STAT 231: Problem Set 6B

Kevin Ma

due by 10 PM on Friday, April 2

This homework assignment is designed to help you further ingest, practice, and expand upon the material covered in class over the past week(s). You are encouraged to work with other students, but all code and text must be written by you, and you must indicate below who you discussed the assignment with (if anyone).

Steps to proceed:

1. In RStudio, go to File > Open Project, navigate to the folder with the course-content repo, select the course-content project (course-content.Rproj), and click "Open"
2. Pull the course-content repo (e.g. using the blue-ish down arrow in the Git tab in upper right window)
3. Copy ps6B.Rmd from the course repo to your repo (see page 6 of the GitHub Classroom Guide for Stat231 if needed)
4. Close the course-content repo project in RStudio
5. Open YOUR repo project in RStudio
6. In the ps6B.Rmd file in YOUR repo, replace "YOUR NAME HERE" with your name
7. Add in your responses, committing and pushing to YOUR repo in appropriate places along the way
8. Run "Knit PDF"
9. Upload the pdf to Gradescope. Don't forget to select which of your pages are associated with each problem. *You will not get credit for work on unassigned pages (e.g., if you only selected the first page but your solution spans two pages, you would lose points for any part on the second page that the grader can't see).*

If you discussed this assignment with any of your peers, please list who here:

ANSWER: TA Hours with Andrea

Trump Tweets

David Robinson, Chief Data Scientist at DataCamp, wrote a blog post “Text analysis of Trump’s tweets confirms he writes only the (angrier) Android half”.

He provides a dataset with over 1,500 tweets from the account realDonaldTrump between 12/14/2015 and 8/8/2016. We’ll use this dataset to explore the tweeting behavior of realDonaldTrump during this time period.

First, read in the file. Note that there is a `TwitteR` package which provides an interface to the Twitter web API. We’ll use this R dataset David created using that package so that you don’t have to set up Twitter authentication.

```
load(url("http://varianceexplained.org/files/trump_tweets_df.rda"))
#Loading doesn't work for some reason but this works:
load("/Users/kevinma/Downloads/trump_tweets_df.rda")
```

A little wrangling to warm-up

1a. There are a number of variables in the dataset we won’t need.

- First, confirm that all the observations in the dataset are from the screen-name `realDonaldTrump`.
- Then, create a new dataset called `tweets` that only includes the following variables:
- `text`
- `created`
- `statusSource`

```
#Check if column 11 are all the same
length(unique(trump_tweets_df[,11]))==1
```

```
## [1] TRUE
```

```
#select for wanted columns
tweets <- trump_tweets_df %>%
  select("text", "created", "statusSource")
glimpse(tweets)
```

```
## Rows: 1,512
## Columns: 3
## $ text         <chr> "My economic policy speech will be carried live at 12:...
## $ created      <dttm> 2016-08-08 15:20:44, 2016-08-08 13:28:20, 2016-08-08 ...
## $ statusSource <chr> "<a href=\"http://twitter.com/download/android\" rel=\...
```

1b. Using the `statusSource` variable, compute the number of tweets from each source. How many different sources are there? How often are each used?

> ANSWER: There 5 different sources. There was 1 tweet from Instagram, 762 from Android, 1 from iPad, 628 from iPhone, and 120 from Web Client.

```
#Data wrangling, remove unwanted string characters
sources <- tweets %>%
  mutate(
    statusSource = gsub("</a>", "", statusSource),
    statusSource = gsub(".*>", "", statusSource)
  )
#Get number of unique sources, create a table of those shources and how many tweets from each source
unique(sources$statusSource)
```

```
## [1] "Twitter for Android" "Twitter for iPhone"  "Twitter Web Client"
## [4] "Twitter for iPad"    "Instagram"
```

```
table(sources$statusSource)
```

```
##
##           Instagram Twitter for Android    Twitter for iPad  Twitter for iPhone
##                   1                 762                   1                 628
##  Twitter Web Client
##                 120
```

1c. We're going to compare the language used between the Android and iPhone sources, so only want to keep tweets coming from those sources. Explain what the `extract` function (from the `tidyverse` package) is doing below. Include in your own words what each argument is doing. (Note that "regex" stands for "regular expression".)

