Tweets

Igor Adamiec

7/31/2019

Background

This analysis base on **Democrat Vs. Republican Tweets** dataset. It contains 200 lastest tweets for hundreds US politicians (tweets were gathered in May 2018).

Analysis

Libraries

```
library(tidyverse)
library(tidytext)
library(tm)
library(qdap)
library(rebus)
```

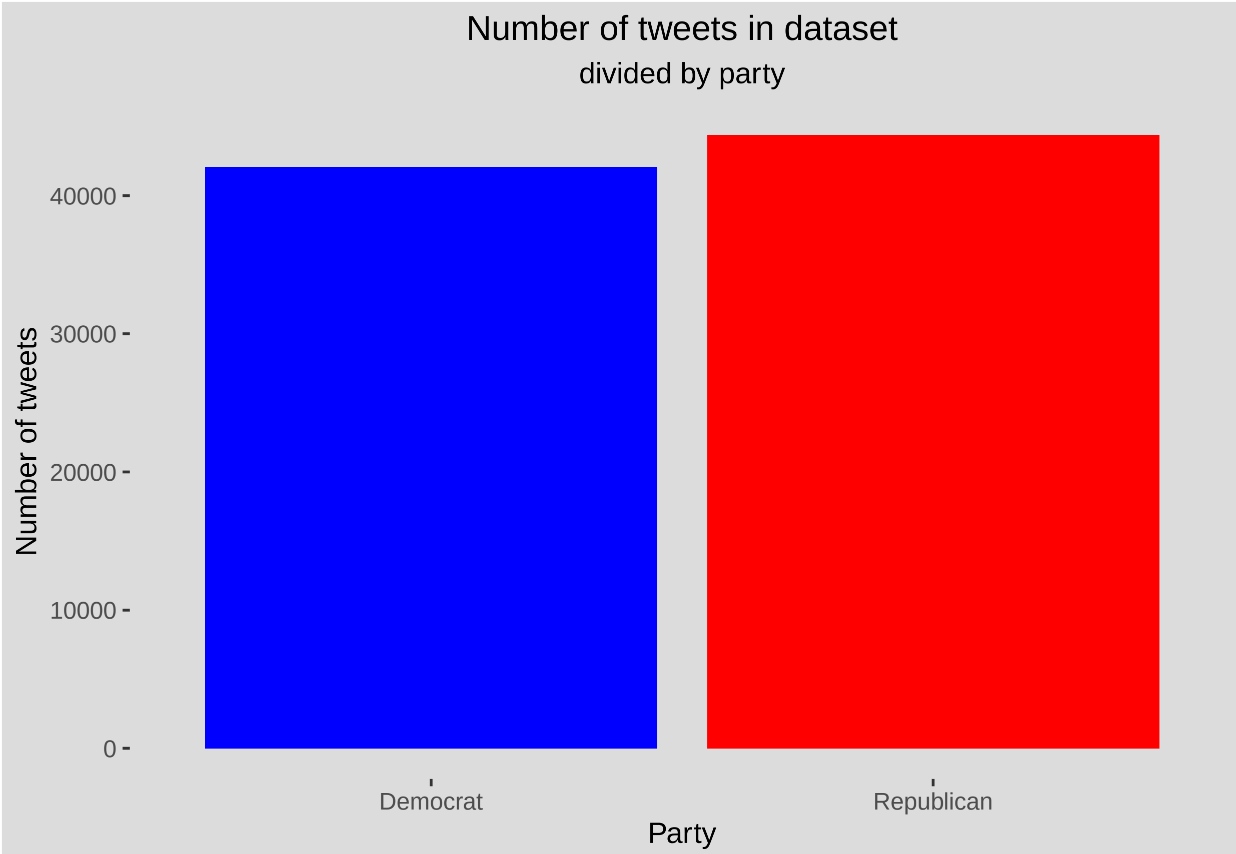
Loading file and exploratory analysis

```
tweets <- read_csv("./ExtractedTweets.csv")
```

```
tweets %>% glimpse()
```

```
## Observations: 86,460
## Variables: 3
## $ Party  <chr> "Democrat", "Democrat", "Democrat", "Democrat", "Democr...
## $ Handle <chr> "RepDarrenSoto", "RepDarrenSoto", "RepDarrenSoto", "Rep...
## $ Tweet  <chr> "Today, Senate Dems vote to #SaveTheInternet. Proud to ...
```

We can see that dataset contains three columns. First is party of user (either Dem or Rep), second is their account name and the last one is the tweet itself. Dataset contains almost 86,5 thousand tweets. Let's take a closer look which party twitts more.



Above plot shows that Republicans tweet a bit more than democrats. We can see exact numbers below:

```
## # A tibble: 2 x 3
##   Party          n percentage
##   <chr>      <int>      <dbl>
## 1 Democrat   42068      0.487
## 2 Republican 44392      0.513
```

Total difference is around 2300 tweets that is about 2% of all of them.

