FPOTUS_Tweets_Analysis

Shweta Mishra

15/04/2022

Library calls:

```
library(dplyr)
library(readr)
library(stringr)
```

```
## Warning: package 'stringr' was built under R version 4.1.3
```

```
library(tidytext)
```

```
## Warning: package 'tidytext' was built under R version 4.1.3
```

```
library(ggplot2)
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 4.1.3
```

```
x<-"C:/Users/HP/Documents/"
y<-"NEU_22_23/NEU_SPRING_22/"
z<-"DS 5110 - IDMP/HW6/"
w<-"realDonaldTrump-20201106.csv"
T_data<-read_csv(paste0(x,
                        y,
                        z,
                        w),
                 guess_max=1000)
```

```
## Rows: 55090 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): text, device
## dbl  (3): id, favorites, retweets
## lgl  (2): isRetweet, isDeleted
## dttm (1): date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
T_data$date<-as.Date(T_data$date,
                     format="%Y-%m-%d")
T_data$id<-format(T_data$id,
                  scientific=F)
summary(T_data)
```

```
##       id                 text              isRetweet       isDeleted
##  Length:55090       Length:55090       Mode :logical   Mode :logical
##  Class :character   Class :character   FALSE:45755     FALSE:54050
##  Mode  :character   Mode  :character   TRUE :9335      TRUE :1040
##
##
##
##     device            favorites          retweets            date
##  Length:55090       Min.   :      0   Min.   :     0   Min.   :2009-05-04
##  Class :character   1st Qu.:     11   1st Qu.:    54   1st Qu.:2014-04-07
##  Mode  :character   Median :    154   Median :  2897   Median :2016-04-17
##                     Mean   :  25573   Mean   :  7917   Mean   :2016-10-06
##                     3rd Qu.:  40914   3rd Qu.: 12312   3rd Qu.:2019-10-05
##                     Max.   :1869706   Max.   :408866   Max.   :2020-11-06
```

Removing re-tweets:

```
T_data1<-T_data%>%
  filter(isRetweet==FALSE)
```

Removing tweets without spaces:

```
T_data1<-T_data1[-which(is.na(str_locate(T_data1$text," "))),]
```

Removing URL , username , &amp , change date to year:

```
T_data1$text<-gsub("(f|ht)(tp)(s?)(://)(.*)[.|/](.*)",
                   "  ",
                   T_data1$text)
T_data1$text<-gsub("@\\w+",
                   "",
                   T_data1$text)
T_data1$text= gsub("&amp",
                   "",
                   T_data1$text)
T_data1$text <- tolower(T_data1$text)
T_data1 <- T_data1 %>%
           rename(year = date)
T_data1$year<-str_sub(T_data1$year,
                      1,
                      4)
```

Removing variation from name

```
x   <-'donald*'
donald<-str_subset(T_data1$text,x)
y   <-'trump*'
trump <- str_subset(T_data1$text,y)
T_data1$text  <-  gsub("realdonaldtrump",
                       'dt',
                       T_data1$text)
T_data1$text  <-  gsub("realdonal",
                       "dt",
                       T_data1$text)
T_data1$text  <-  gsub("donaldTrump",
                       "dt",
                       T_data1$text)
T_data1$text  <-  gsub("trump",
                       "dt",
                       T_data1$text)
T_data1$text  <-  gsub("donald",
                       "dt",
                       T_data1$text)
T_data1$text  <-  gsub("trump",
                       "dt",
                       T_data1$text)
T_data1$text  <-  gsub("donaldtrump",
                       "dt",
                       T_data1$text)
T_data1$text  <-  gsub("donal",
                       "dt",
                       T_data1$text)
T_data1$text  <-  gsub("donaldTrump",
                       "dt",
                       T_data1$text)
T_data1$text  <-  gsub("DonaldTrump",
                       "dt",
                       T_data1$text)
T_data1$text  <-  gsub("dt",
                       "",
                       T_data1$text)
T_data1 <- T_data1[!(T_data1$text == ""), ]
```

