Clustering

Part 2

Prof. Bisbee

Vanderbilt University

Lecture Date: 2023/04/05

Slides Updated: 2023-03-31

Agenda

1. Tweets as data
2. Words → topics
3. Application

Social Media

- Unprecedented access to our leaders
  - (If they let us)

Social Media

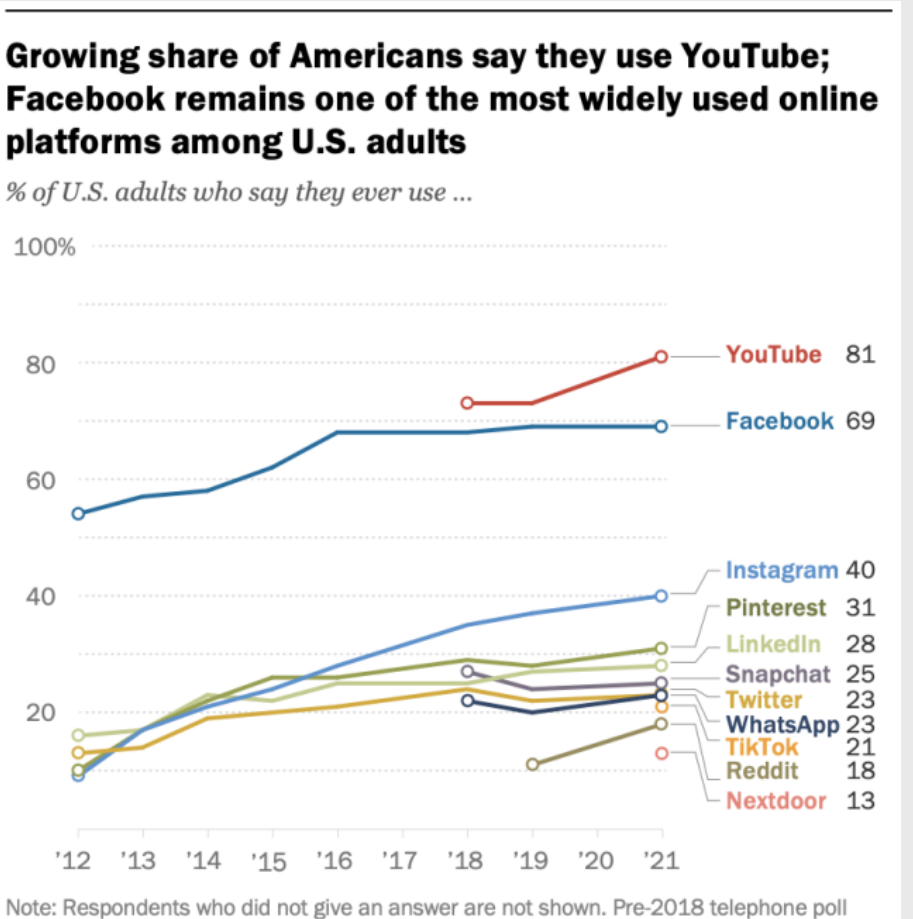
- Unprecedented access to our leaders
  - (If they let us)

Social Media

- For researchers, social media is two things
    1. A source of data
    2. An object of interest

Twitter as Data

- Not the most popular social media app



Twitter as Data

- But an outsized platform for the elite
- As of 2020
  - every U.S. governor had a Twitter account
  - 49 had a Facebook account
  - 44 had an Instagram account
  - 44 had a YouTube account
- In professional networks, particularly media, Twitter is almost *lingua franca*

Twitter as Data

- Today?

Trump and Twitter

- Today, looking at Trump's tweets
  - Treating it as a data source
  - What can his tweets tell us about the man?

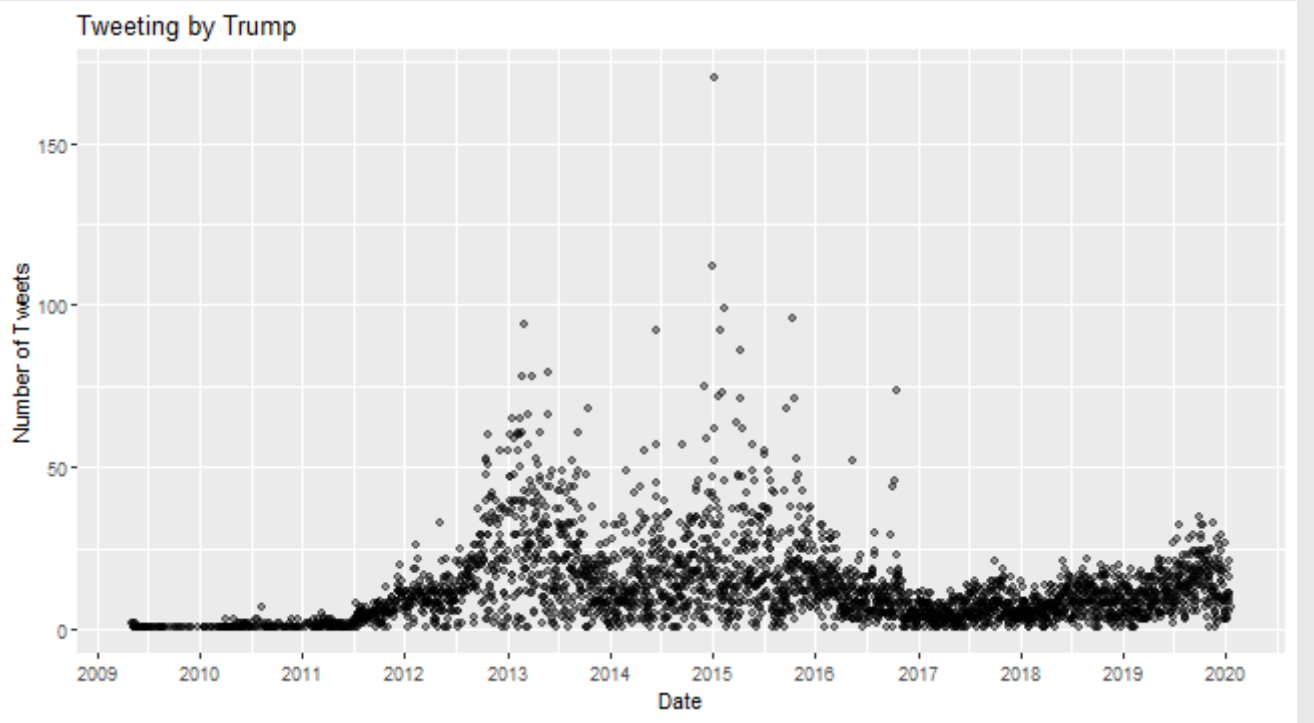
```
require(tidyverse)
tweets <- readRDS(file="../data/Trumptweets.Rds")
```

- The **process**
  - Univariate visualisation: how often does he tweet?

```
p <- tweets %>%
  count(Tweeting.date) %>%
  ggplot() +
  geom_point(aes(x=Tweeting.date,y=n),alpha=.4) +
  scale_x_date(date_breaks = 'years',date_labels = '%Y') +
  labs(x="Date",y="Number of Tweets",title="Tweeting by Trump")
```

Trump and Twitter

p



Trump and Twitter

- Research questions abound!