I also wanted to see which user tweeted the most but this dataset contains max 200 latest tweets so let's see how many politicians hitted maximum.

```
## # A tibble: 4 x 2
##   `Number of tweets per author`     n
##                           <int> <int>
## 1                           200   416
## 2                           199    14
## 3                           197     2
## 4                            80     1
```

Total number of politicians in this dataset is 433 and representation of Republicans is bit higher.

```
## # A tibble: 2 x 2
##   Party      politicians
##   <chr>            <int>
## 1 Democrat           211
## 2 Republican         222
```

Word counting

First I added new column in which I tidied all tweets. Below code may not look great but it is much faster than writing my own function and using map from **purrr** package. As a stop words I used stop_words set from **tidytext** package with some additions.

```
custom_stop_words <- tribble(
  ~word, ~lexicon,
  "https", "CUSTOM",
  "t.co", "CUSTOM",
  "rt", "CUSTOM",
  "amp", "CUSTOM"
)

stop_words2 <- stop_words %>%
  bind_rows(custom_stop_words)
```

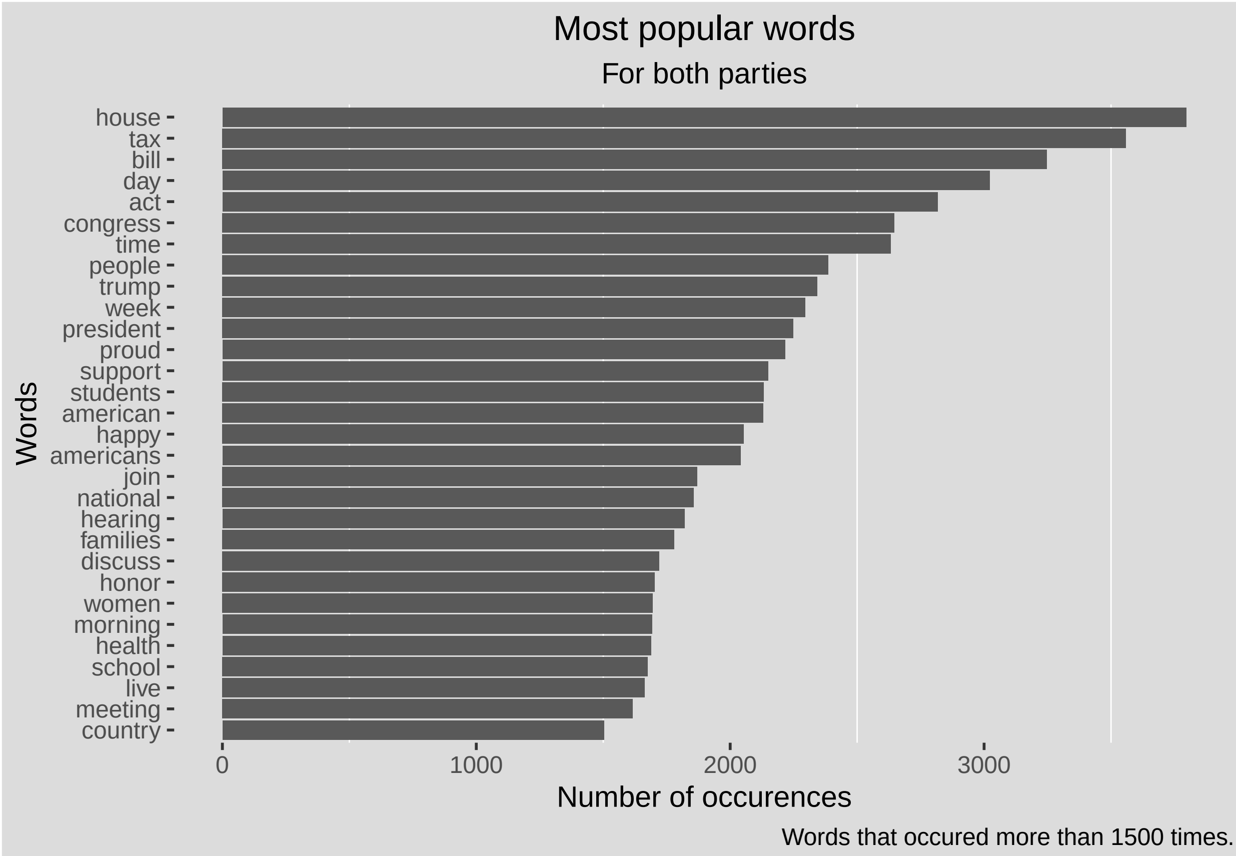
```
http_pattern <- "https" %R% one_or_more(or(ALNUM, PUNCT))

prepared_tweets <- tweets %>%
  mutate(prepared_tweet = str_remove_all(Tweet, http_pattern),
         prepared_tweet = replace_contraction(prepared_tweet),
         prepared_tweet = replace_contraction(prepared_tweet),
         prepared_tweet = str_to_lower(prepared_tweet),
         prepared_tweet = removePunctuation(prepared_tweet),
         prepared_tweet = str_remove_all(prepared_tweet, pattern = "...|’"),
         prepared_tweet = removeWords(prepared_tweet, stop_words2$word),
         prepared_tweet = stripWhitespace(prepared_tweet),
         prepared_tweet = str_trim(prepared_tweet))
```

Division by words

Below we can see top 10 words used in this dataset. We can find there words describing politics related places (house, congress), law related things (tax, bill, act), classic expressions used by politicians (people, day, time, week) and what is most interesting **Trump**.