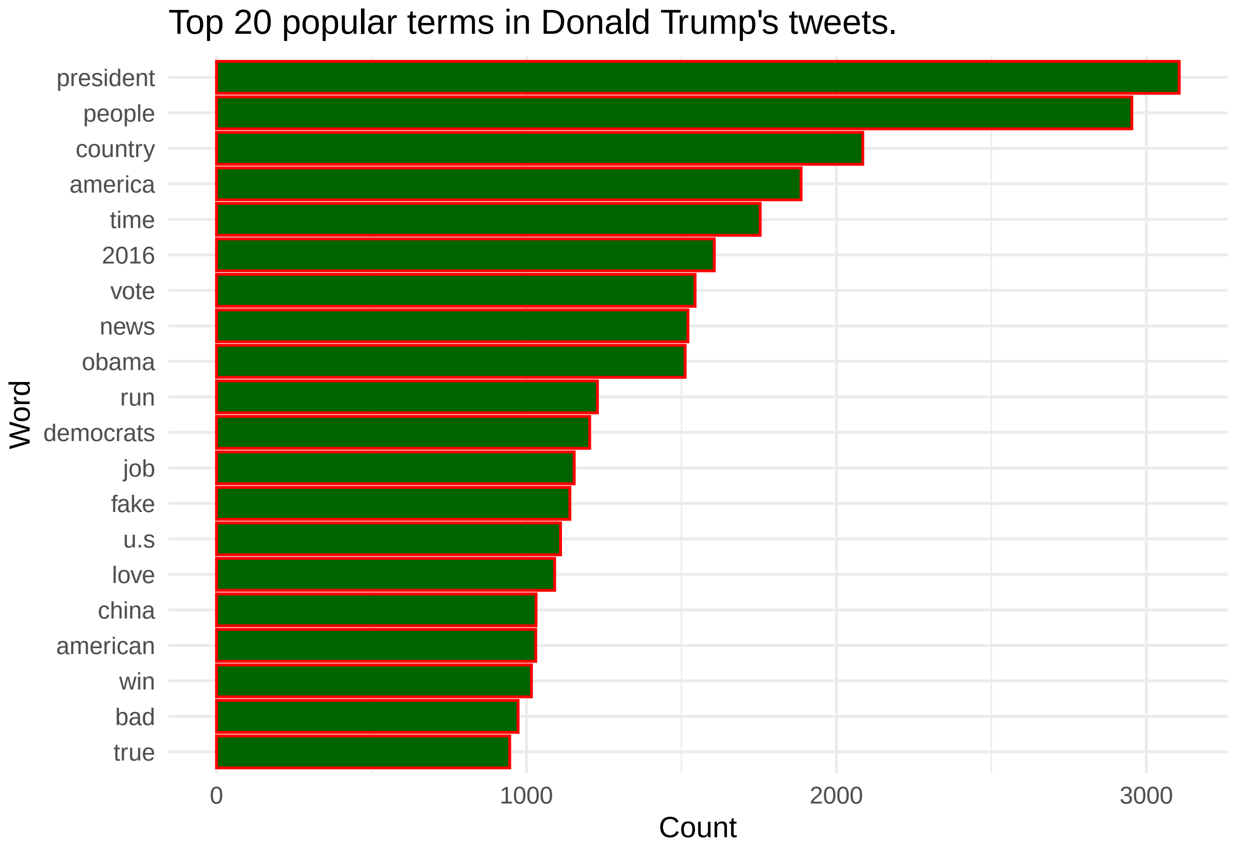
Removing stop words

```
T_data1<-unest_tokens(T_data1,
                       output="word",
                       input=text)
T_data1<-anti_join(T_data1,
                   stop_words,
                   by="word")
```

Top 20 words

```
T_data1 %>%
count(word,
      sort=TRUE)  %>%
top_n(20) %>%
ggplot(aes(x=reorder(word,
                     n),
           y=n,
           fill = word)) +
geom_col(color = "red",
         fill="darkGreen",
         show.legend=FALSE) +
coord_flip() +
labs(x="Word",
     y="Count",
title="Top 20 popular terms in Donald Trump's tweets.") +
theme_minimal()
```