1. What happened in...
    - June of 2011?
    - June of 2012?
    - November of 2016? (duh)
2. Overtime increase during presidency?
3. Others?

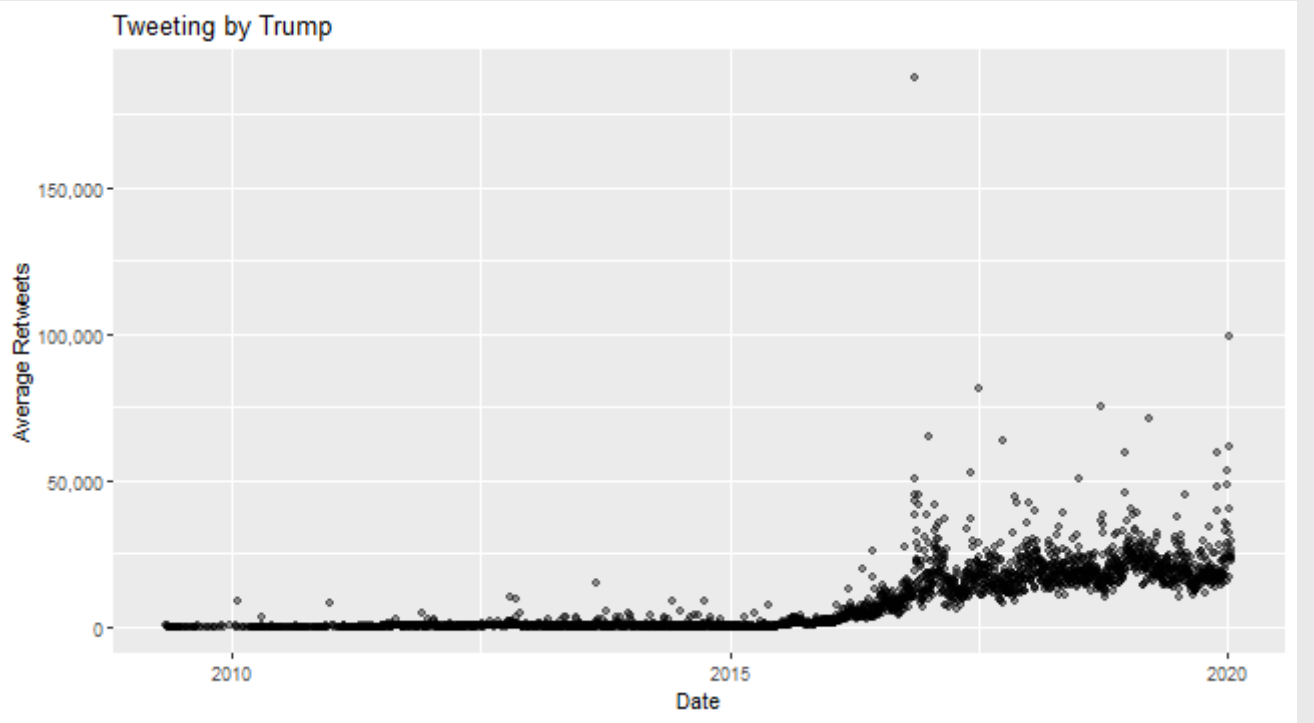
Trump and Twitter

- Research Question: Did Trump's twitter account benefit from his presidency?

```
require(scales)
p <- tweets %>%
  group_by(Tweeting.date) %>%
  summarize(AvgRetweet = mean(retweets)) %>%
  ggplot() +
  geom_point(aes(x=Tweeting.date,y=AvgRetweet),alpha=.4) +
  labs(x="Date",y="Average Retweets",title="Tweeting by Trump") +
  scale_y_continuous(label=comma)
```

Trump and Twitter

- Research Question: Did Trump's twitter account benefit from his presidency?
- **Yes**



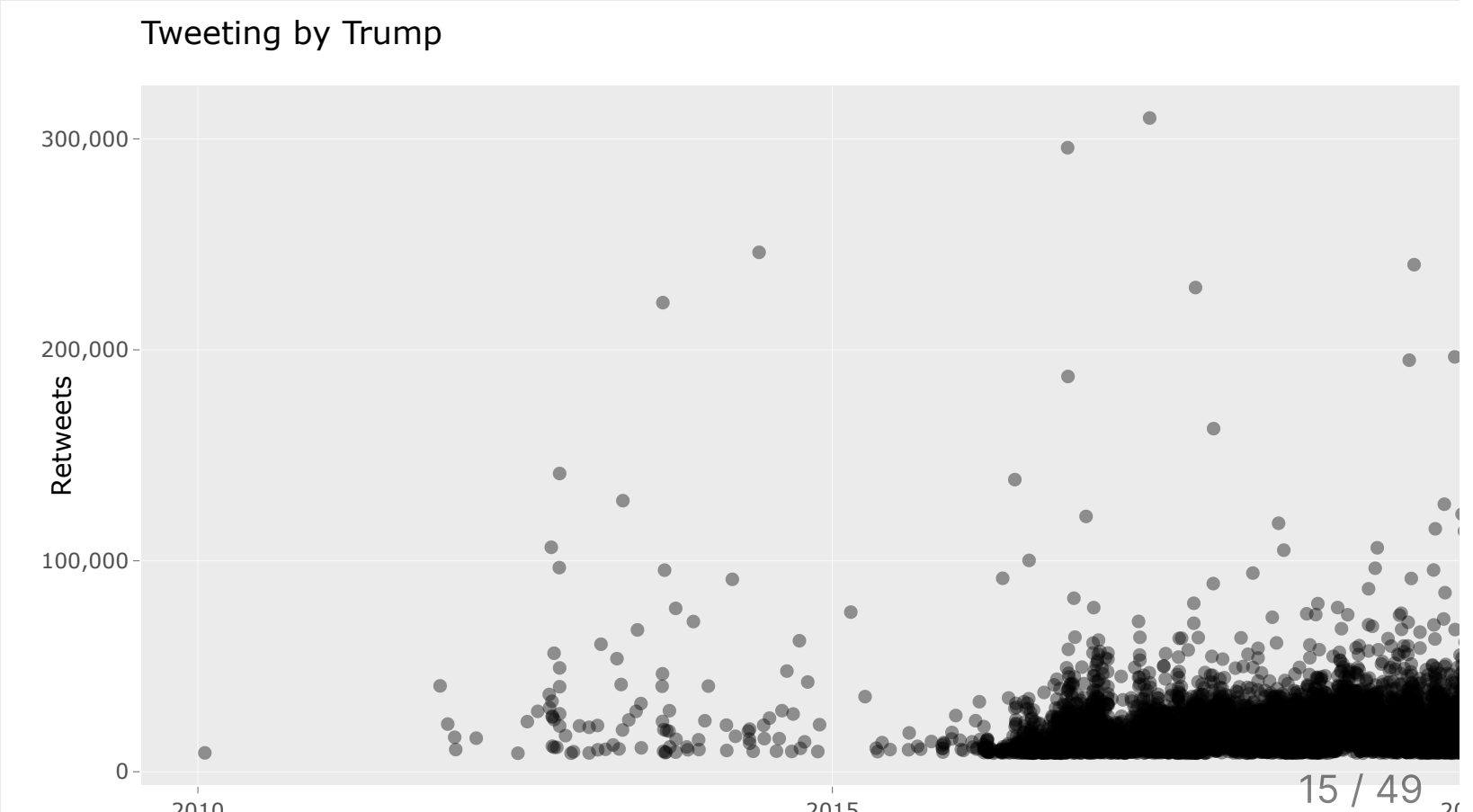
Looking with `plotly`

```
library(plotly)
gg <- tweets %>%
  filter(retweets > quantile(retweets,.75)) %>%

ggplot(aes(x=Tweeting.date,y=retweets,text=stringr::str_wrap(conten
= 60))) +
  geom_point(alpha=.4) +
  labs(x="Date",y="Retweets",title="Tweeting by Trump") +
  scale_y_continuous(label=comma)
```

