2015-2020

```
tweets_tf = count(tweets_tidy3, year, word, sort=TRUE)

tweets_wc = tweets_tidy3 %>%
  group_by(year) %>%
  summarize(word_count=n())

tweets_tf = tweets_tf %>%
  left_join(tweets_wc) %>%
  mutate(tf = n / word_count)
```

```
## Joining, by = "year"
```

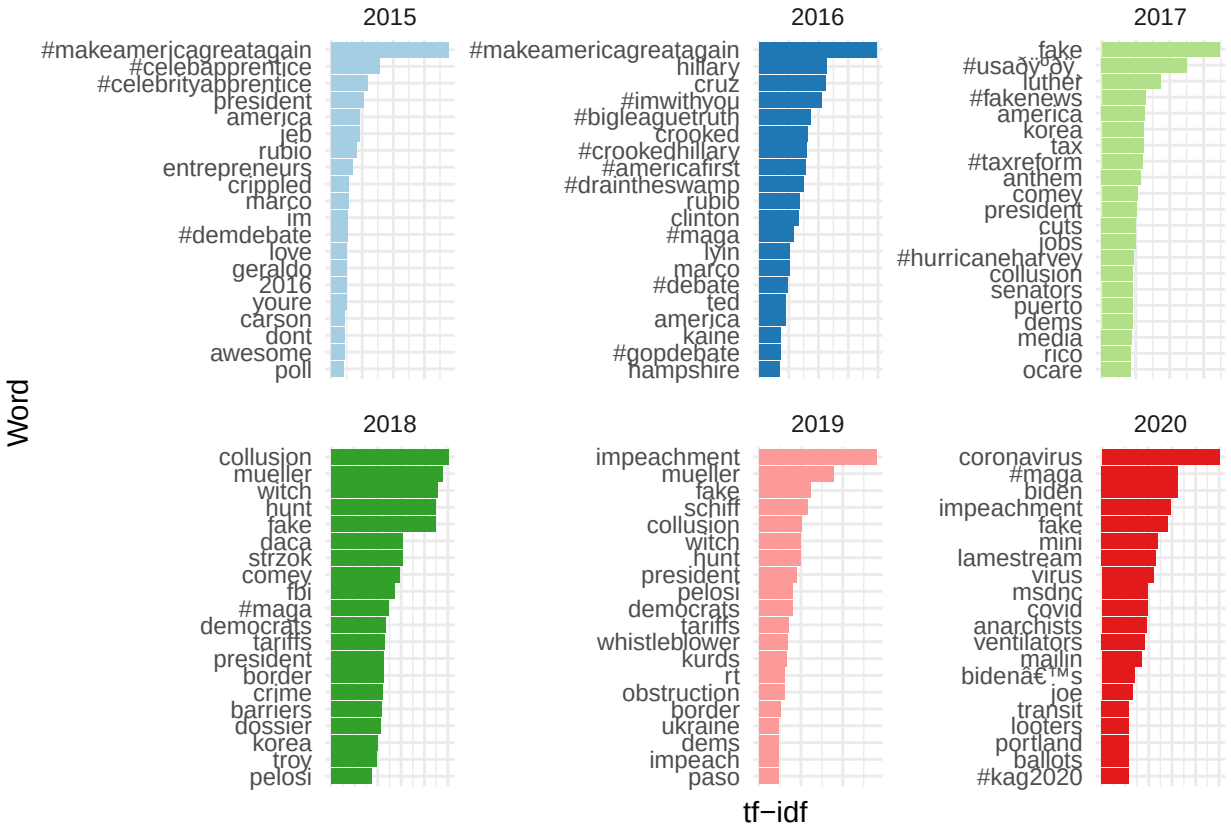
```
tweets_tf <- tweets_tf %>%
  mutate(n_doc = n_distinct(year)) %>%
  group_by(word) %>%
  mutate(idf = log(n_doc / n()),
         tf_idf = tf * idf) %>%
  select(-n_doc, -word_count) %>%
  ungroup()

tweets_tf = arrange(tweets_tf, desc(tf_idf))

tweets_tf_idf = tweets_tidy3 %>%
  count(year, word, sort=TRUE) %>%
  bind_tf_idf(term=word, document=year, n=n)

tweets_tf_idf = arrange(tweets_tf_idf, desc(tf_idf))

tweets_tf_idf %>%
  filter(year %in% c("2015", "2016", "2017", "2018", "2019", "2020")) %>%
  group_by(year) %>%
  top_n(20, wt=tf_idf) %>%
  ggplot(aes(x=reorder_within(word, tf_idf, year),
             y=tf_idf, fill=factor(year))) +
  geom_col(position="dodge", show.legend=FALSE) +
  coord_flip() +
  facet_wrap(~year, scales="free") +
  labs(x="Word", y="tf-idf",
       fill="Year") +
  scale_fill_brewer(palette="Paired") +
  scale_x_reordered() +
  scale_y_continuous(labels=NULL) +
  theme_minimal()
```