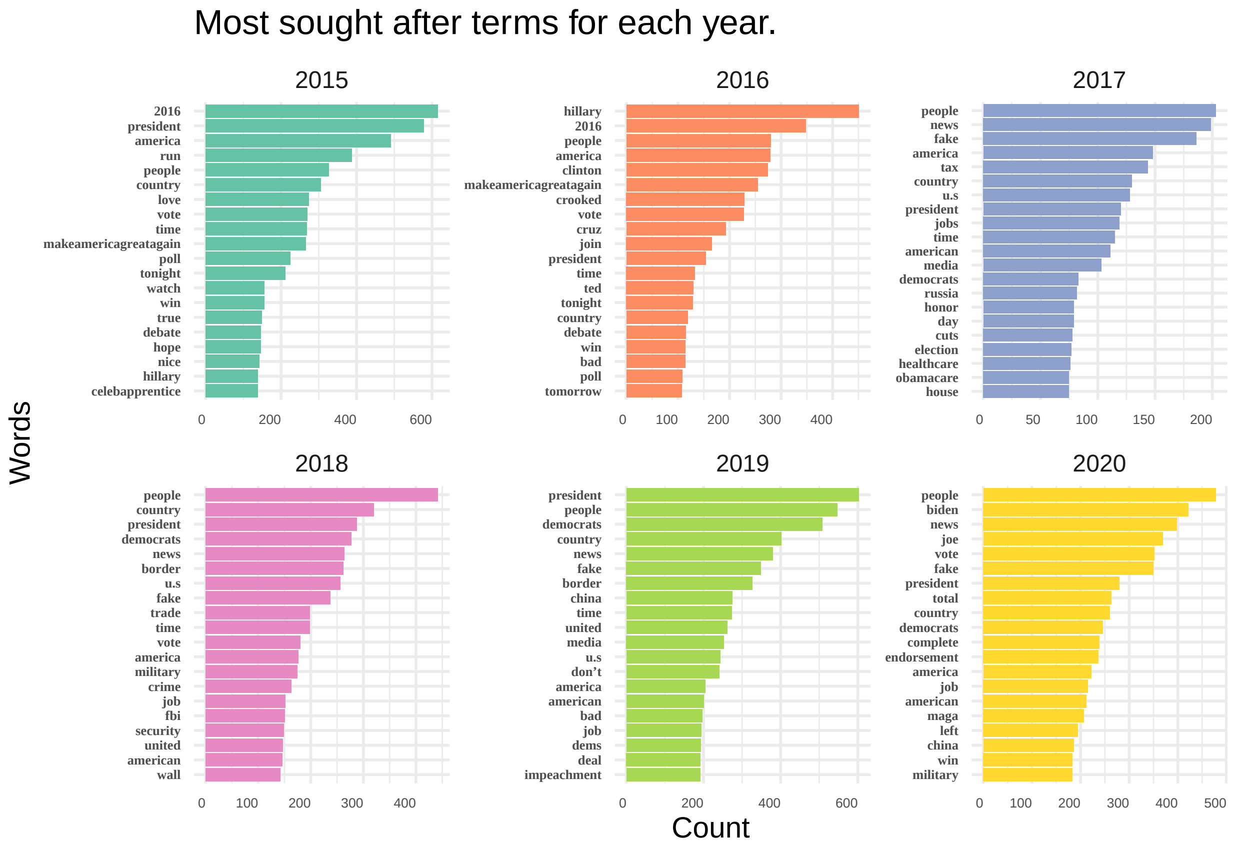
```
## Selecting by n
```



The word "president" appears frequently in Donald Trump's tweets.

```
T_data1 <-  T_data1  %>%
filter(year >= 2015)
T_data1 %>%
count(word,
      year,
      sort=TRUE) %>%
group_by(year) %>%
top_n(20) %>%
ggplot(aes(x=reorder_within(word,
                            n,
                            year),
           y=n,
           fill=year)) +
geom_col(show.legend=FALSE) +
facet_wrap(~year,
           scales="free") +
coord_flip() +
labs(x="Words",
     y="Count",
     title="Most sought after terms for each year.",
     fill="Year") +
scale_fill_brewer(palette="Set2") +
scale_x_reordered() +
theme_minimal()+
theme(axis.text.x = element_text(angle=0,
                                 hjust=1,
                                 vjust=0.5,
                                 size=5))+
theme(axis.text.y=element_text(angle=0,
                               hjust=1,
                               vjust=0.5,
                               size=5,
                               face="bold",
                               family = "Times"))
```

```
## Selecting by n
```



“People” has indeed been the most commonly used word for most of the years.