Looking with `plotly`

```
ggplotly(gg,tooltip = "text")
```



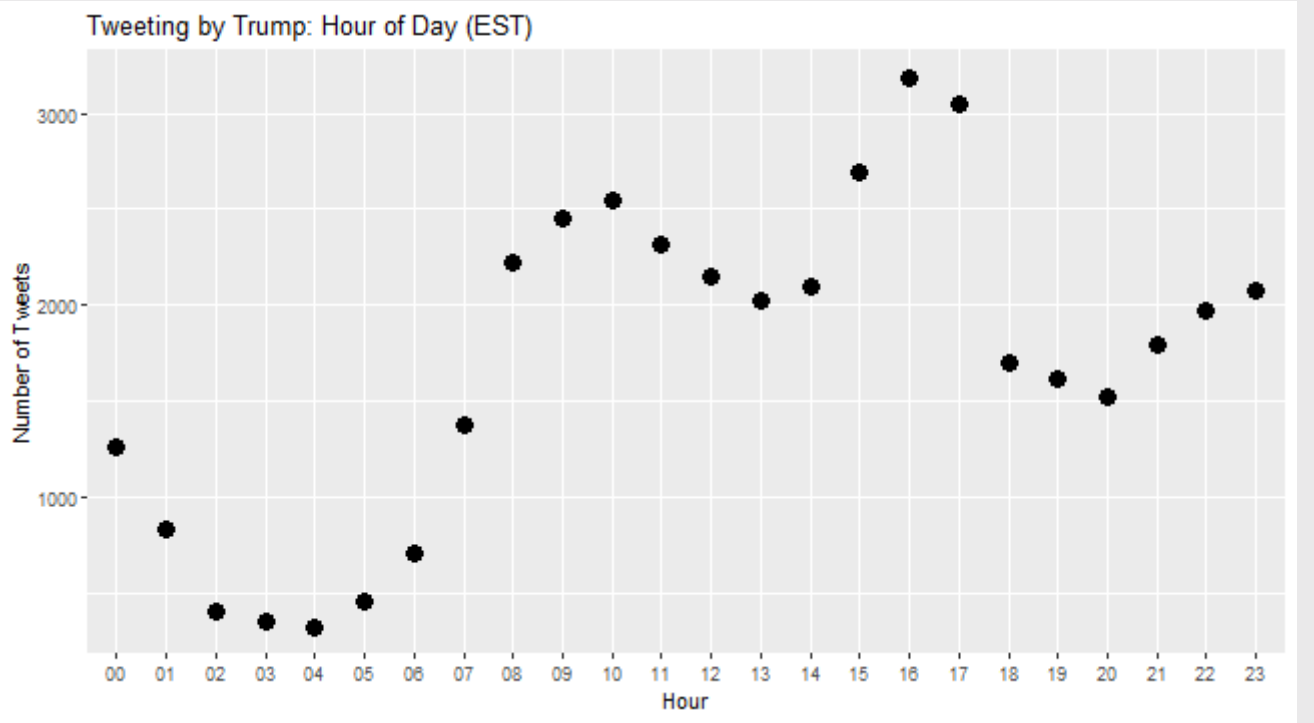
When is Trump online?

- Visualize posts by hour

```
p <- tweets %>%
  group_by(Tweeting.hour) %>%
  count() %>%
  ggplot() +
  geom_point(aes(x=Tweeting.hour,y=n),size = 4) +
  labs(x="Hour",y="Number of Tweets",title="Tweeting by Trump:
Hour of Day (EST)")
```

When is Trump online?

p



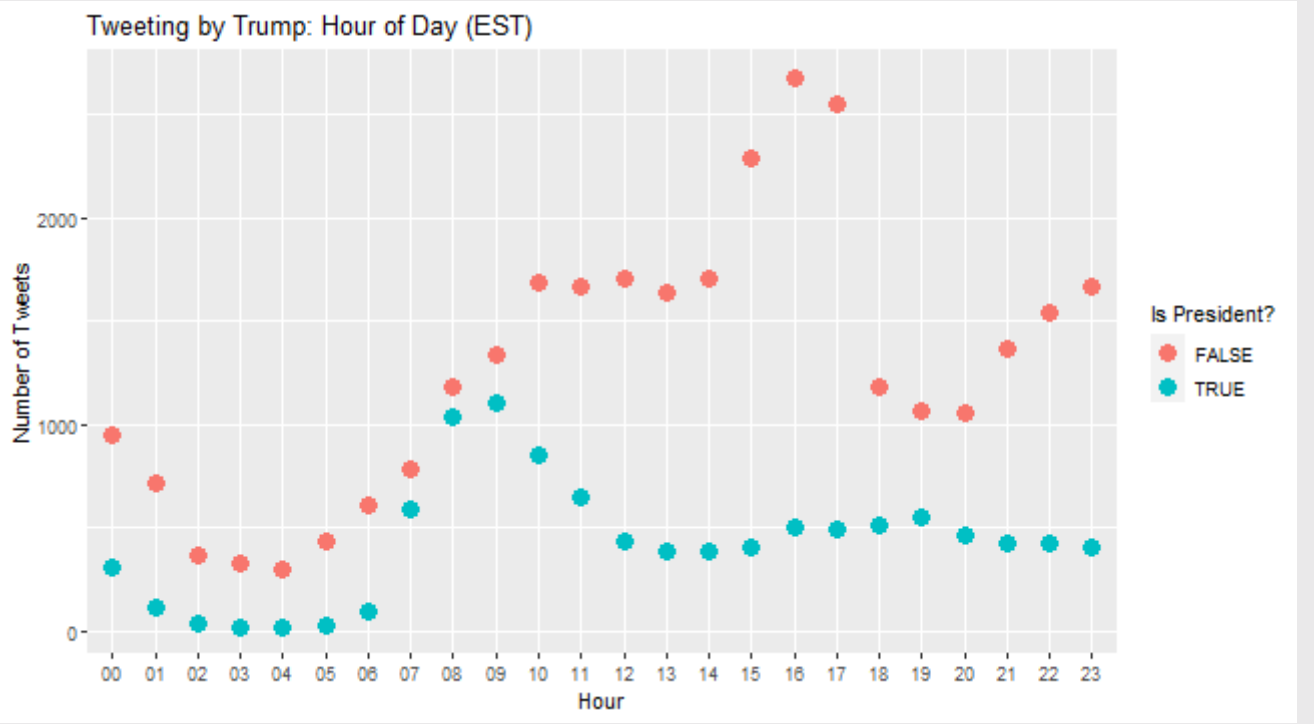
When is Trump online?

