Word cloud

Taehee Kim

12/6/2020

```
library(dplyr)
library(wordcloud)
library(RColorBrewer)
library(tidytext)
library(SnowballC)

load("rigged_election.RData")
```

Preperation

In this session, we create a word cloud of two goups we detected last session. To do that, we need to extract tweet texts of each group members. Since some of user's texts are stroed in retweet related columns, I create a new dataframe which contain all original tweets as a first step.

```
fields <- c("user_id", "status_id", "created_at", "screen_name", "text", "retweet_
count",
            "is_retweet", "is_quote", "reply_to_user_id")

fields_rt <- c("retweet_user_id", "retweet_status_id", "retweet_created_at",
               "retweet_screen_name", "retweet_text", "retweet_retweet_count")


tweets <- rigged %>%
  select(all_of(fields)) %>% # When you write select(fields), you get a warning me
ssage.
  filter(is_retweet == FALSE) %>% # Remove retweets
  filter(is_quote == FALSE) %>% # Remove quotes
  filter(is.na(reply_to_user_id) == TRUE) %>% # Remove replies
  select(-c(is_retweet, is_quote, reply_to_user_id)) # Remove columns we don't nee
d anymore

head(tweets)
```

```
## # A tibble: 6 x 6
##   user_id   status_id  created_at          screen_name text        retweet_coun
t
##   <chr>     <chr>      <dttm>              <chr>       <chr>               <int
>
## 1 78814863… 133492477… 2020-12-04 18:17:23 Donlusin    "\"RIGGED …
7
## 2 30517896… 133492948… 2020-12-04 18:36:05 robertc3d   "Georgia R…
0
## 3 39441344  133492948… 2020-12-04 18:36:04 smalls777   "Yes, yes,…
0
## 4 97367856… 133492946… 2020-12-04 18:35:59 BMccaughrin "Nothing w…
0
## 5 14748990… 133492942… 2020-12-04 18:35:50 WriterLDud… "Trump's c…
0
## 6 11911578… 133492942… 2020-12-04 18:35:50 rlouis82    "Since Don…
0
```

```
tweets_rt <- rigged %>%
  select(all_of(fields_rt)) %>%
  filter(is.na(retweet_user_id) == FALSE) # remove if retweet_user_id is NA

head(tweets_rt)
```

```
## # A tibble: 6 x 6
##   retweet_user_id retweet_status_… retweet_created_at  retweet_screen_…
##   <chr>           <chr>            <dttm>              <chr>
## 1 32871086        133465490301729… 2020-12-04 00:25:00 kylegriffin1
## 2 25073877        133361515190439… 2020-12-01 03:33:24 realDonaldTrump
## 3 25073877        133397599151818… 2020-12-02 03:27:15 realDonaldTrump
## 4 25073877        133485885233707… 2020-12-04 13:55:25 realDonaldTrump
## 5 74303349        133486838395589… 2020-12-04 14:33:18 MarkFinchem
## 6 25073877        133485885233707… 2020-12-04 13:55:25 realDonaldTrump
## # … with 2 more variables: retweet_text <chr>, retweet_retweet_count <int>
```

```
names(tweets_rt) <- names(tweets) # change names of columns of tweets_rt to bind t
wo data.frame.

all_tweets <- bind_rows(tweets, tweets_rt)
dim(all_tweets)
```

```
## [1] 46741     6
```

```
all_tweets <- distinct(all_tweets, status_id, .keep_all = TRUE) # Remove duplicate
d ones.
dim(all_tweets)
```

```
## [1] 3705    6
```

```
# Check tweet text
trump <- all_tweets %>%
  filter(screen_name == "realDonaldTrump")
trump$text
```

```
##  [1] "RIGGED ELECTION!"
##  [2] "Not statistically possible. Rigged Election! https://t.co/Yw8roUTJhy"
##  [3] "Rigged Election. Show signatures and envelopes. Expose the massive voter
fraud in Georgia. What is Secretary of State and @BrianKempGA afraid of. They know
what we'll find!!! https://t.co/Km7tRm2s1A"
##  [4] "I gave a long news conference today after wishing the military a Happy Th
anksgiving, &amp;amp; realized once again that the Fake News Media coordinates so that
the real message of such a conference never gets out. Primary point made was that
the 2020 Election was RIGGED, and that I WON!"
##  [5] "We have some big things happening in our various litigations on the Elect
ion Hoax. Everybody knows it was Rigged. They know Biden didn't get more votes fro
m the Black community than Obama, &amp;amp; certainly didn't get 80,000,000 votes. Loo
k what happened in Detroit, Philadelphia, plus!"