> ANSWER: The extract function is taking from the column "statusSource" and extracting into a new column called "source". The regex argument is helping to extract matching patterns from a string. The remove argument prevents the deletion of the column statusSource which is automatic if you don't specify not to delete.

```
tweets2 <- tweets %>%
  extract(col = statusSource, into = "source"
          , regex = "Twitter for (.*)<"
          , remove = FALSE) %>%
  filter(source %in% c("Android", "iPhone"))
```

How does the language of the tweets differ by source?

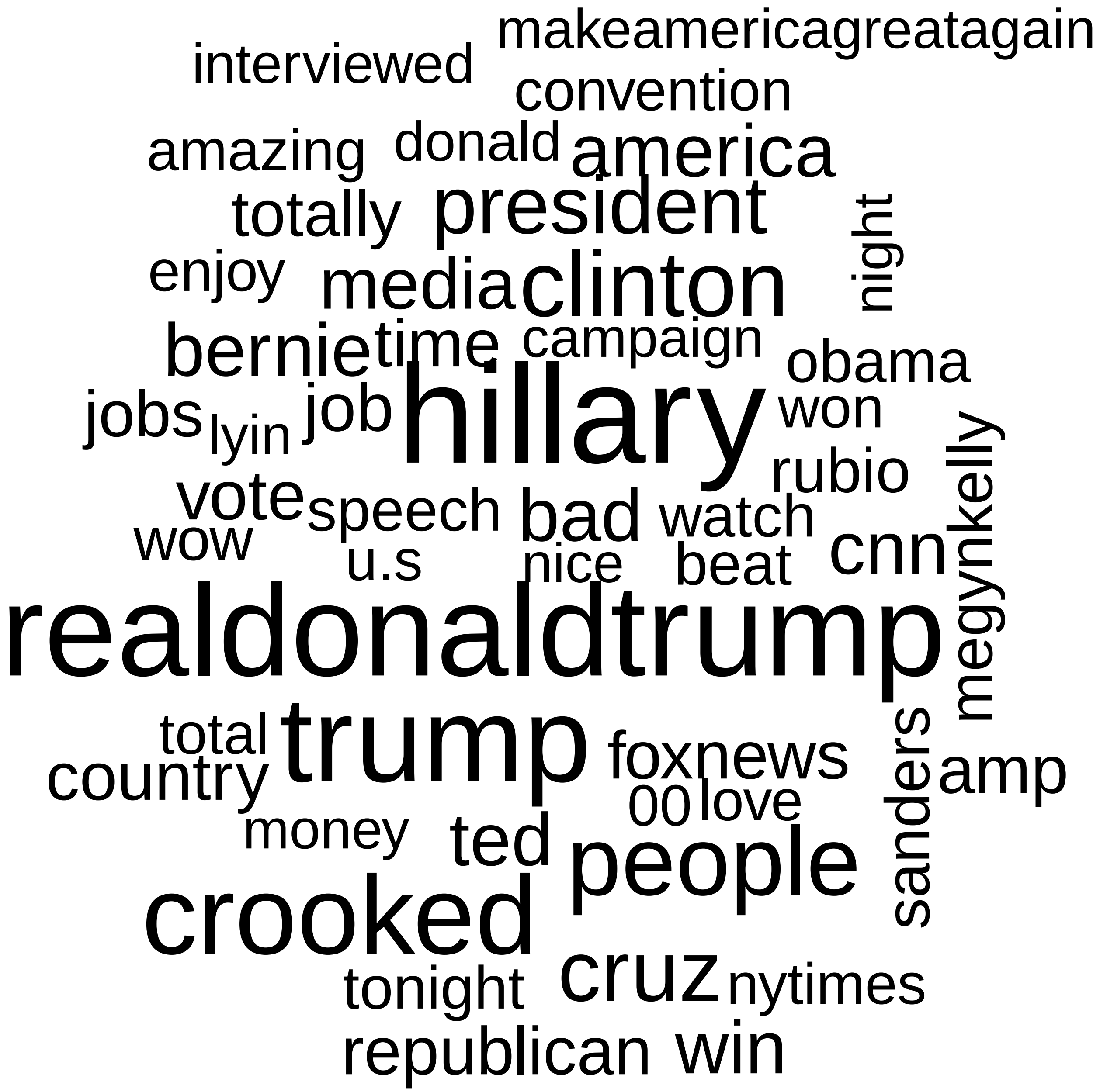
2a. Create a word cloud for the top 50 words used in tweets sent from the Android. Create a second word cloud for the top 50 words used in tweets sent from the iPhone. How do these word clouds compare? (Are there some common words frequently used from both sources? Are the most common words different between the sources?)