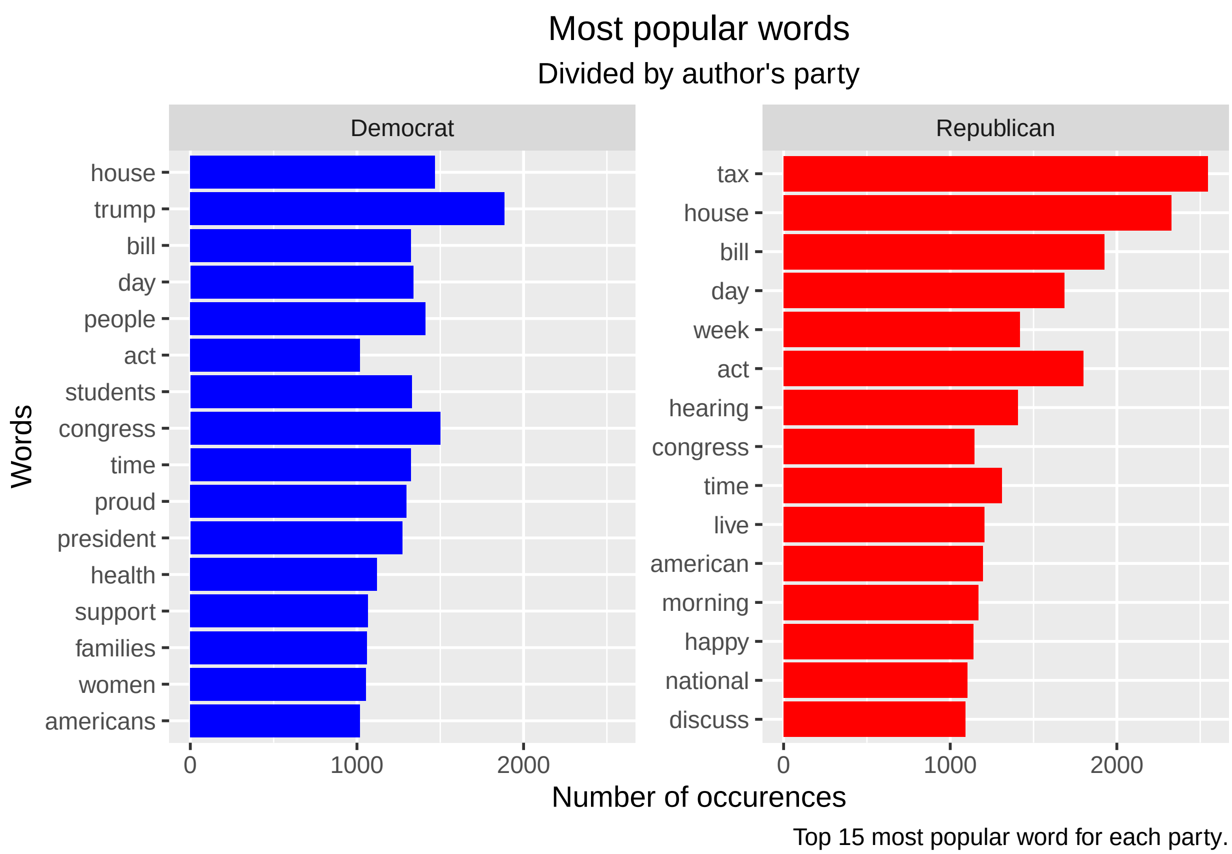
```
## # A tibble: 10 x 2
##    word         n
##    <chr>    <int>
##  1 house     3796
##  2 tax       3558
##  3 bill      3248
##  4 day       3022
##  5 act       2817
##  6 congress  2646
##  7 time      2633
##  8 people    2386
##  9 trump     2343
## 10 week      2296
```

Below plot expands above table and shows all most popular words that occured in the dataset more than 1500 times.



Below I divided dataset by party. We can see that Republicans more often repeat words than Democrats (none of "Democratic" words has more than 2000 occurences).

Republicans in their tweets mostly focus on financial stuff (taxes, bills) ans Democrats focus on Donald Trump. On lower position in Democrats' tweets we can see progressive values such as health, families, support and women. Republicans focuses on words like american and national.

