Lecture 19 Notes

2024-04-09

```
require(tidyverse)
```

```
## Loading required package: tidyverse
```

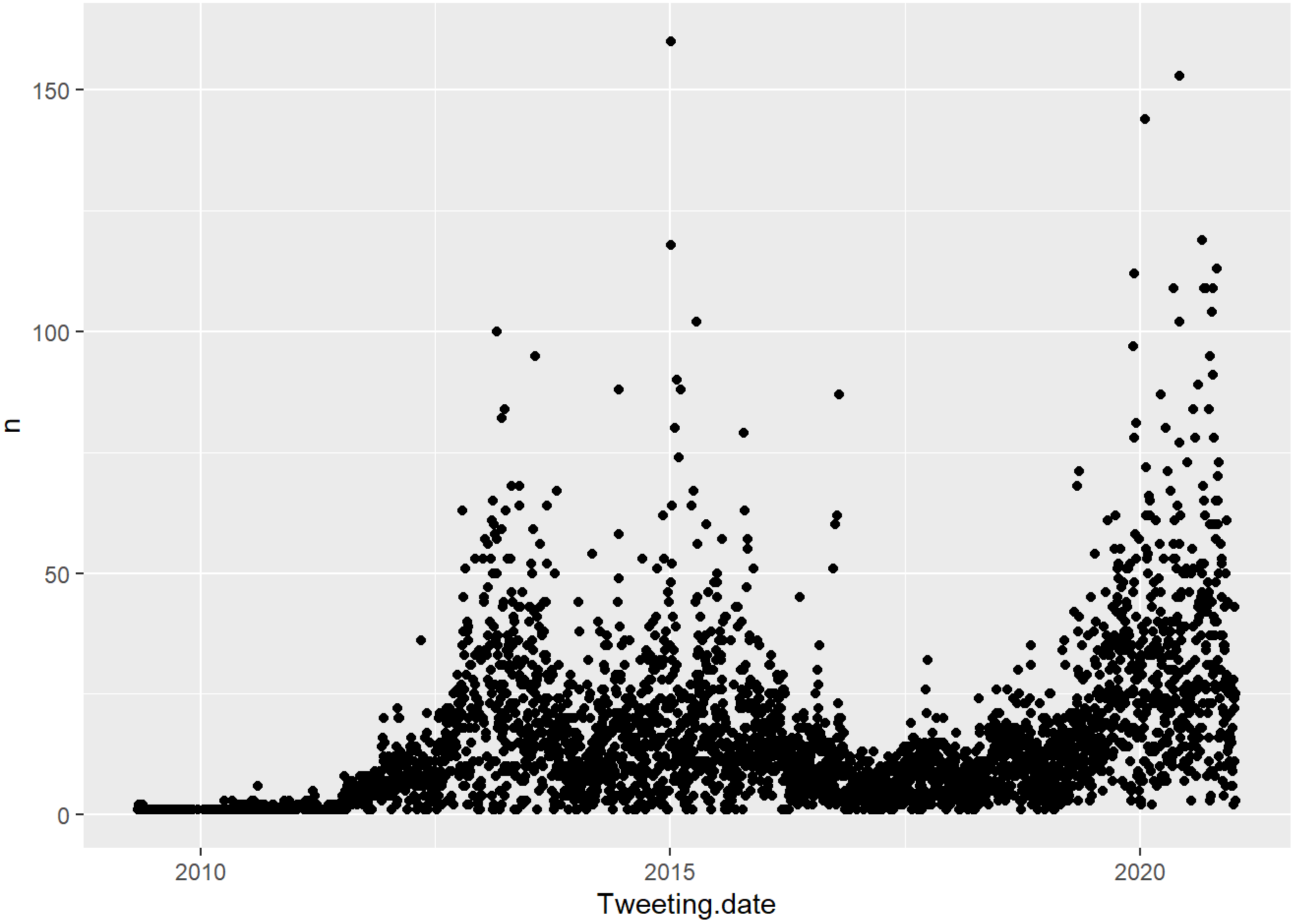
```
## Warning: package 'tidyverse' was built under R version 4.3.2
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.2     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.4     ✔ tibble    3.2.1
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ✔ purrr     1.0.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
tweets <- read_rds('https://github.com/jbisbee1/DS1000_S2024/raw/main/data/Trumptweets.Rds')

# Looking at tweet volume over time
tweets %>%
  count(Tweeting.date) %>%
  ggplot(aes(x = Tweeting.date,
             y = n)) +
  geom_point()
```