We discovered “2016” to be the most widely used word in 2015 due to the presidential elections in 2016.

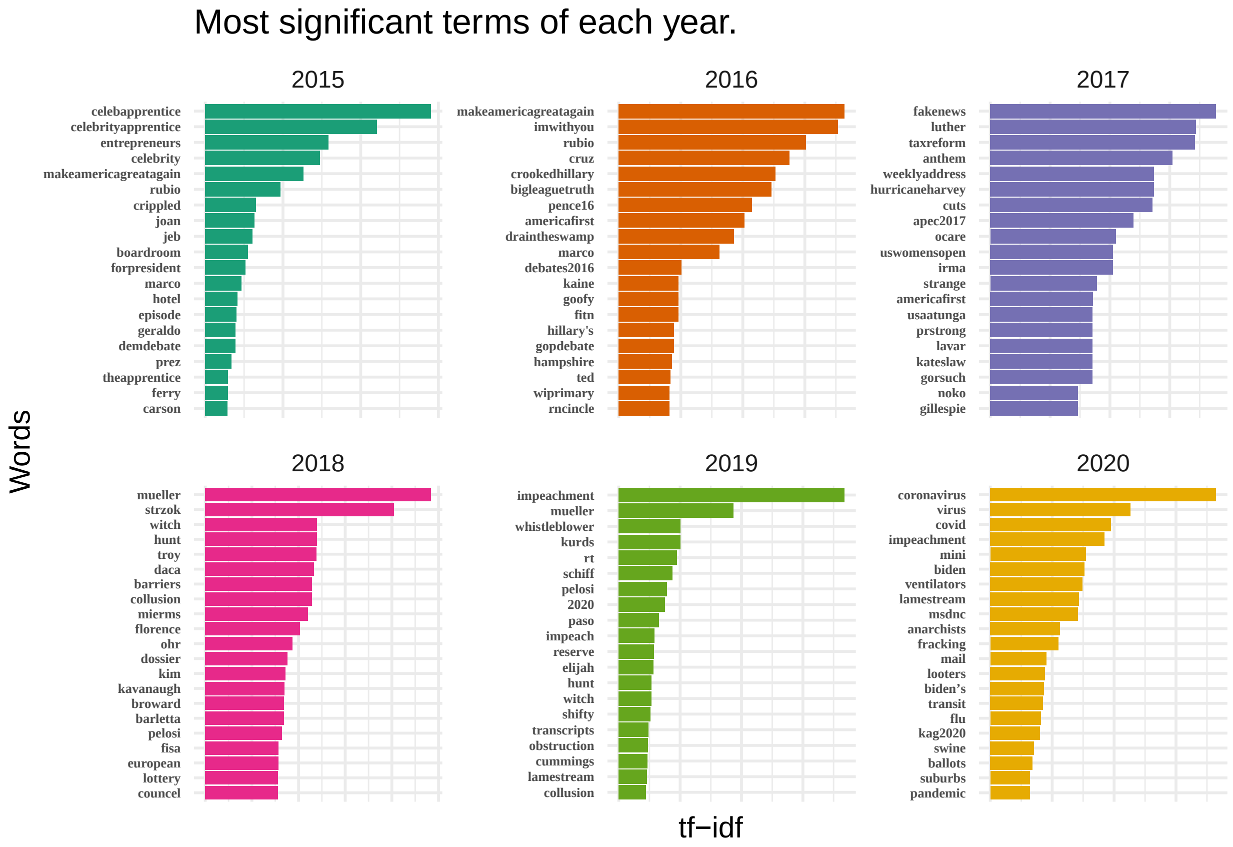
Hillary had become the most widely used word in 2016 since her presidential campaign. Similarly, Joe Biden was the second most popular word in 2020.

The title “President” has been bandied around a lot in 2019 because 2020 was the year of presidential elections.