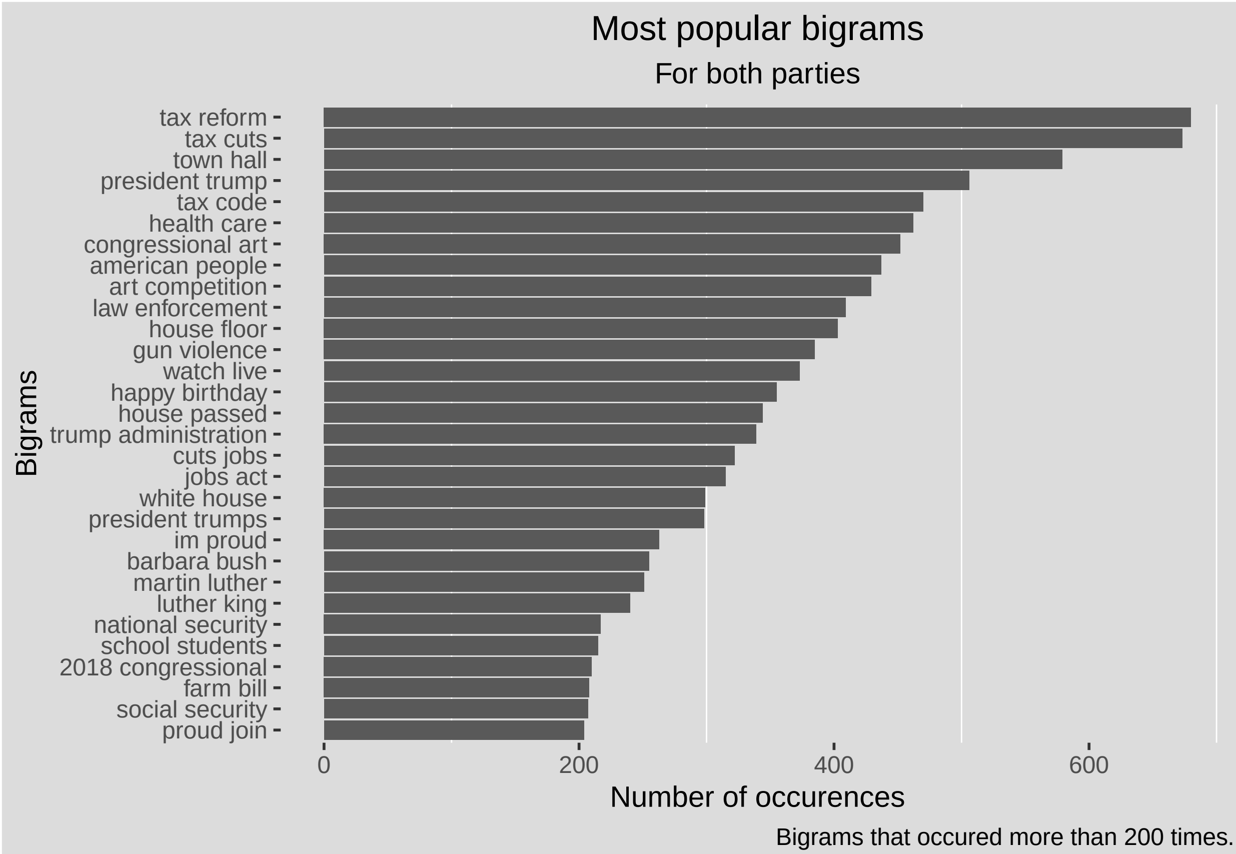


Division by bigrams

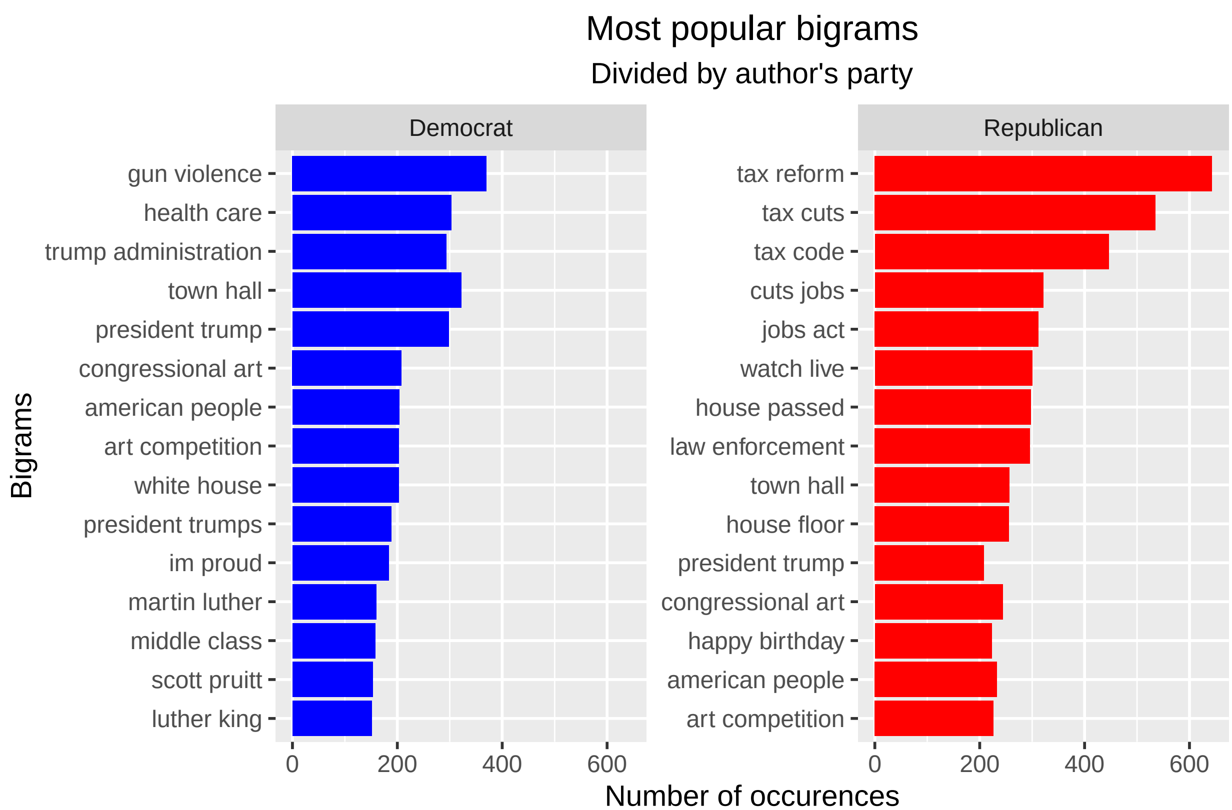
Below I counted every bigram (combination of two words) from tweets. Now the image is much clearer than earlier. We can see that most popular topics are tax reform, tax cuts, town hall and ... President Trump.

```
## # A tibble: 10 x 2
##    token                 n
##    <chr>             <int>
##  1 tax reform          680
##  2 tax cuts            673
##  3 town hall           579
##  4 president trump     506
##  5 tax code            470
##  6 health care         462
##  7 congressional art   452
##  8 american people     437
##  9 art competition     429
## 10 law enforcement     409
```