- Did his use of Twitter change with the presidency?
    - Certainly his popularity did

```
p <- tweets %>%
  mutate(PostPresident = date > "2016-11-03") %>%
  group_by(PostPresident,Tweeting.hour) %>%
  count() %>%
  ggplot() +
  geom_point(aes(x=Tweeting.hour,y=n,color=PostPresident),size =
4) +
  labs(x="Hour",y="Number of Tweets",title="Tweeting by Trump:
Hour of Day (EST)",color="Is President?")
```

When is Trump online?

p



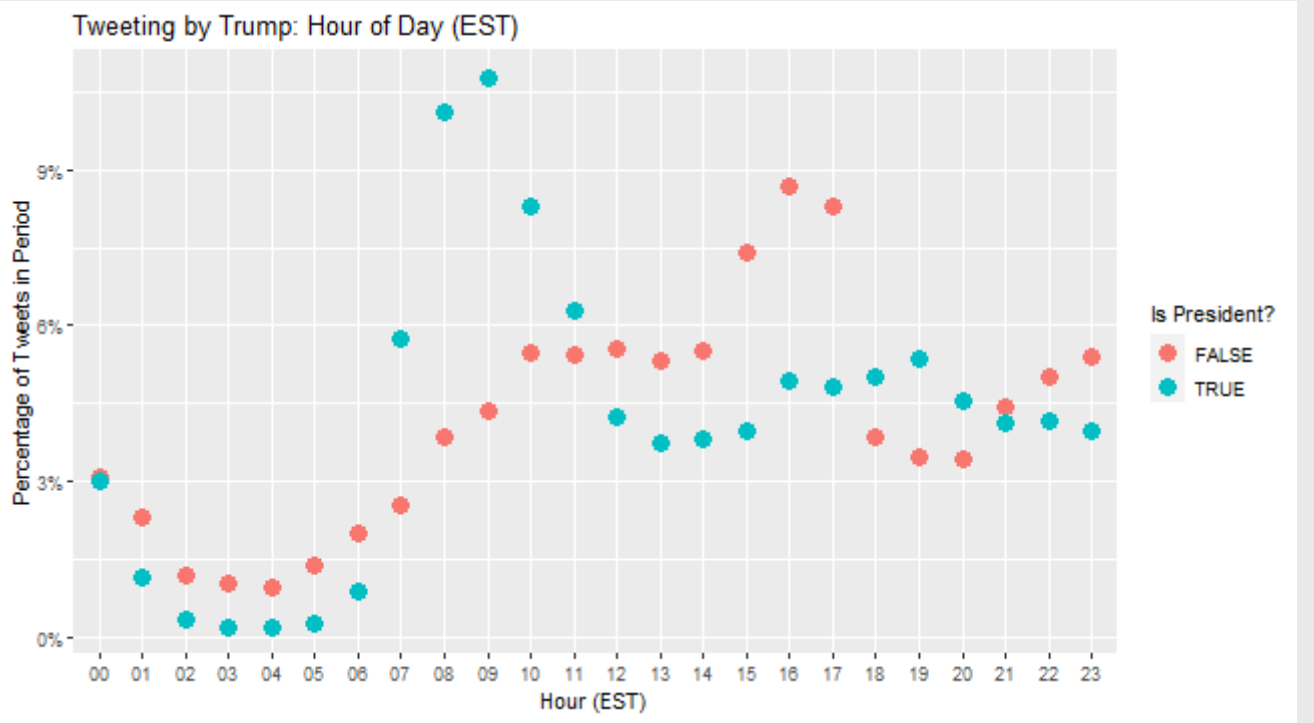
When is Trump online?

- Is the total count of tweets useful here?
    - Want the **proportion** instead

```
p <- tweets %>%
  mutate(PostPresident = date > "2016-11-03") %>%
  group_by(PostPresident,Tweeting.hour) %>%
  count() %>%
  ungroup(Tweeting.hour) %>%
  mutate(Prop = n/sum(n)) %>%
  ggplot() +
  geom_point(aes(x=Tweeting.hour,y=Prop,
                 color=PostPresident),size = 4) +
  labs(x="Hour (EST)",y="Percentage of Tweets in Period",
       title="Tweeting by Trump: Hour of Day (EST)",
       color="Is President?") +
  scale_y_continuous(labels = scales::percent_format(accuracy =
1))
```

When is Trump online?

p



How does Trump tweet?

- Interested in the content of Trump's tweets
    1. Tweets as data: we can understand him better
    2. Tweets as object of interest: how do his tweets influence public discourse?
- We will be using pre-processed tweets
    - Lots of **data wrangling** went into this
    - See the `trump_preprocess.Rmd` file for more details

```
tweet_words <- readRDS(file="../data/Trump_tweet_words.Rds")
```

NLP Definitions

- Before we dig in, some important definitions

1. **Word / Term:** The core unit of interest
    - Often pre-processed to remove "stop words" and to "stem" the words
    - "Stop word": an uninteresting, commonly used word
    - "Stem / Lemmatize": the core component of a word that contains its meaning (eat, ate, eaten → eat)
2. **Document:** A collection of words with a single purpose / idea (i.e., a tweet, an essay)
3. **Corpus:** A collection of documents
4. **BOW:** Bag-of-words. Convert a **document** into a count of how many times **words** appear
5. **DTM:** Document-term matrix. A dataset where rows are **documents**,

What does Trump tweet about?