##  [6] "...And there are many such articles. Rigged Election! https://t.co/P8scMa
uhcI"
##  [7] "RIGGED 2020 ELECTION: MILLIONS OF MAIL-IN BALLOTS WILL BE PRINTED BY FORE
IGN COUNTRIES, AND OTHERS. IT WILL BE THE SCANDAL OF OUR TIMES!"
##  [8] "RIGGED ELECTION!"
##  [9] "“Democrats suffered crushing down-ballot loss across America.” @nytimes.
This is true. All statehouses won, and in Washington we did great. So I led this g
reat charge, and I'm the only one that lost? No, it doesn't work that way. This wa
s a massive fraud, a RIGGED ELECTION!"
## [10] "RIGGED ELECTION. WE WILL WIN!"
## [11] "He only won in the eyes of the FAKE NEWS MEDIA. I concede NOTHING! We hav
e a long way to go. This was a RIGGED ELECTION!"
## [12] "Just saw the vote tabulations. There is NO WAY Biden got 80,000,000 votes
!!! This was a 100% RIGGED ELECTION."
## [13] "People will not accept this Rigged Election! https://t.co/XQAOIt5ZwU"
## [14] "Poll: 79 Percent of Trump Voters Believe 'Election Was Stolen' https://t.
co/PmMBmt05AI via @BreitbartNews They are 100% correct, but we are fighting hard.
Our big lawsuit, which spells out in great detail all of the ballot fraud and more
, will soon be filled. RIGGED ELECTION!"
## [15] "This Election was RIGGED, but we will WIN! https://t.co/luS6SnFscx"
## [16] "Many mostly Democrat States refused to hand over data from the 2016 Elect
ion to the Commission On Voter Fraud. They fought hard that the Commission not see
their records or methods because they know that many people are voting illegally.
System is rigged, must go to Voter I.D."
## [17] "Heartwarming to see all of the tremendous support out there, especially t
he organic Rallies that are springing up all over the Country, including a big one
on Saturday in D.C. I may even try to stop by and say hello. This Election was Rig
ged, from Dominion all the way up &amp;amp; down!"
```

```
## [18] "A Rigged Election! https://t.co/dAviFrkEP4"
## [19] "The “Republican” Governor of Georgia, @BrianKempGA, and the Secretary of
State, MUST immediately allow a signature verification match on the Presidential E
lection. If that happens, we quickly and easily win the State and importantly, pav
e the way for a big David and Kelly WIN!"
## [20] "Breaking News: 50,000 OHIO VOTERS getting WRONG ABSENTEE BALLOTS. Out of
control. A Rigged Election!!!"
## [21] "Dominion is running our Election. Rigged! https://t.co/xvwrpLpAZa"
## [22] "....Absentee Ballots are fine because you have to go through a precise pr
ocess to get your voting privilege. Not so with Mail-Ins. Rigged Election!!! 20% f
raudulent ballots?"
## [23] "He won because the Election was Rigged. NO VOTE WATCHERS OR OBSERVERS all
owed, vote tabulated by a Radical Left privately owned company, Dominion, with a b
ad reputation &amp;amp; bum equipment that couldn’t even qualify for Texas (which I wo
n by a lot!), the Fake &amp;amp; Silent Media, &amp;amp; more! https://t.co/Exb3C1mAPg"
## [24] "Hundreds of thousands of people showing their support in D.C. They will n
ot stand for a Rigged and Corrupt Election! https://t.co/tr35WKTKM8"
## [25] "Top US Pollster and Statistician Richard Baris — People's Pundit — SUSPEN
DED from Twitter for Reporting on Disputed Election — Political 'WrongThink' Not A
llowed https://t.co/uHBZnJJn1I via Said 10,000 DEAD PEOPLE VOTED IN MICHIGAN. When
will this RIGGED ELECTION be overturned!"
```

```
sanders <- all_tweets %>%
  filter(screen_name == "BernieSanders")
sanders$text
```

```
## [1] "Trump's rants about a \"fraudulent election\" are not a joke. They are the
most significant attack against our democracy in history. If the election system's
\"rigged,\" if the media's \"fake,\" if federal officials are part of a \"deep sta
te,\" who can you trust? You got it. A dictator."
```

```
# Find red group's tweets
red_tweets <- all_tweets %>%
  filter(user_id %in% red)

blue_tweets <- all_tweets %>%
  filter(user_id %in% blue)

dim(red_tweets)
```

```
## [1] 189   6
```

```
dim(blue_tweets)
```

```
## [1] 352   6
```