Because “People” was the most commonly used phrase in the years between the elections and during Trump’s administration, it makes sense that it was the most commonly used term in the two years.

```
trump_tf_idf <- T_data1 %>%
                count(year,
                      word,
                      sort=TRUE)  %>%
                bind_tf_idf(term=word,
                            document=year,
                            n=n)

trump_tf_idf %>%
group_by(year) %>%
top_n(20,wt=tf_idf)  %>%
ggplot(aes(x=reorder_within(word,
                            tf_idf,
                            year),
           y=tf_idf,
           fill=factor(year))) +
scale_fill_brewer(palette="Dark2") +
geom_col(position="dodge",
         show.legend=FALSE)  +
coord_flip() +
facet_wrap(~year, scales="free") +
labs(x="Words",
     y="tf-idf",
     title="Most significant terms of each year.",
     fill="Year") +
scale_x_reordered() +
scale_y_continuous(labels=NULL)  +
theme_minimal() +
theme(axis.text.x=element_text(angle=0,
                               hjust=1,
                               vjust=0.5,
                               size=5))+
theme(axis.text.y=element_text(angle=0,
                               hjust=1,
                               vjust=0.5,
                               size=5,
                               family="Times",
                               face="bold"))
```



As a result of the 2016 presidential elections, the terms celebapprentice and makeamericagreatagain were used in 2015.

In 2017, the phrase “fakenews” was used the most.

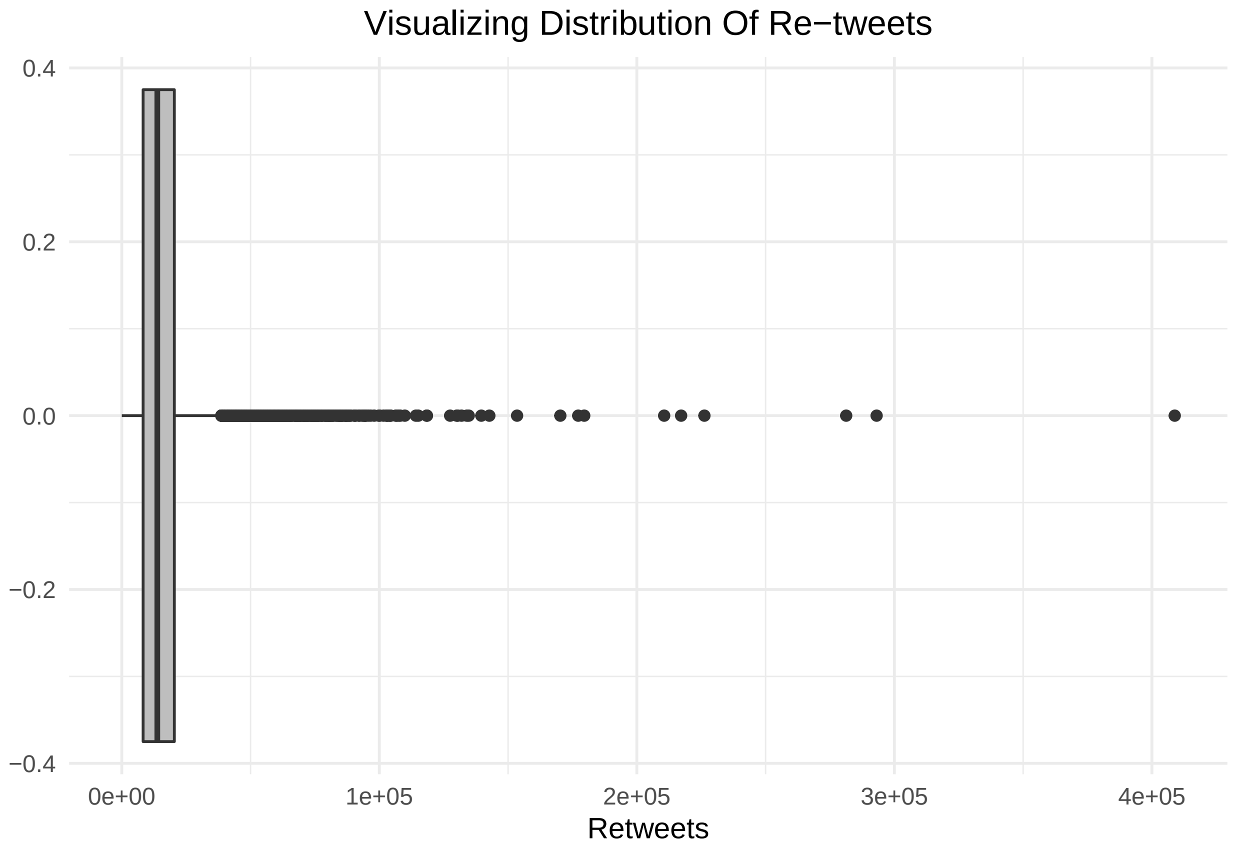
We use mueller as the first word of the year because Robert Muller’s Russian investigation started in 2018. Since the research was finished in December of 2018, mueller is the second word of the year.