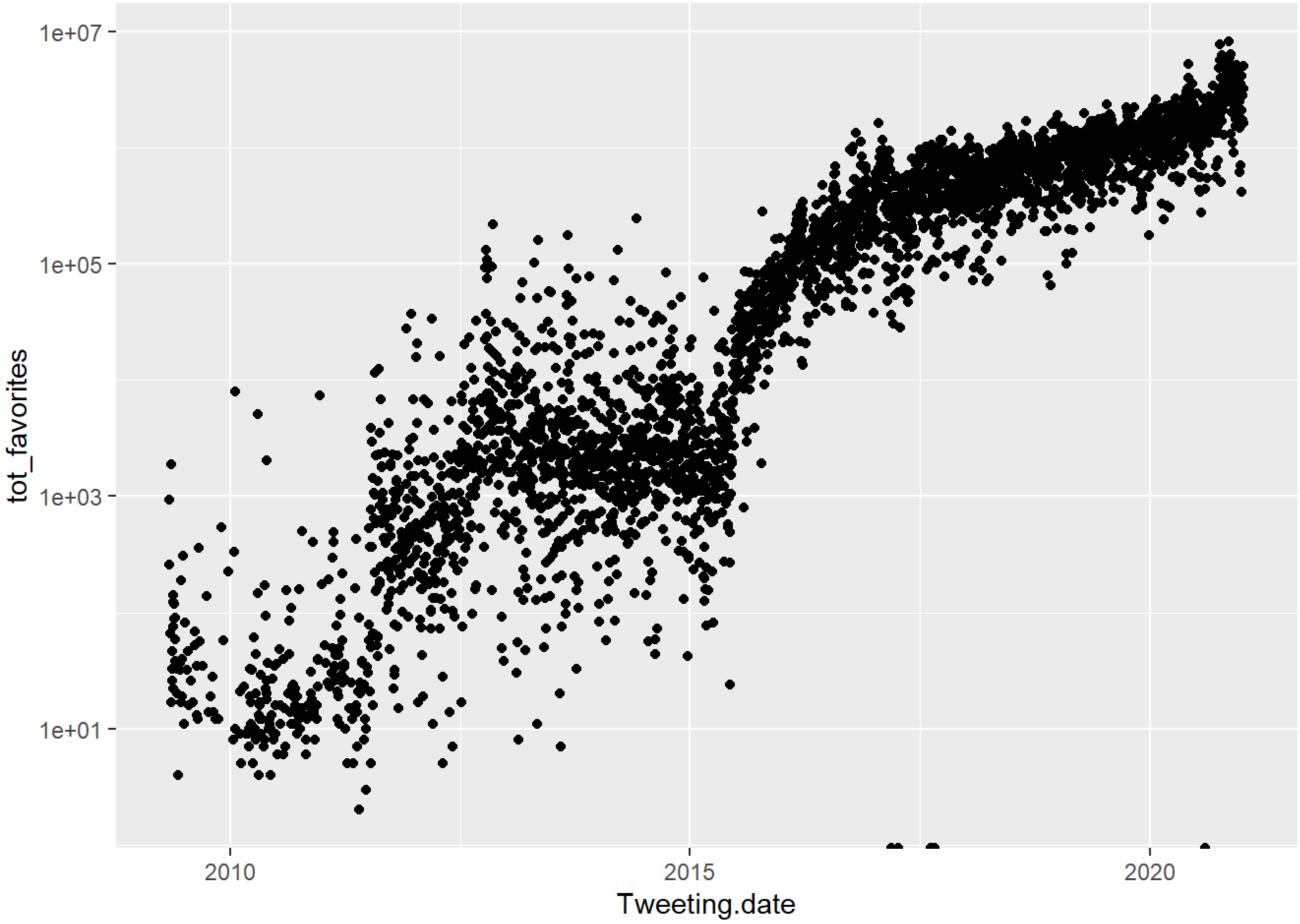


```
summary(tweets)
```

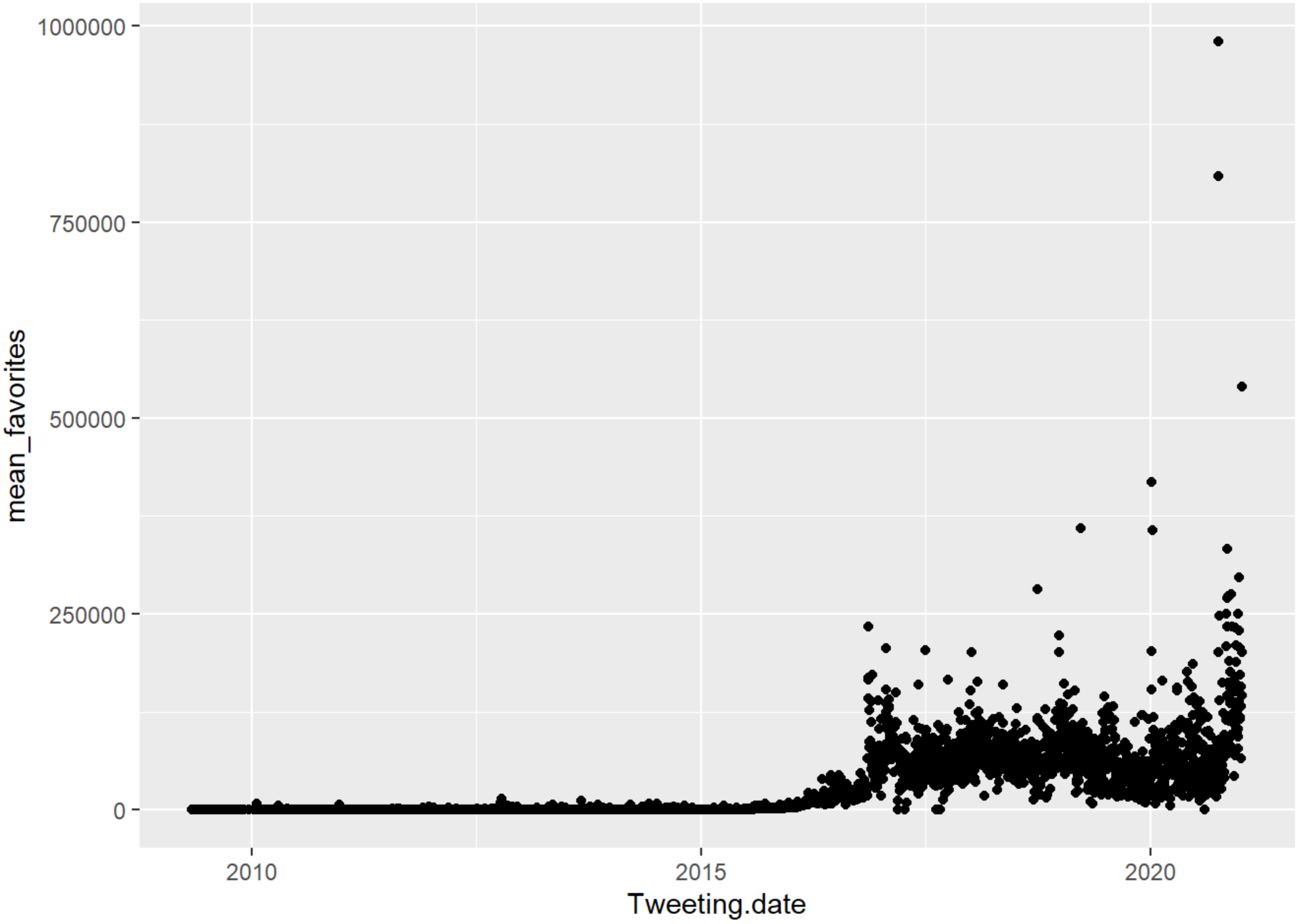
```
##        id              content           is_deleted      is_flagged
##  Min.   :1.698e+09   Length:56571       Mode :logical   Mode :logical
##  1st Qu.:4.606e+17   Class :character   FALSE:55479     FALSE:56267
##  Median :7.471e+17   Mode  :character   TRUE :1092      TRUE :304
##  Mean   :7.988e+17
##  3rd Qu.:1.193e+18
##  Max.   :1.348e+18
##
##     datetime                         retweets        favorites
##  Min.   :2009-05-04 18:54:25.00   Min.   :     0   Min.   :      0
##  1st Qu.:2014-04-28 03:19:18.50   1st Qu.:    59   1st Qu.:     10
##  Median :2016-06-26 16:21:15.00   Median :  3450   Median :    164
##  Mean   :2016-11-15 12:38:31.19   Mean   :  8619   Mean   :  28350
##  3rd Qu.:2019-11-09 11:51:37.00   3rd Qu.: 13014   3rd Qu.:  43939
##  Max.   :2021-01-08 15:44:28.00   Max.   :408866   Max.   :1869706
##
##  Tweeting.date        Tweeting.hour      Tweeting.year   content_clean
##  Min.   :2009-05-04   Length:56571       2020   :12236   Length:56571
##  1st Qu.:2014-04-27   Class :character   2013   : 8144   Class :character
##  Median :2016-06-26   Mode  :character   2019   : 7818   Mode  :character
##  Mean   :2016-11-14                      2015   : 7536
##  3rd Qu.:2019-11-09                      2014   : 5784
##  Max.   :2021-01-08                      2016   : 4225
##                                          (Other):10828
```

```
# Look at tweet popularity over time
tweets %>%
  group_by(Tweeting.date) %>%
  summarise(tot_favorites = sum(favorites,na.rm=T),
            mean_favorites = mean(favorites,na.rm=T)) %>%
  ggplot(aes(x = Tweeting.date,y = tot_favorites)) +
  geom_point() +
  scale_y_log10()
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```



