# Word finding in Tweets

Shimeng Dai

March 22, 2023

## Contents



## Load Package

```
library(tidyverse)
```

## Keyword data

```
# a list of words describing critical race theory
word_set <- word_list$ALL_CRT

head(word_set,20)
```

```
##  [1] "CRT"               "chardonnayantifa"  "VirginiaDemocrats"
##  [4] "activism"          "activist"          "activists"
##  [7] "anti cop"          "anti racist"       "anti white"
## [10] "anti-american"     "anti-american"     "anti-CRT"
## [13] "ANTIFA"            "antifa chardonnay" "anti-free speech"
## [16] "anti-racism"       "anti-racist"       "anti-western"
## [19] "anti-white"        "bias"
```

```
# make the words lower case
ALL_CRT <- tolower(word_set)
head(ALL_CRT, n = 20)
```

```
##  [1] "crt"               "chardonnayantifa"  "virginiademocrats"
##  [4] "activism"          "activist"          "activists"
##  [7] "anti cop"          "anti racist"       "anti white"
## [10] "anti-american"     "anti-american"     "anti-crt"
## [13] "antifa"            "antifa chardonnay" "anti-free speech"
## [16] "anti-racism"       "anti-racist"       "anti-western"
## [19] "anti-white"        "bias"
```

## Tweet Data

```
# scraped Twitter data
tweets_processed <- filtered_tweets
head(tweets_processed,5)
```

```
## # A tibble: 5 x 90
##   user_id   status_id           created_at          screen_name   text    source
##   <chr>     <chr>               <dttm>              <chr>         <chr>   <chr>
## 1 378776885 1531284775733972995 2022-05-30 14:41:55 JimMcNichols1 ".@Hun~ Twitt~
## 2 378776885 1514267542054965258 2022-04-13 15:41:30 JimMcNichols1 "Multi~ Twitt~
## 3 378776885 1491376235968040960 2022-02-09 11:39:37 JimMcNichols1 "Stude~ Twitt~
## 4 378776885 1452133856895766535 2021-10-24 04:44:25 JimMcNichols1 "EXCLU~ Twitt~
## 5 378776885 1456287648222416896 2021-11-04 15:50:06 JimMcNichols1 "If yo~ Twitt~
## # ... with 84 more variables: display_text_width <dbl>,
## #   reply_to_status_id <chr>, reply_to_user_id <chr>,
## #   reply_to_screen_name <chr>, is_quote <lgl>, is_retweet <lgl>,
## #   favorite_count <int>, retweet_count <int>, quote_count <int>,
## #   reply_count <int>, hashtags <list>, symbols <list>, urls_url <list>,
## #   urls_t.co <list>, urls_expanded_url <list>, media_url <list>,
## #   media_t.co <list>, media_expanded_url <list>, media_type <list>, ...
```

```
# take a look of the tweets

text <- tweets_processed$text
head(text, n = 5)
```

```
## [1] ".@HungCaoCongress escaped Vietnam, went to Thomas Jefferson High School, the Naval Academy, ther
## [2] "Multiple sources have confirmed that a special grand jury has been convened re: the investigatio
## [3] "Students whose families oppose masks arrive at a Loudoun County school board meeting carrying ba
## [4] "EXCLUSIVE: .@GlennYoungkin demands resignations from the Loudoun County School Board and Superin
## [5] "If you are actually interested in the truth, unlike Philip, this is the true story of how the Cl
```

## Process Tweet Data for analysis

```
# filter out non-English tweets
tweets <-  tweets_processed %>% filter(lang == "en")


# Exclude URL patterns in Text
tweets$text <- gsub("http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+]|[!*\\(\\),]|(?:%[0-9a-fA-F][0-9a-fA-F]))+",

# Exclude @user patterns in Text
tweets$text <- gsub("@\\w+", "" ,tweets$text)
#tweets$text <- gsub("\\w+", "" ,tweets$text)


# Exclude hashtags patterns in Text
tweets$text <- gsub("#", "" ,tweets$text)
```

## Check if a tweet contains any of the words of a wordlist

```
find_keywords <- function(text, keywords){
  text <- tolower(text)
  count <- 0
    for (ch in keywords) {
      if (grepl(ch, text)){

        if (ch != ""){
          count <- count + 1
        }
      }
   }
  return(count)
}
```

```
# apply the function to the data
tweets$ALL_CRT <- lapply(tweets$text, FUN = find_keywords, keywords = ALL_CRT)
```

```
# each tweet has a unique status id
# ALL_CRT: the number of keywords found in a particular tweet

df_test <- cbind(unlist(tweets$status_id), unlist(tweets$ALL_CRT))
head(df_test, n = 10)
```

```
##       [,1]                  [,2]
##  [1,] "1531284775733972995" "1"
##  [2,] "1514267542054965258" "0"
##  [3,] "1491376235968040960" "0"
##  [4,] "1452133856895766535" "1"
##  [5,] "1456287648222416896" "2"
##  [6,] "1452083977460797442" "0"
##  [7,] "1528797913299681284" "1"
##  [8,] "1487213380586905601" "0"
##  [9,] "1531613472924086272" "0"
## [10,] "1519472316115259392" "0"
```

```
# Recode all positive numbers as 1 and 0 as 0
tweets$ALL_CRT  <- ifelse(tweets$ALL_CRT  > 0, 1, 0)
df_test <- cbind(unlist(tweets$status_id), unlist(tweets$ALL_CRT))
head(df_test, n = 10)
```

```
##       [,1]                  [,2]
##  [1,] "1531284775733972995" "1"
##  [2,] "1514267542054965258" "0"
##  [3,] "1491376235968040960" "0"
##  [4,] "1452133856895766535" "1"
##  [5,] "1456287648222416896" "1"
##  [6,] "1452083977460797442" "0"
##  [7,] "1528797913299681284" "1"
##  [8,] "1487213380586905601" "0"
##  [9,] "1531613472924086272" "0"
## [10,] "1519472316115259392" "0"
```