Don't forget to remove stop words before creating the word cloud. Also remove the terms "https" and "t.co".

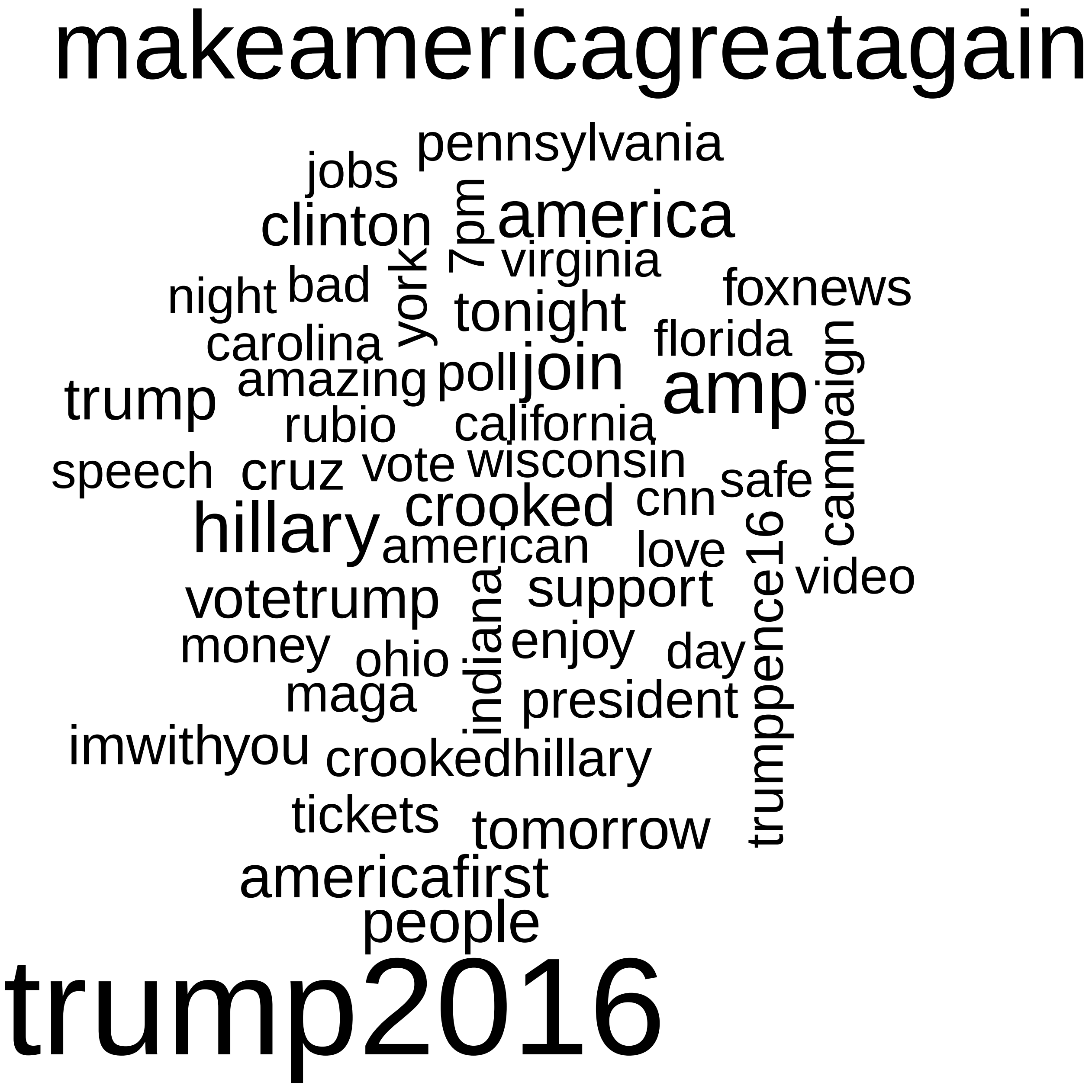
> ANSWER: It appears that the word clouds are significantly different. Some common words that appear in both word clouds are "hillary", "foxnews", and "cnn". However, the most common words are different. For android, the most common words are "realdonaldtrump", "hillary", and "crooked". For iPhone, the most common words are "trump2016", "makeamericagreatagain", and "hillary". It appears more of the words from Android are about his political opponents, while his iPhone tweets relate more to his campaign events. Thus it seems the tone for these two text groups are somewhat different.

```r
#Modifying stopwords
word <- c("https", "t.co")
lexicon <- c("Link", "Link")
df <- data.frame(word, lexicon)
stop_words <- rbind(stop_words, df)

#Android wordcloud
android <- tweets2 %>%
  #Filter for android
  filter(source == "Android") %>%
  select(text) %>%
  unnest_tokens(output = word, input = text) %>%
  #removing stopwords
  anti_join(stop_words, by="word") %>%
  count(word, sort = TRUE)
#generate wordcloud
wordcloud(words = android$word
        , freq = android$n, max.words=50, , scale=c(5,1.6))
```



```r
#iPhone WordCloud
iPhone <- tweets2 %>%
  #filter for iPhone source
  filter(source == "iPhone") %>%
  select(text) %>%
  unnest_tokens(output = word, input = text) %>%
  #remove stopwords
  anti_join(stop_words, by="word") %>%
  count(word, sort = TRUE)
#generate wordcloud
wordcloud(words = iPhone$word
        , freq = iPhone$n, max.words=50, , scale=c(5,1.6))
```