Below plots shows most popular bigrams that occured more than 200 times in dataset. We can see that phrase "happy birthday" is quite popular. Also we can find well known figures such as Barbara Bush and Martin Luther King. Let's see who mentions them more.



As we would suspect, Democrats tweet about bad things (gun violence, trump administration) while Republicans mention tax reforms. Both parties mention President Trump (Democrats a bit more) but we can assume that they have different reasons. Democrats mention Martin Luther King but also Scott Pruitt - former chief of EPA that resigned in July 2018 after some scandals.

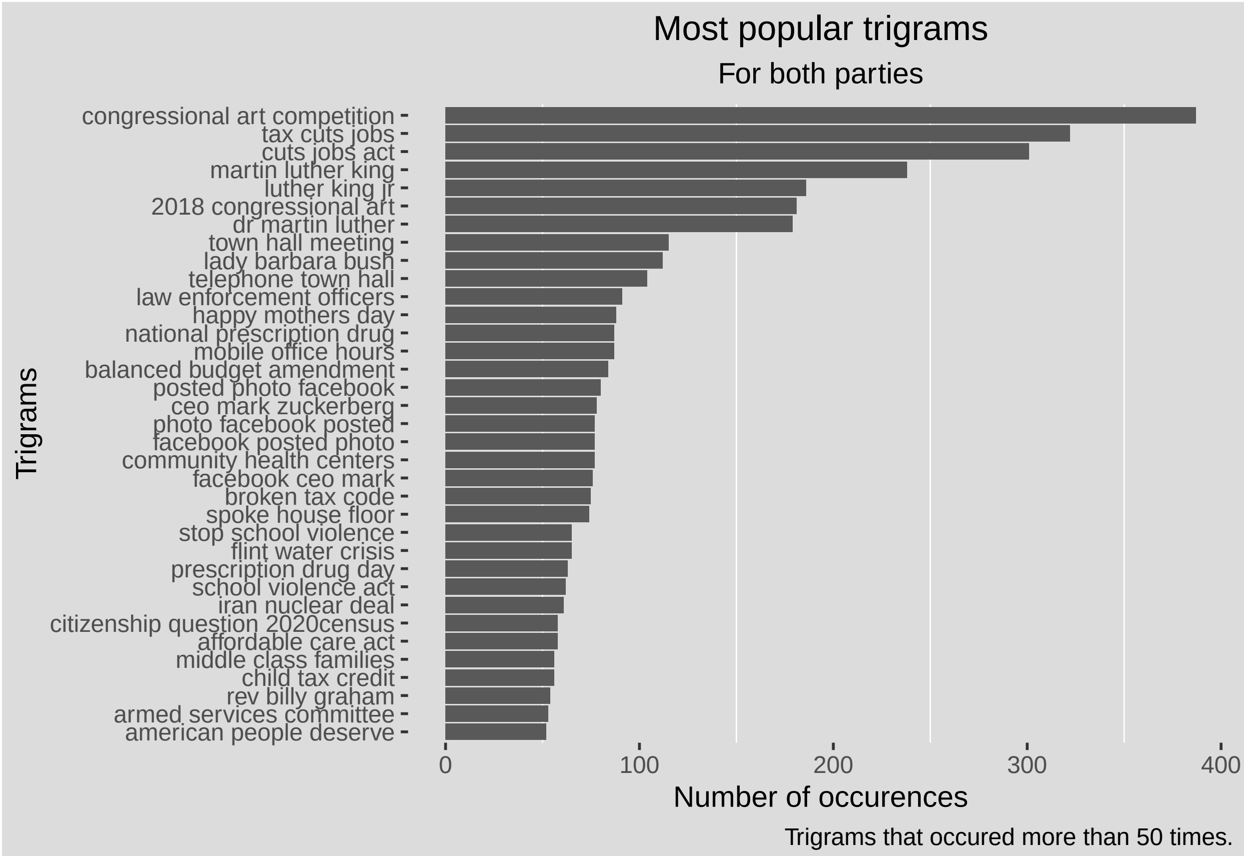


Top 15 most popular bigrams for each party.