- Core idea: word frequencies can help:
    - Help us understand **documents**
    - Which help us understand **authors**

```
counts <- tweet_words %>%
  count(word) %>%
  arrange(-n)
```

What does Trump tweet about?

```
counts
```

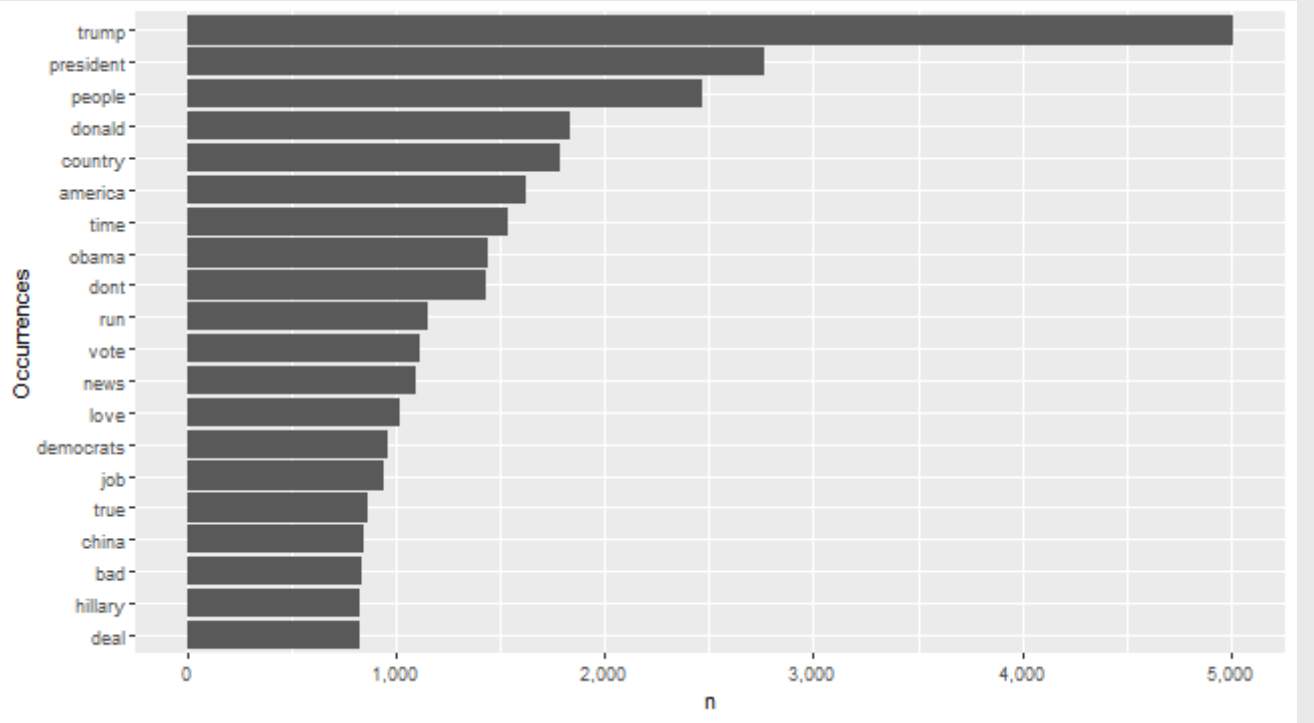
```
## # A tibble: 23,453 × 2
##    word           n
##    <chr>      <int>
##  1 trump       5010
##  2 president   2766
##  3 people      2465
##  4 donald      1833
##  5 country     1788
##  6 america     1627
##  7 time        1540
##  8 obama       1440
##  9 dont        1434
## 10 run         1152
## # … with 23,443 more rows
```

What does Trump tweet about?

```
p <- tweet_words %>%
  count(word, sort = TRUE) %>% # New ways of doing old things
  head(20) %>%
  ggplot(aes(x = n,y = reorder(word, n))) +
  geom_bar(stat = "identity") +
  ylab("Occurrences") +
  scale_x_continuous(label=comma)
```

What does Trump tweet about?

p



Effect of becoming president

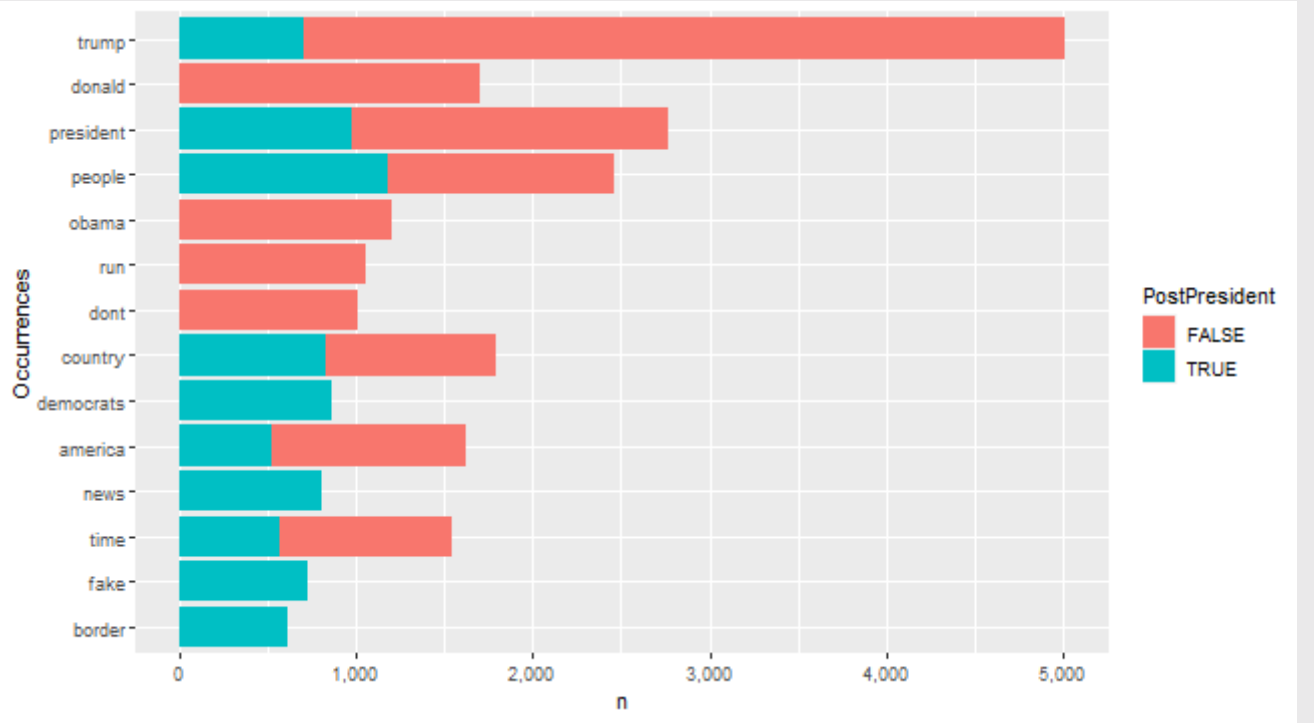
- Did his focus change when he became president?

```
tweet_words <- tweet_words %>%
  mutate(PostPresident = Tweeting.date > as.Date("2016-11-03"))
```

```
p <- tweet_words %>%
  count(PostPresident,word) %>%
  group_by(PostPresident) %>%
  arrange(-n) %>%
  slice(1:10) %>%
  ggplot(aes(x = n,y = reorder(word, n),
             fill = PostPresident)) +
  geom_bar(stat = "identity") +
  ylab("Occurrences") +
  scale_x_continuous(label=comma)
```