Observations - 1. Coronavirus appears extremely important in the year 2020. The word fake also is quite significant 2. In 2019, the major discussion is around impeachment and mueller 3. In 2016, significant discussion occured around make america great again and hillary 4. The most significant word used in the 2017 tweets was fake in the top 20 5. The most significant word in 2015 and 2016 is the hashtag Make America Great Again

Fitting a sparse regression model to predict the number of retweets that a tweet will get. Visualizing the terms with the strongest positive relationship with the number of re-tweets

```
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 4.1.2
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## Loaded glmnet 4.1-3
```

```
library(coefplot)
```

```
## Warning: package 'coefplot' was built under R version 4.1.2
```

```
tweets_tidy4 = tweets_tidy3 %>%
  filter(year %in% c("2016", "2017", "2018", "2019", "2020"));
```

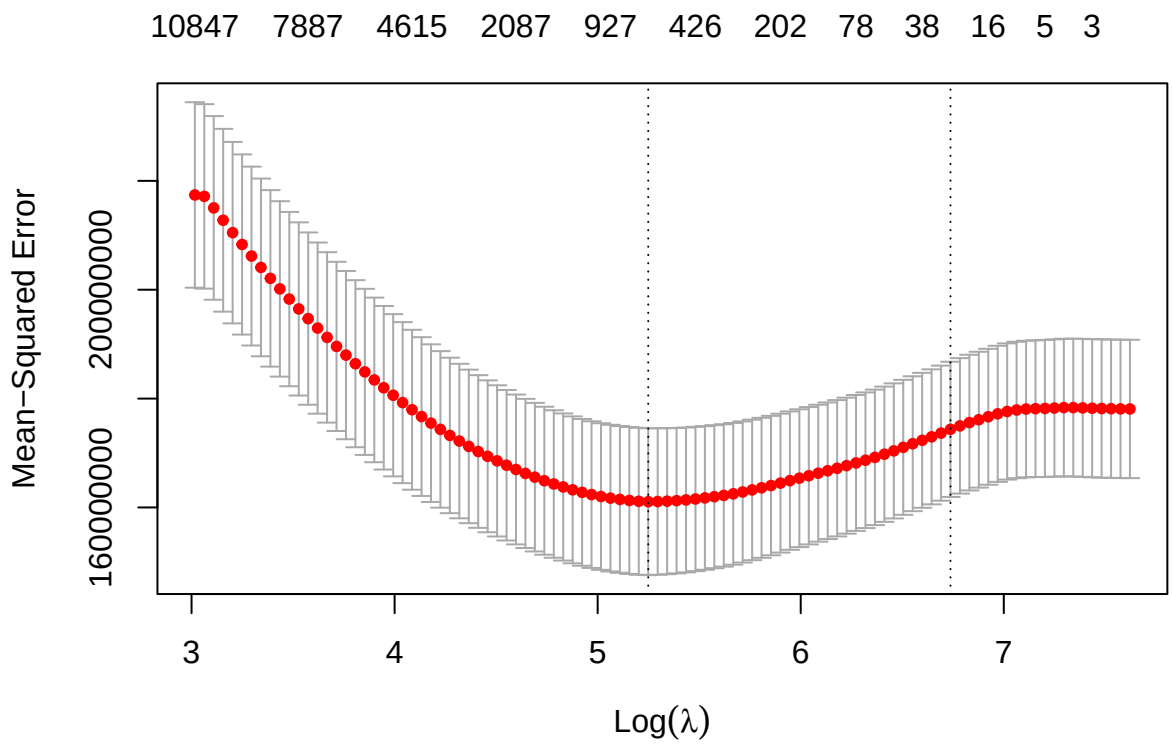
```
temp_tweet = tweets_tidy4 %>%
  count(id, word) %>%
  inner_join(tweets_tidy4, on = id) %>%
  select(id, retweets, year) %>%
  group_by(id) %>%
  distinct()
```

```
## Joining, by = c("id", "word")
```

```
tweets_dtm <- tweets_tidy4 %>%
  count(id, word) %>%
  cast_sparse(row=id, column=word, value=n)
tweets_dtm = cbind(tweets_dtm, temp_tweet$retweets)

cvfit = cv.glmnet(tweets_dtm[, 1:ncol(tweets_dtm)-1], tweets_dtm[, ncol(tweets_dtm)])

plot(cvfit)
```