```
tweets %>%
  group_by(Tweeting.date) %>%
  summarise(tot_favorites = sum(favorites,na.rm=T),
            mean_favorites = mean(favorites,na.rm=T)) %>%
  ggplot(aes(x = Tweeting.date,y = mean_favorites)) +
  geom_point()
```



Load the word-level data

```
tweet_words <- read_rds(file="https://github.com/jbisbee1/DS1000_S2024/raw/main/data/Trump_tweet_words.Rds")
```

Convert to DTM

```
tweet_words %>%
  count(word) %>%
  arrange(desc(n))
```

```
## # A tibble: 45,221 × 2
##    word          n
##    <chr>     <int>
##  1 trump      6269
##  2 president  4637
##  3 amp        4306
##  4 people     3475
##  5 country    2302
##  6 america    2211
##  7 time       1913
##  8 donald     1891
##  9 news       1842
## 10 democrats  1824
## # ℹ 45,211 more rows
```

```
# Create DTM
dtm <- tweet_words %>%
  filter(Tweeting.year == 2017) %>%
  count(document,word)

dtm %>%
  arrange(desc(n))
```

```
## # A tibble: 22,971 × 3
##    document word         n
##       <dbl> <chr>    <int>
##  1  8.23e17 bring        4
##  2  8.71e17 american     4
##  3  9.29e17 security     4
##  4  9.29e17 security     4
##  5  9.30e17 amp          4
##  6  9.41e17 amp          4
##  7  8.17e17 build        3
##  8  8.17e17 season       3
##  9  8.20e17 talk         3
## 10  8.28e17 charge       3
## # ℹ 22,961 more rows
```

Calculate TF-IDF

```
require(tidytext)
```

```
## Loading required package: tidytext
```

```
## Warning: package 'tidytext' was built under R version 4.3.2
```

```
dtm.tfidf <- bind_tf_idf(tbl = dtm,
            term = word,
            document = document,
            n = n)

dtm.tfidf %>%
  select(word,tf_idf) %>%
  arrange(desc(tf_idf))
```

```
## # A tibble: 22,971 × 2
##    word               tf_idf
##    <chr>               <dbl>
##  1 httpstcohoumbxgnpe   7.86
##  2 cpac                 7.86
##  3 httpstcordojtpip     7.86
##  4 httpstcogkockgndtc   7.86
##  5 httpstcowkqhymcya    7.86
##  6 httpstcoodlvpgjq     7.86
##  7 httpstcojpexqtvre    7.86
##  8 httpstcoeqnktg       7.86
##  9 httpstcoqckkpbtcr    7.86
## 10 httpstcowddeoivos    7.86
## # ℹ 22,961 more rows
```

Convert to wide format

```
wide_dtm <- cast_dtm(data = dtm.tfidf %>%
                       select(document,word,tf_idf),
                     document = document,
                     term = word,
                     value = tf_idf)

set.seed(123)
km_out <- kmeans(x = wide_dtm,
                 center = 50,
                 nstart = 5)
```

```
## Warning: did not converge in 10 iterations
```

```
## Warning: Quick-TRANSfer stage steps exceeded maximum (= 129850)
```