makeamericagreatagain
pennsylvania
jobs
clinton
7pm
america
virginia
york
night
bad
foxnews
tonight
carolina
florida
amazing
poll
join
amp
campaign
trump
rubio
california
speech
cruz
vote
wisconsin
safe
cnn
crooked
hillary
american
trumppence16
love
video
votetrump
indiana
support
money
ohio
enjoy
day
maga
president
imwithyou
crookedhillary
tickets
tomorrow
americafirst
people
trump2016

2b. Create a visualization that compares the top 10 *bigrams* appearing in tweets by each source (that is, facet by source). After creating a dataset with one row per bigram, you should remove any rows that contain a stop word within the bigram.

How do the top used bigrams compare between the two sources?

> ANSWER: While some phrases appear at the top of both lists, such as “crooked hillary” and “ted cruz”, the words for Android refer more directly to people while for iPhone the most used words include more slogans and what appear to be campaign event hashtags.

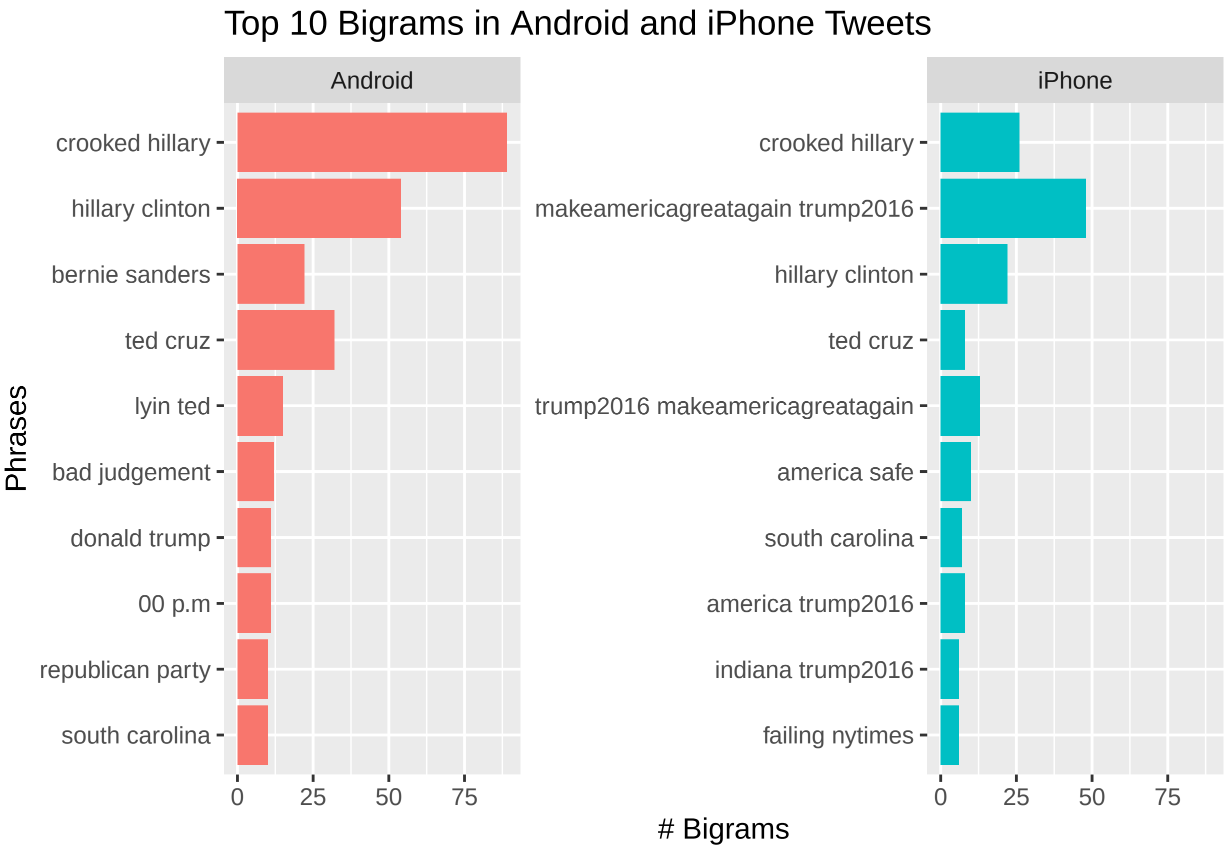
```
# Bigram code
trump_bigrams <- tweets2 %>%