Trump was the subject of an impeachment inquiry in 2019. The first word we hear in the year 2019 is impeachment.

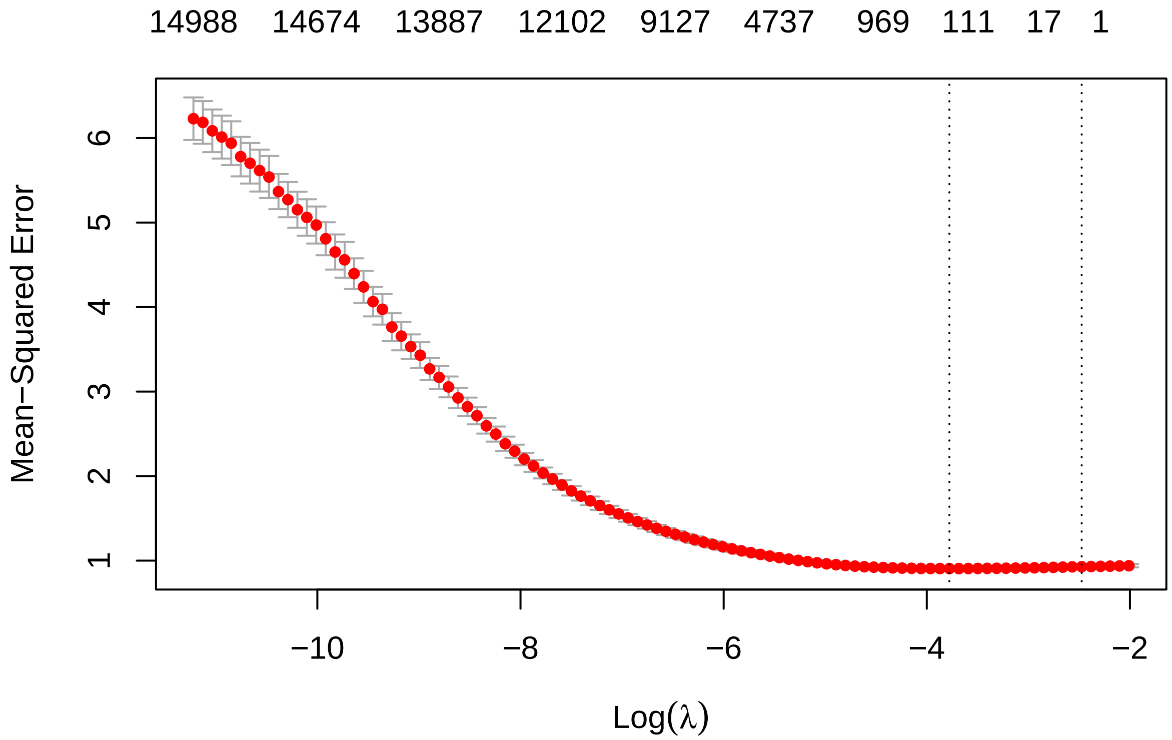
In the year 2020, the COVID-19 Coronavirus is often referenced. We find a range of terms referring to the pandemic and health in the year 2020.

```
T_data1 <- T_data1 %>%
           filter(year >= 2016)
df  <-left_join(T_data1,
                trump_tf_idf,
                by=c("year",
                     "word"))  %>%
select(c("id",
         "retweets",
         "word",
         "n"))
Y <- df %>%
group_by(id)%>%
summarise(retweets=mean(retweets))

Y %>%
ggplot() +
geom_boxplot(aes(x=(retweets)),
             fill  =  "gray")+
labs(x = "Retweets",
     title="Visualizing Distribution Of Re-tweets")+
theme_minimal()+
theme(plot.title =element_text(hjust = 0.5))
```