Effect of becoming president

p



Document Term Matrix

- "DTM" counts all the words in each document
- In this case, a "document" is a tweet

```
dtm <- tweet_words %>%
  count(document,word)
glimpse(dtm)
```

```
## Rows: 289,377
## Columns: 3
## $ document <dbl> 1698308935, 1698308935, 1698308935, 16983…
## $ word     <chr> "david", "donald", "late", "letterman", "…
## $ n        <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,…
```

Document Term Matrix

- However, each tweet is very short
- Let's consider the `Tweeting.date` the document
    - Concept: What is Trump tweeting about on a given day?

```
dtm <- tweet_words %>%
  count(Tweeting.date,word) %>%
  group_by(word) %>%
  mutate(tot_n = sum(n)) %>% # Also calculate TOTAL number of
times a word appears
  ungroup()
```

Wrangle

- Extremely rare words should be dropped (typos, etc.)

```
dtm %>%
  arrange(tot_n) %>% head() # Trump tweeted "barnesandnoblecom"
only once in his life
```

```
## # A tibble: 6 × 4
##   Tweeting.date word                n tot_n
##   <date>        <chr>           <int> <int>
## 1 2009-05-12    cling               1     1
## 2 2009-05-12    tara                1     1
## 3 2009-05-12    wallflower          1     1
## 4 2009-05-16    achiever            1     1
## 5 2009-05-16    achieves            1     1
## 6 2009-05-19    barnesandnoblecom   1     1
```

```
dtm <- dtm %>%
  filter(tot_n > 20) # Drop these rarely occurring words
```

Wordcloud

- One of the cheesiest "visualizations"
- Summarizes what corpus is "about"

```
library(wordcloud)
wordcloud(words = dtm$word,
          freq = dtm$n,
          max.words = 200,
          random.order=FALSE,
          rot.per=0.35)
```

Wordcloud

- Can apply to a subset of documents

```
# Pre-presidency
dtmPre <- dtm %>%
  filter(Tweeting.date < as.Date('2016-01-01'))  %>%
  arrange(-n)
wordcloud(words = dtmPre$word,
          freq = dtmPre$n,
          max.words = 20,
          rot.per=0.35)
```

Wordcloud

- Can apply to a subset of documents

```
# Post-presidency
dtmPost <- dtm %>%
  filter(Tweeting.date > as.Date('2016-11-01'))  %>%
  arrange(-n)
wordcloud(words = dtmPost$word,
          freq = dtmPost$n,
          max.words = 20,
          rot.per=0.35)
```

Conclusion?

- What might you infer from this comparison?
- The presidency made Trump less vain?

Analyzing BOW

- Some words are frequently found in many documents
- We want to find words that are **uniquely** used
  - "Unique" → used frequently in one document but not in any others
- "TF-IDF": Term frequency-inverse document frequency
- "TF": $\frac{\text{word count}}{\text{total words}}$
- "DF": $\frac{\text{documents with word}}{\text{total documents}}$
  - "IDF": Just invert it $\frac{\text{total documents}}{\text{documents with word}}$

$$tf - idf(w, d) = tf(w, d) \times log\left(\frac{N}{df(w)}\right)$$

TF-IDF

```
require(tidytext) # Required to calculate TF-IDF
dtm.tfidf <- bind_tf_idf(tbl = dtm, term = word, document =
Tweeting.date, n = n) # Calculate TF-IDF
dtm.tfidf  %>%
  select(word,tf_idf) %>%
  distinct() %>%
  arrange(-tf_idf) %>%
  slice(1:10)
```

```
## # A tibble: 10 × 2
##    word       tf_idf
##    <chr>       <dbl>
##  1 cleveland    2.61
##  2 ego          2.58
##  3 weekly       2.46
##  4 apprentice   2.46
##  5 august       2.35
##  6 appearance   2.25
##  7 wsj          2.18
##  8 fame         2.11
##  9 defeat       2.07
```

K-means

- How to summarize this? k-means clustering!
- Recall that `kmeans()` function clusters over every column in a data frame
- But our `dtm.tfidf` is organized "long" (i.e., each row is a word-by-document)
- Want to convert to "wide" (i.e., rows are documents, columns are words)
- In the past, we have used `spread()`, but for *k*-means with text: `cast_dtm()`

```
castdtm <- cast_dtm(data = dtm.tfidf, document = Tweeting.date,
term = word, value = tf_idf)
```

- Now let's calculate `kmeans()` (this will take a few seconds)

```
set.seed(42)
```

Looking at Clusters

- Some quick wrangling

```
km_out_tidy <- tidy(km_out) %>%
  gather(word,mean_tfidf,-size,-cluster,-withinss) %>% # Convert to long data
  mutate(mean_tfidf = as.numeric(mean_tfidf)) # Calculate average TF-IDF
km_out_tidy
```

```
## # A tibble: 118,350 × 5
##     size withinss cluster word       mean_tfidf
##    <int>    <dbl> <fct>   <chr>           <dbl>
##  1     1     0    1       apprentice   0
##  2    63    44.5  2       apprentice   0
##  3  1076   289.   3       apprentice   0.000296
##  4     3     4.75 4       apprentice   0
##  5   163    53.1  5       apprentice   0.00394
##  6     1     0    6       apprentice   0
##  7     3     5.18 7       apprentice   0
##  8    15     8.41 8       apprentice   0.0193
##  9     4    12.0  9       apprentice   0
```