Division by trigrams

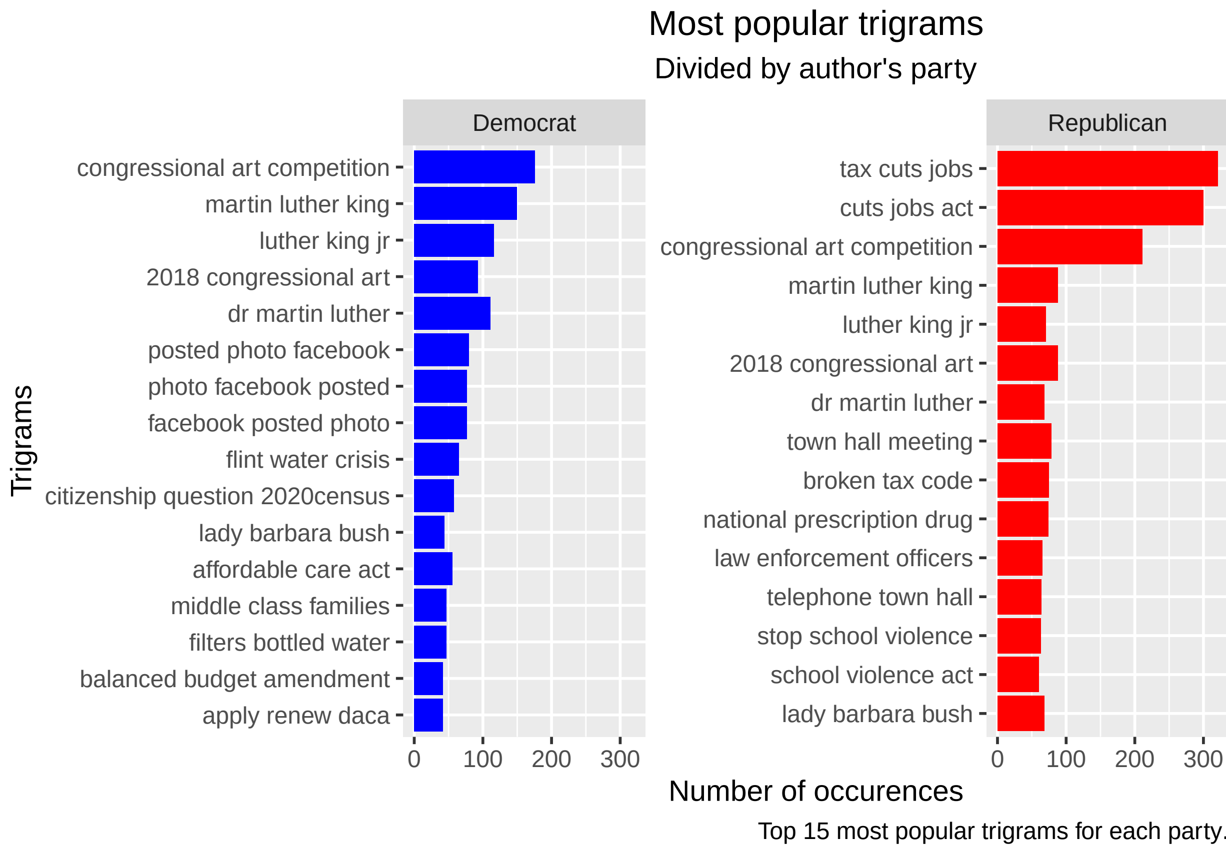
Below I extracted all trigrams - each combination of 3 words.

Below we can see most popular trigrams. They differ than single words and bigrams. Except phrases that we knew from previous plots now we can see mentions about Mark Zuckerberg.



Same trigrams, of course, occure less than bigrams so this analysis can be less accurate but we can see that Republicans also mentioned Martin Luter King quite often.

```
unnested_trigrams %>%
  count(Party, token) %>%
  group_by(Party) %>%
  top_n(15, n) %>%
  ggplot(aes(x = reorder(token, n), y = n)) +
  geom_col(aes(fill = Party), show.legend = F) +
  facet_wrap(.~Party, scales = "free_y")+
  coord_flip() +
  scale_fill_manual(values = c("blue", "red")) +
  theme(plot.title = element_text(hjust = .5),
        plot.subtitle = element_text(hjust = .5)) +
  labs(title = "Most popular trigrams", subtitle = "Divided by author's party",
       x = "Trigrams", y = "Number of occurences",
       caption = "Top 15 most popular trigrams for each party.")
```