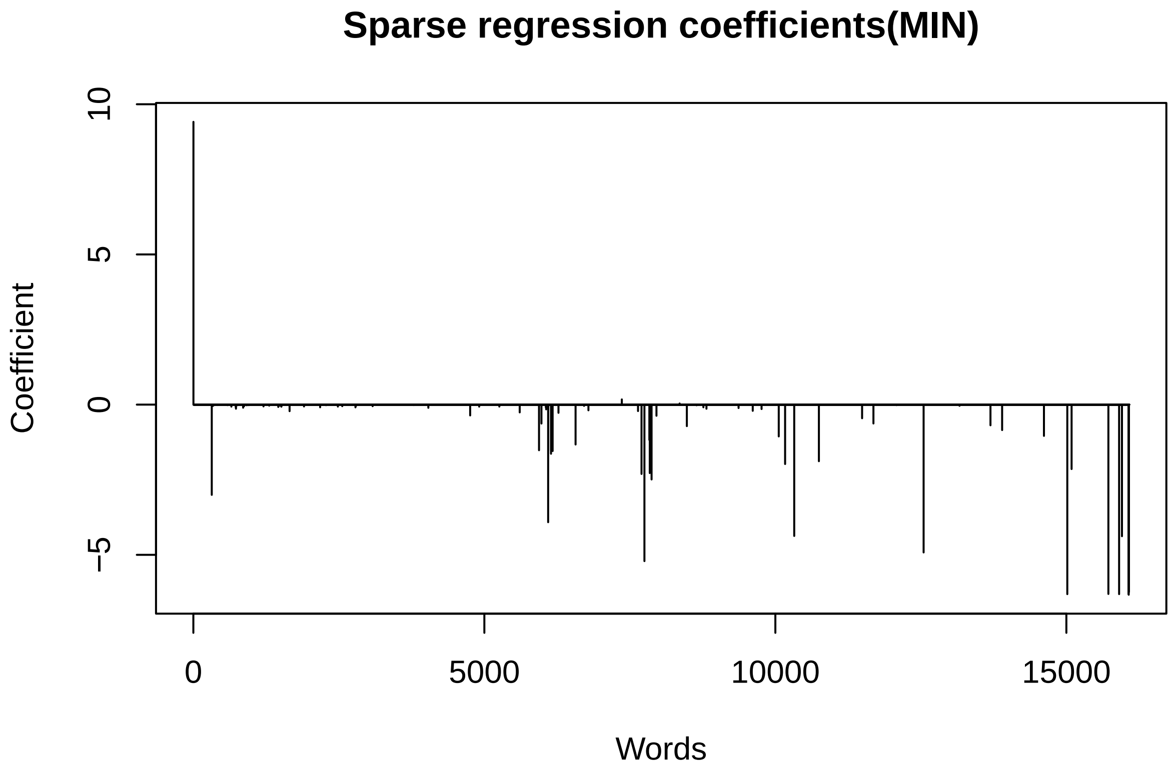
```
X <- cast_sparse(data  =  df,
                 row  =  id,
                 column  =  word,
                 value  =  n)
Y <- as.matrix(log1p(Y$retweets))
set.seed(1234)
cvfit <- cv.glmnet(X,
                   Y,
                   family="gaussian")
plot(cvfit)
```



```
c1<- coef(cvfit,
          s="lambda.min")
sum(c1 != 0)
```

```
## [1] 189
```

```
plot(c1,
     type='h',
xlab="Words",
ylab="Coefficient",
main="Sparse regression coefficients(MIN)")
```