  unnest_tokens(bigram, text, token = "ngrams", n = 2) %>%
  group_by(source) %>%
  count(bigram, sort = TRUE)

# Remove stop words and unite back, split in two
bigrams_filtered <- trump_bigrams %>%
  separate(bigram, c("first", "second"), sep = " ") %>%
  filter(!first %in% stop_words$word) %>%
  filter(!second %in% stop_words$word) %>%
  unite(bigram, first, second, sep = " ")

#Take top 10 for android:
android_ten <- bigrams_filtered %>%
  filter(source == "Android") %>%
  arrange(desc(n)) %>%
  top_n(10)
```

```
## Selecting by n
```

```
#Take top 10 for iPhone
iPhone1 <- bigrams_filtered %>%
  filter(source == "iPhone") %>%
  arrange(desc(n))
iPhone_ten <- iPhone1[1:10,]
#combine
trump_proc <- rbind(android_ten, iPhone_ten)
#Visualization
trump_proc %>%
  ggplot(aes(reorder(bigram, n), n, fill = source)) +
  geom_col(show.legend = FALSE) +
  #Second argument makes separate y-axis for the two graphs
  facet_wrap(~source, scales = "free_y") +
  labs(y = "# Bigrams", x = "Phrases") +
  ggtitle("Top 10 Bigrams in Android and iPhone Tweets") +
  coord_flip()
```