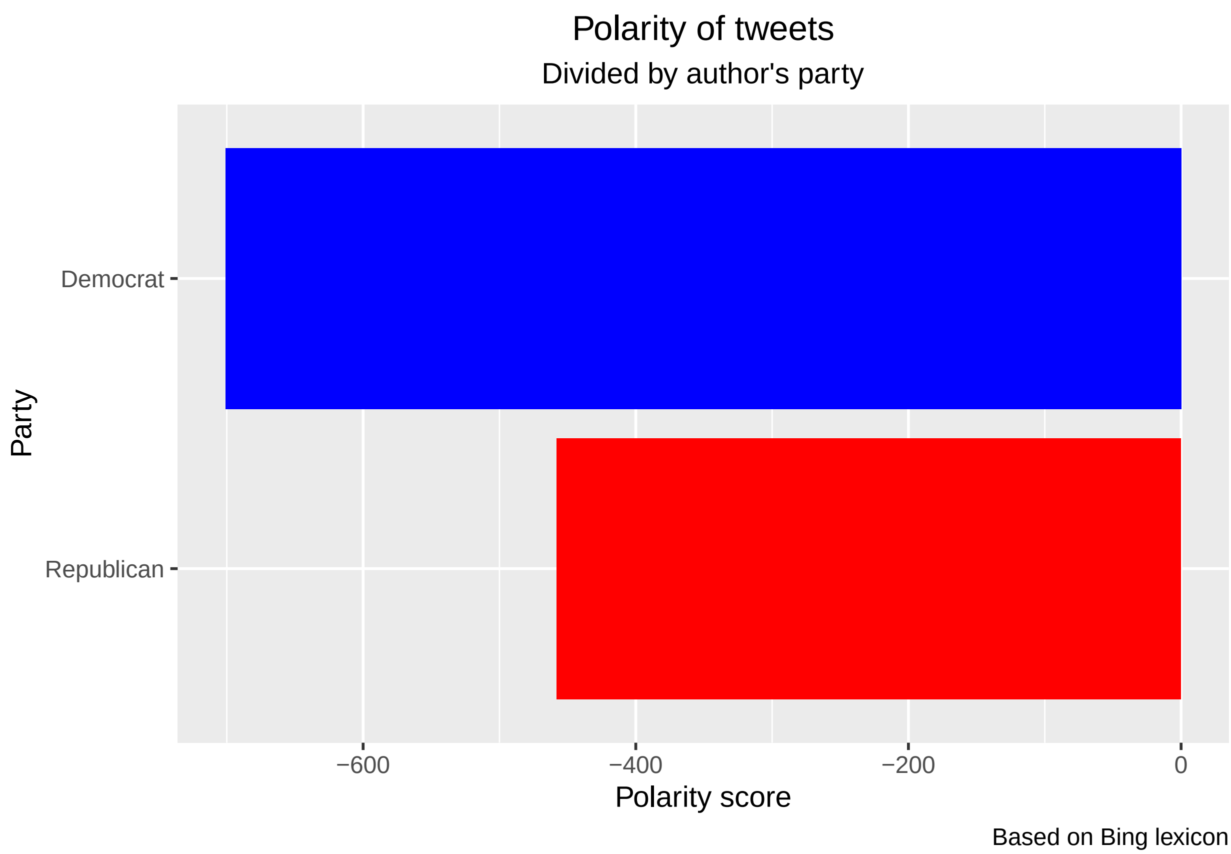


Sentiment analysis

By Party

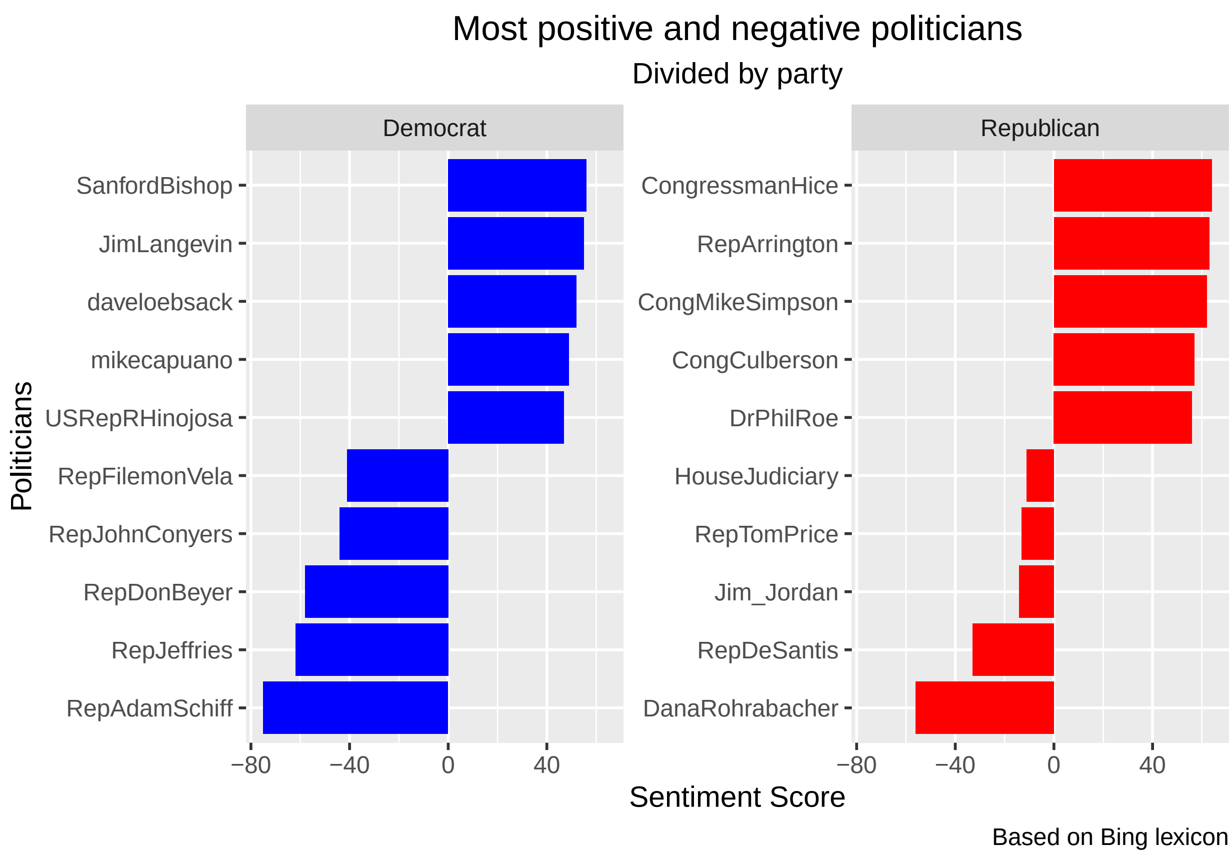
To calculate polarity I used Bing lexicon. It rates word as either positive or negative. Polarity is the difference between positive and negative words. Sad conclusion is fact that tweets of both parties are strongly negative. I would really prefer politician that rather join than divide. But as we know complaining and attacking opponents sells more.

```
## # A tibble: 2 x 4
##   Party      negative positive polarity
##   <chr>         <int>    <int>    <int>
## 1 Democrat       1582      881     -701
## 2 Republican     1324      866     -458
```