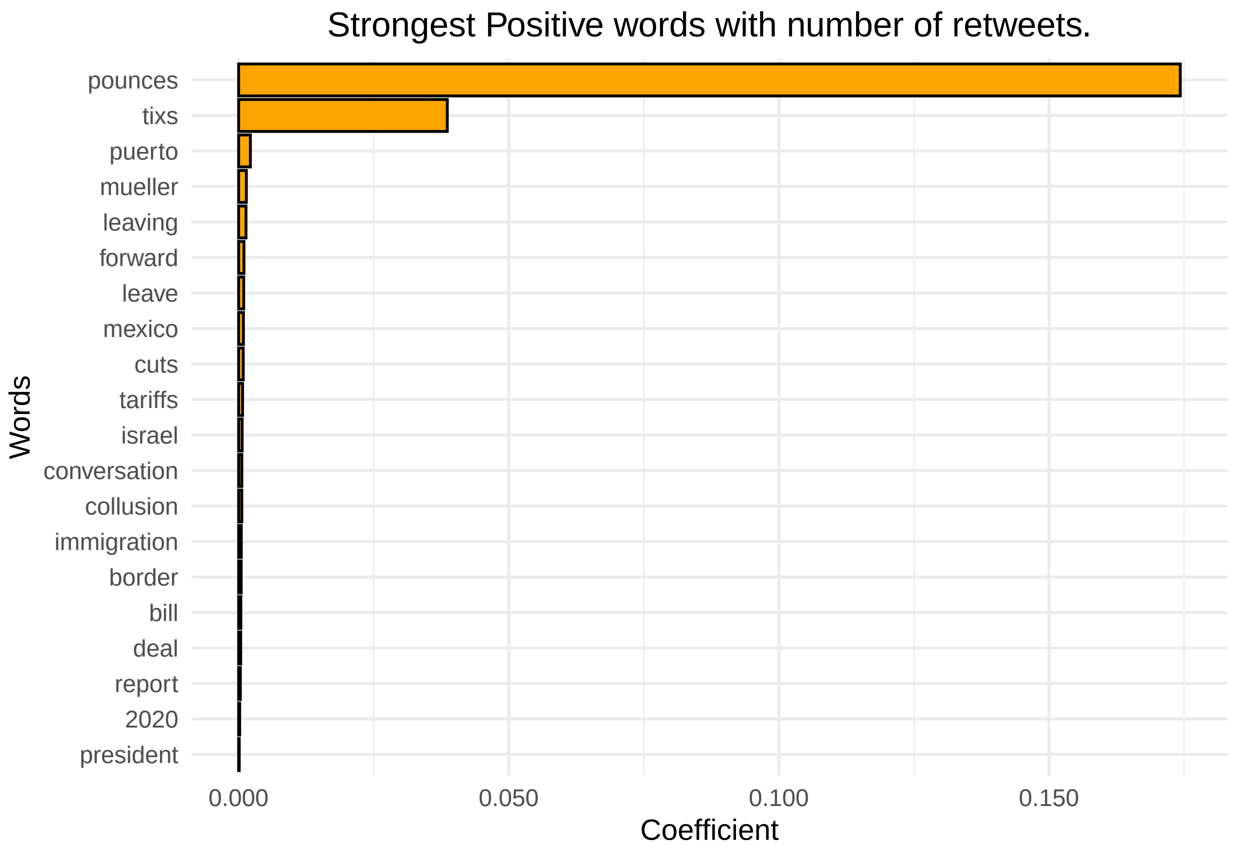
```
cvfit
```

```
##
## Call:  cv.glmnet(x = X, y = Y, family = "gaussian")
##
## Measure: Mean-Squared Error
##
##      Lambda Index Measure     SE Nonzero
## min 0.02288    20  0.9057 0.0237     188
## 1se 0.08416     6  0.9282 0.0200       2
```

Lambda's minimum value is 0.02288, and there are 188 non-zero coefficients.

```
coeff_vars <-rownames(c1)[which(c1 >0)]
coeff <- c1[which(c1>0)]
model <- as.data.frame(coeff_vars)
model$coeff <- coeff
model <- model %>%
filter(coeff_vars != "(Intercept)")
model  <-  model[order(model$coeff,
                       decreasing  =  TRUE),]
model %>%
top_n(20) %>%
ggplot(aes(y = reorder(coeff_vars,
                       coeff),
           x = coeff))+
geom_col(color="black",
         fill="orange")+
scale_x_continuous(labels=scales::label_comma())+
labs(x = "Coefficient",
    y  =  "Words",
    title = "Strongest Positive words with number of retweets.")+
theme_minimal()+
theme(plot.title = element_text(hjust = 0.5))
```

```
## Selecting by coeff
```



We can see that pounces is the strongest positive word with the most retweets when we look at the top 20 strongest positive terms by number of tweets.

Out of these 20, 5 words have higher coefficients which include pounces, tixs, puerto, mueller and leaving.

All the other words have a small value of coefficients and less positive relationship with the number of retweets.