Plotting word cloud

Now let's create a word cloud. To do that, we first need to clean up tweet texts. Following example demonstrates cleaning tweets text using regular expression and `gsub()` function. `gsub()` function replace all the matches of a pattern from a string. `gsub(pattern, replacement, string)`.

```
# A function for cleaning tweet text
clean_tweet <- function(tweet_df){
  text <- tweet_df$text
  text <- gsub("https\\S*", "", text)  # remove https
  text <- gsub("@\\S*", "", text) # remove screen_name
  text <- gsub("[\n]", "", text) # remove line breaks
  text <- gsub("[[:punct:]]", "", text) # remove punctuation
  return(text)
}
```

First, we plot word cloud of the blue group.

```
# Clean text
blue_tweets$text <- clean_tweet(blue_tweets)


tweets_words <- blue_tweets %>%
  select(text) %>%
  unnest_tokens(word, text) # Tokenization: break the text into individual tokens

head(tweets_words)
```

```
## # A tibble: 6 x 1
##   word
##   <chr>
## 1 thats
## 2 the
## 3 other
## 4 part
## 5 that
## 6 makes
```

```
tweets_words <- tweets_words %>%
  anti_join(stop_words) # remove stop words: a commonly used word (eg., a, the, th
at) that does not give us much information
```

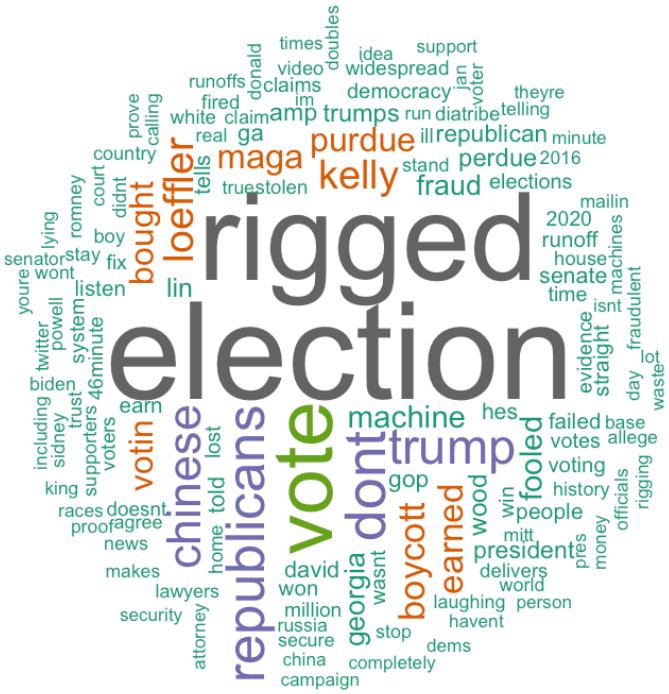
```
## Joining, by = "word"
```

```
words <- tweets_words %>%
  count(word, sort = TRUE)

head(count)
```

```
##
## 1 function (x, ..., wt = NULL, sort = FALSE, name = NULL, .drop = group_by_drop
_default(x))
## 2 {
## 3     if (!missing(...)) {
## 4         out <- group_by(x, ..., .add = TRUE, .drop = .drop)
## 5     }
## 6     else {
```

```
set.seed(3)
wordcloud(words = words$word,
          freq = words$n,
          random.order = FALSE,
          rot.per = 0.35,
          colors = brewer.pal(8, "Dark2"),
          scale = c(4,0.5),
          min.freq = 5)
```



Exercise

1. Plot word cloud of `red` group.

(2-1) Plot word cloud using Donald Trump's recent tweets.

(2-2) Plot word cloud using Joe Biden's recent tweets.

(2-3) Compare the results.

BONUS: Stemming

```
# Stemmed version

tweets_words <- blue_tweets %>%
  select(text) %>%
  unnest_tokens(word, text)

tweets_words <- tweets_words %>%
  anti_join(stop_words)
```

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

```
words <- tweets_words %>%
  mutate(stem = wordStem(word)) %>%
  count(stem, sort = TRUE)

set.seed(5)
wordcloud(words = words$stem, freq = words$n, random.order = FALSE, rot.per = 0.35
,
          colors = brewer.pal(8, "Dark2"), scale = c(4,0.5), min.freq = 5)
```