By users

I chose 5 most negative and most positive politicians of both parties. Below are the results



Positive Democrats

Looking on 10 most popular bigrams from top 2 positive Democrats, we can see phrase "I'm proud".

```
## # A tibble: 10 x 2
##    token                          n
##    <chr>                      <int>
##  1 rhode island                  19
##  2 im proud                      10
##  3 rhode islands                  8
##  4 albany georgia                 6
##  5 military family                6
##  6 senjackreed senwhitehouse      6
##  7 health insurance               5
##  8 town hall                      5
##  9 townhall meeting               5
## 10 2nd congressional              4
```

Looking at words we can see "congratulations", "pleasure" and mentioned previously - "proud".

```
## # A tibble: 10 x 2
##    token               n
##    <chr>           <int>
##  1 georgia            30
##  2 rhode              30
##  3 im                 24
```

```
##  4 congratulations    21
##  5 island             19
##  6 proud              18
##  7 students           18
##  8 meeting            17
##  9 time               17
## 10 pleasure           16
```

Positive Reps

2 most positive Republicans also were proud. Moreover they were tweeting about some brave women and wished happy.

```
## # A tibble: 10 x 2
##    token                        n
##    <chr>                    <int>
##  1 service academy              8
##  2 covnews congressmanhice      7
##  3 dee dee                      7
##  4 art competition              5
##  5 brave women                  5
##  6 congressional art            5
##  7 im proud                     5
##  8 wishing happy                5
##  9 women uniform                5
## 10 barbara bush                 4
```

```
## # A tibble: 10 x 2
##    token                n
##    <chr>            <int>
##  1 idaho               41
##  2 congressmanhice     29
##  3 congratulations     21
##  4 ga10                21
##  5 service             21
##  6 meeting             18
##  7 week                18
##  8 bill                17
##  9 congmikesimpson     17
## 10 day                 15
```

Negative Democrats

Two most negative Democrats were complaining about white house, living situation of american people, gun violence and… female rap? Also popular phrase is "so called president".

```
## # A tibble: 10 x 2
##    token                       n
##    <chr>                   <int>
##  1 american people            17
##  2 white house                11
##  3 female rap                 10
```

```
##   4 gun violence               8
##   5 intelligence committee     8
##   6 rap collaboration          8
##   7 socalled president         8
##   8 stock market               7
##   9 infrastructure plan        6
##  10 president trump            6
```

```
## # A tibble: 10 x 2
##    token           n
##    <chr>       <int>
##  1 president      68
##  2 trump          50
##  3 house          42
##  4 people         31
##  5 american       26
##  6 time           21
##  7 republicans    20
##  8 america        19
##  9 socalled       19
## 10 republican     18
```

Negative Reps

```
prepared_tweets %>%
  filter(Handle %in% c("DanaRohrabacher", "RepDeSantis")) %>%
  select(-Tweet) %>%
  unnest_tokens(token, prepared_tweet, token = "ngrams", n = 2) %>%
  count(token) %>%
  arrange(desc(n)) %>%
  head(10)
```

```
## # A tibble: 10 x 2
##    token                               n
##    <chr>                           <int>
##  1 hush fund                          12
##  2 accountability hush                 8
##  3 congressional accountability        8
##  4 president trump                     7
##  5 elimination act                     6
##  6 forward joining                     6
##  7 fund elimination                    6
##  8 lwherron reprohrabacher             6
##  9 american people                     5
## 10 andrew mccabe                       4
```

```
## # A tibble: 10 x 2
##    token           n
##    <chr>       <int>
##  1 repdesantis    30
##  2 congress       25
##  3 people         22
##  4 election       20
```

```
##  5 house            20
##  6 president        19
##  7 trump            19
##  8 2                18
##  9 congressional    18
## 10 constituents     18
```