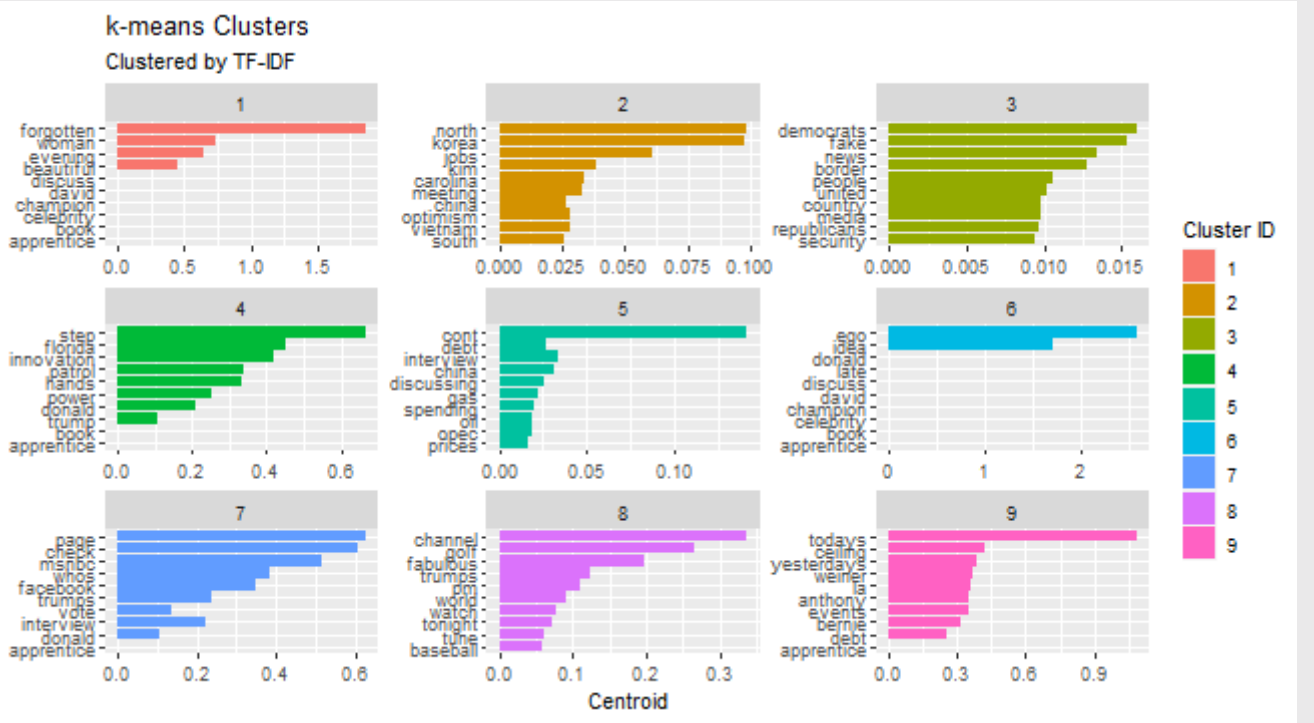
Looking at Clusters

- And can plot! (Just look at first 10 "topics")

```
p <- km_out_tidy %>%
  filter(cluster %in% 1:9) %>%
  group_by(cluster) %>%
  arrange(-mean_tfidf) %>%
  slice(1:10) %>%
  ggplot(aes(x = mean_tfidf,y = reorder(word,mean_tfidf),
             fill = factor(cluster))) +
  geom_bar(stat = 'identity') +
  facet_wrap(~cluster,scales = 'free') +
  labs(title = 'k-means Clusters',
       subtitle = 'Clustered by TF-IDF',
       x = 'Centroid',
       y = NULL,
       fill = 'Cluster ID')
```

Looking at Clusters

p



Looking at clusters

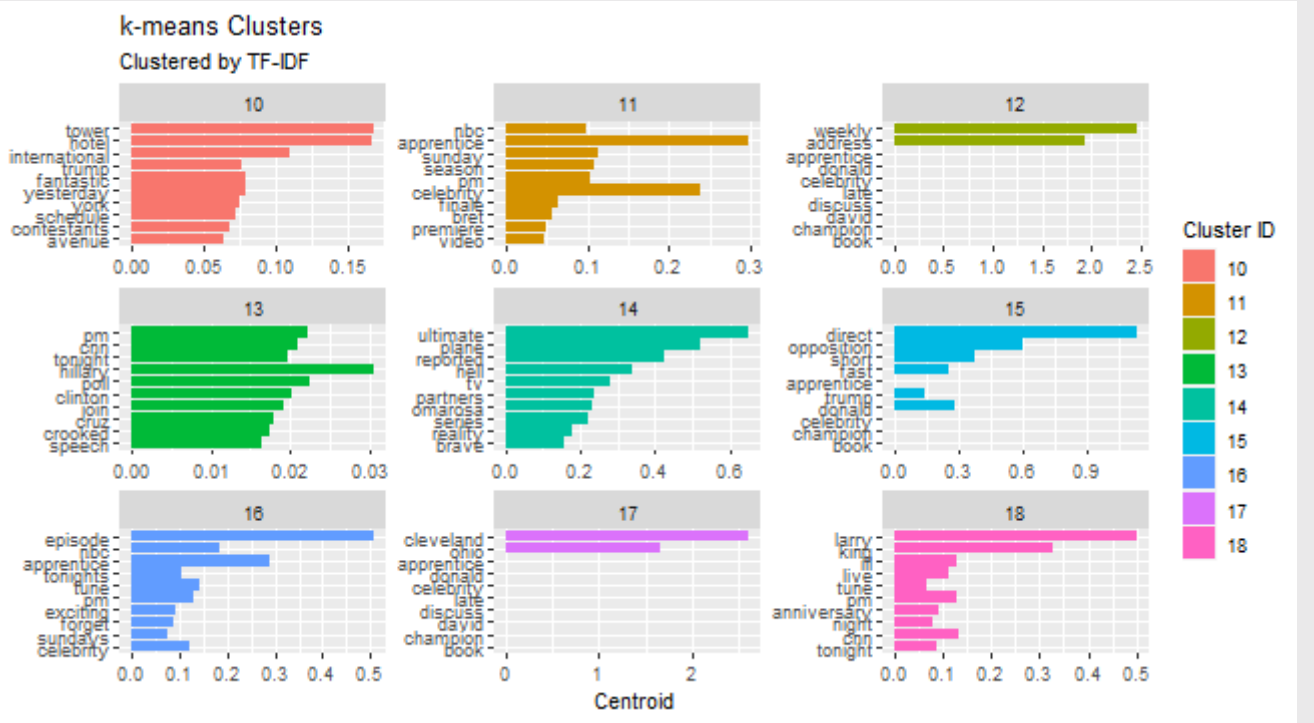
- What are these clusters?
- Topic 2?
  - Looks like foreign policy, specifically Asia and the U.S. economy!
- Topic 3?
  - Looks like domestic policy, specifically partisanship and the border!
- Topic 8?
  - Sports?

Other topics?

```
p <- km_out_tidy %>%
  filter(cluster %in% 10:18) %>%
  group_by(cluster) %>%
  arrange(-mean_tfidf) %>%
  slice(1:10) %>%
  ggplot(aes(x = mean_tfidf,y = reorder(word,mean_tfidf),
             fill = factor(cluster))) +
  geom_bar(stat = 'identity') +
  facet_wrap(~cluster,scales = 'free') +
  labs(title = 'k-means Clusters',
       subtitle = 'Clustered by TF-IDF',
       x = 'Centroid',
       y = NULL,
       fill = 'Cluster ID')
```

Other topics?