Top 10 Bigrams in Android and iPhone Tweets
Android
iPhone
Phrases
crooked hillary
hillary clinton
bernie sanders
ted cruz
lyin ted
bad judgement
donald trump
00 p.m
republican party
south carolina
crooked hillary
makeamericagreatagain trump2016
hillary clinton
ted cruz
trump2016 makeamericagreatagain
america safe
south carolina
america trump2016
indiana trump2016
failing nytimes
0
25
50
75
0
25
50
75
# Bigrams

2c. Consider the sentiment. Compute the proportion of words among the tweets within each source classified as “angry” and the proportion of words classified as “joy” based on the NRC lexicon. How does the proportion of “angry” and “joy” words compare between the two sources? What about “positive” and “negative” words?

> ANSWER: The proportion of words that are classified as angry is higher for Android than for iPhone, although the difference is not very much, around 1 percentage point. The proportion of words that are classified as “joy” is lower for Android than for iPhone with a difference of around 2 percentage points. Android is less positive and more negative compared to iPhone, about 3 percentage points more negative and 3 percentage points less positive.

```r
#get the sentiments
nrc1 <- get_sentiments("nrc")
#vector of sentiments
sentiments <- c("anger", "joy", "negative", "positive")

#Get number of sentiments for Android tweets
android_sentiments1 <- nrc1 %>%
  inner_join(android) %>%
  group_by(sentiment) %>%
  summarise(total = sum(n)) %>%
  mutate(proportion = total/sum(total),
         source = "Android")
```

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

```r
#Get number of sentiments for iPhone tweets
iPhone_sentiments1 <- nrc1 %>%
  inner_join(iPhone) %>%
  group_by(sentiment) %>%
  summarise(total = sum(n)) %>%
  mutate(proportion = total/sum(total),
         source = "iPhone")
```

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

```r
#Combine data sets, filter for wanted sentiments
all_sentiments <- rbind(android_sentiments1, iPhone_sentiments1)
all_sentiments <- all_sentiments %>%
  filter(sentiment %in% sentiments) %>%
  select(-total)

all_sentiments
```

```
## # A tibble: 8 x 3
##   sentiment proportion source
##   <chr>          <dbl> <chr>
## 1 anger         0.0911 Android
## 2 joy           0.0670 Android
## 3 negative      0.162  Android
## 4 positive      0.184  Android
## 5 anger         0.0871 iPhone
## 6 joy           0.0835 iPhone
## 7 negative      0.134  iPhone
## 8 positive      0.215  iPhone
```

2d. Lastly, based on your responses above, do you think there is evidence to support Robinson's claim that Trump only writes the (angrier) Android half of the tweets from realDonaldTrump? In 2-4 sentences, please explain.

> ANSWER: While there is a difference between Android and iPhone tweets from Donald Trump, the difference in tone is very small. In addition, while there are differences in the most used words between Android and iPhone tweets, the two groups still share many words. As such, I don't think the assertion that Trump only writes the Android half of the tweets from realDonaldTrump is well-supported. I think it could be that his campaign only tweets from an iPhone as his slogans of "make america great again" and "trump 2016" only appear in the iPhone word cloud. Thus, I do not believe in Robinson's claim that Trump ONLY writes the Android half, but I do support the idea that a higher proportion of the Android tweets are from Trump compared to iPhone.