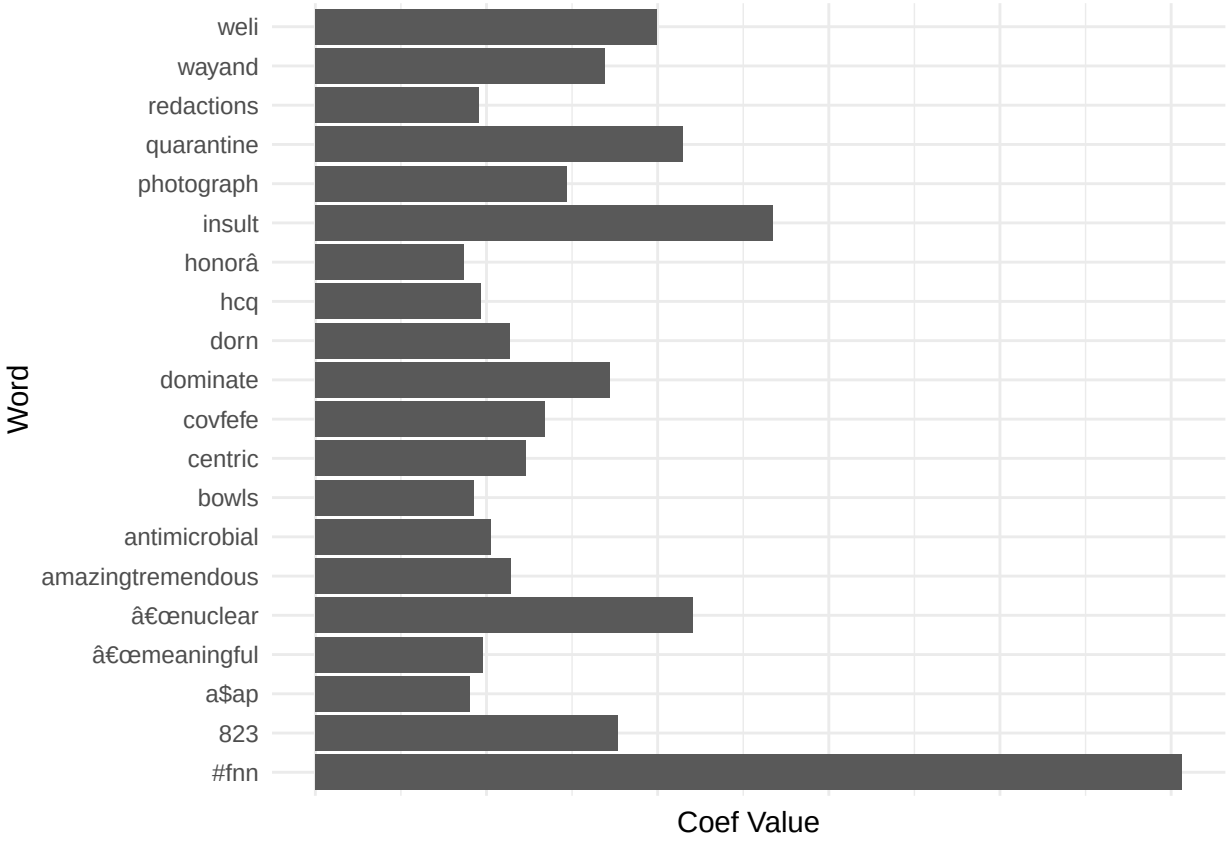



```
cvfit$lambda.min
```

```
## [1] 190.384
```

```
extract.coef(cvfit) %>%
  arrange(desc(Value)) %>%
  top_n(20, wt=Value) %>%
  ggplot(aes(x=Coefficient,
             y=Value)) +
```

```
geom_col() +
coord_flip() +
labs(x="Word", y="Coef Value") +
scale_y_continuous(labels=NULL) +
theme_minimal()
```



Observations - 1. #fnn has the highest positive coefficient noted that impacts retweets 2. Some other key terms noted here are photographs, quarantine, nuclear in the top 20 coefficients that positively impact retweets