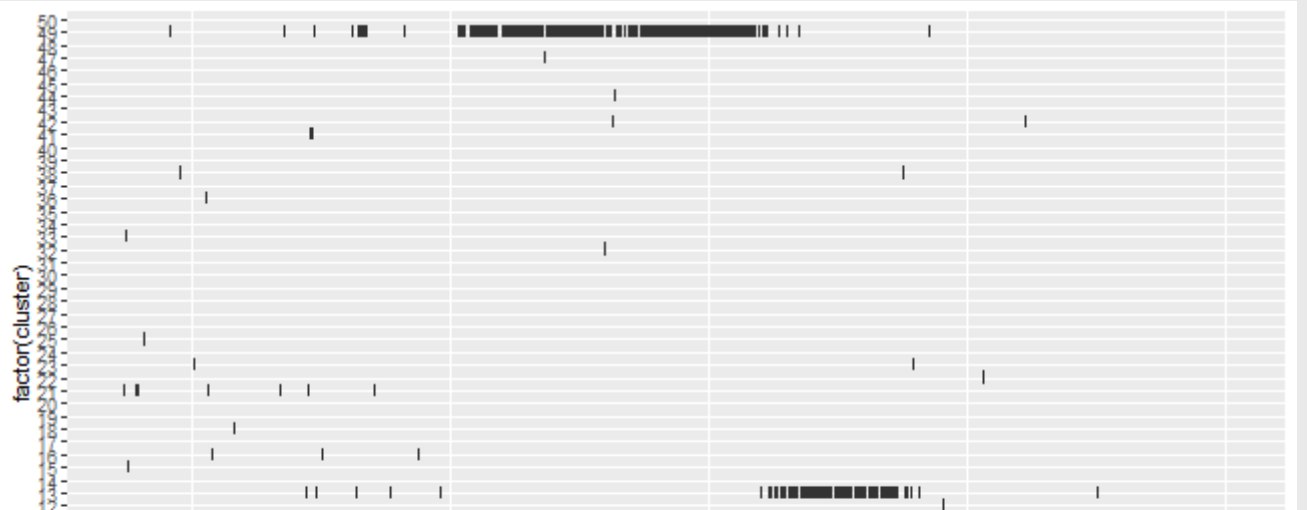
p



Apply to documents

- See when different topics are emphasized

```
data.frame(Tweeting.date = castdtm$dimnames$Docs,
           cluster = km_out$cluster) %>%
  as_tibble() %>%
  mutate(Tweeting.date = as.Date(as.numeric(Tweeting.date),origin
= '1970-01-01')) %>%
  ggplot(aes(x = Tweeting.date,y = factor(cluster))) +
  geom_tile()
```



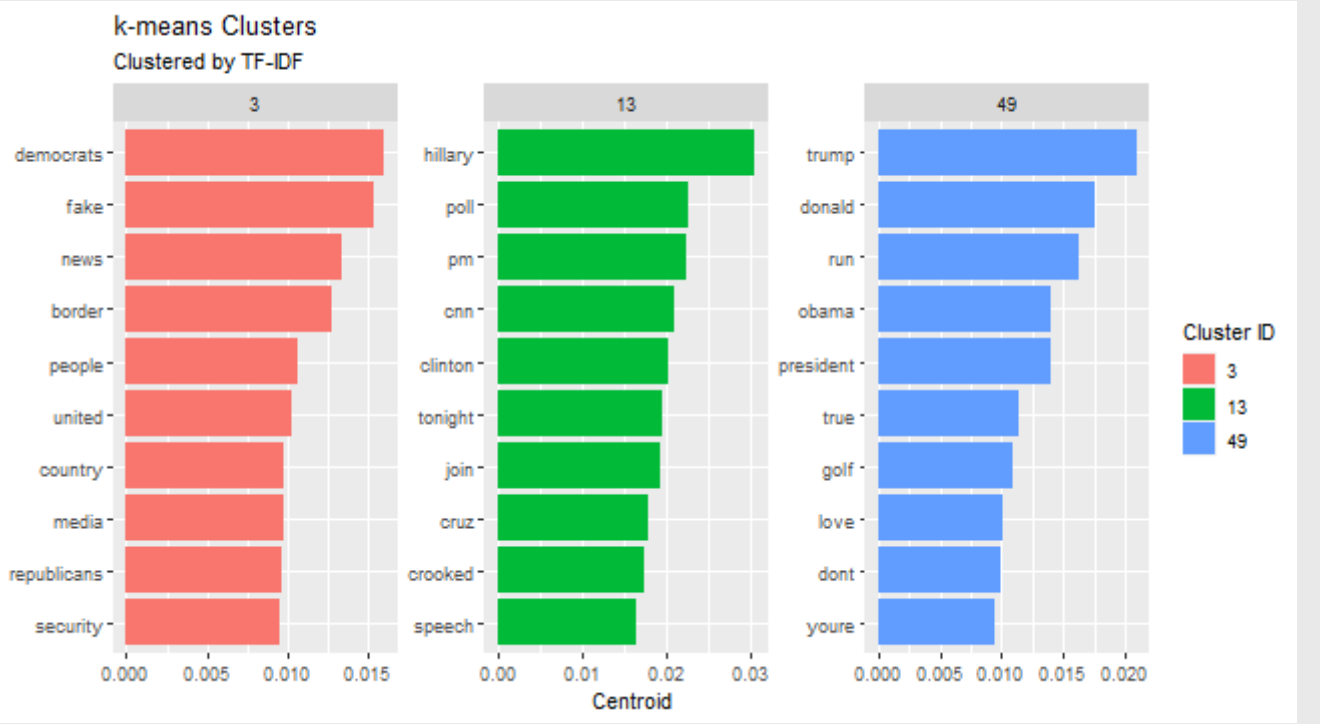
Checking on topics

```
p <- km_out_tidy %>%
  filter(cluster %in% c(3,13,49)) %>%
  group_by(cluster) %>%
  arrange(-mean_tfidf) %>%
  slice(1:10) %>%
  ggplot(aes(x = mean_tfidf,y = reorder(word,mean_tfidf),
             fill = factor(cluster))) +
  geom_bar(stat = 'identity') +
  facet_wrap(~cluster,scales = 'free') +
  labs(title = 'k-means Clusters',
       subtitle = 'Clustered by TF-IDF',
       x = 'Centroid',
       y = NULL,
       fill = 'Cluster ID')
```

Checking on topics

- Trump talks about 49 prior to presidency, 13 during 2016 campaign, and 3 after becoming president

p



Conclusion

- k-means clustering on text → **topics**
- As always, this is a deep area of study
    - Superior methods are out there (Latent Dirichlet Allocation, Structural Topic Modeling, etc.)
- NOTE: even with text, always start with simple descriptives
    - **Looking** at your data is the heart of data science