STAT 231: Problem Set 6B

Jett Knight

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:

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.

```r
load(url("http://varianceexplained.org/files/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`

```r
trump_tweets_df <- trump_tweets_df %>%
  filter(screenName == "realDonaldTrump")
tweets <- trump_tweets_df %>%
  select(text, created, statusSource)
```

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 are 5 different sources. Instagram is used once, Twitter is used 120 times from a laptop, once from an iPad, 762 times from an android phone and 628 times from an iPhone.

```
tweets %>%
  group_by(statusSource) %>%
  summarize(n = n())
```

```
## # A tibble: 5 x 2
##   statusSource                                                                n
## * <chr>                                                                   <int>
## 1 "<a href=\"http://instagram.com\" rel=\"nofollow\">Instagram</a>"           1
## 2 "<a href=\"http://twitter.com\" rel=\"nofollow\">Twitter Web Client</a>"   120
## 3 "<a href=\"http://twitter.com/#!/download/ipad\" rel=\"nofollow\">Twitt~     1
## 4 "<a href=\"http://twitter.com/download/android\" rel=\"nofollow\">Twitt~   762
## 5 "<a href=\"http://twitter.com/download/iphone\" rel=\"nofollow\">Twitte~   628
```

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 command creates a fourth column that identifies whether a tweet in the statusSource column is either from an android or an iPhone.

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

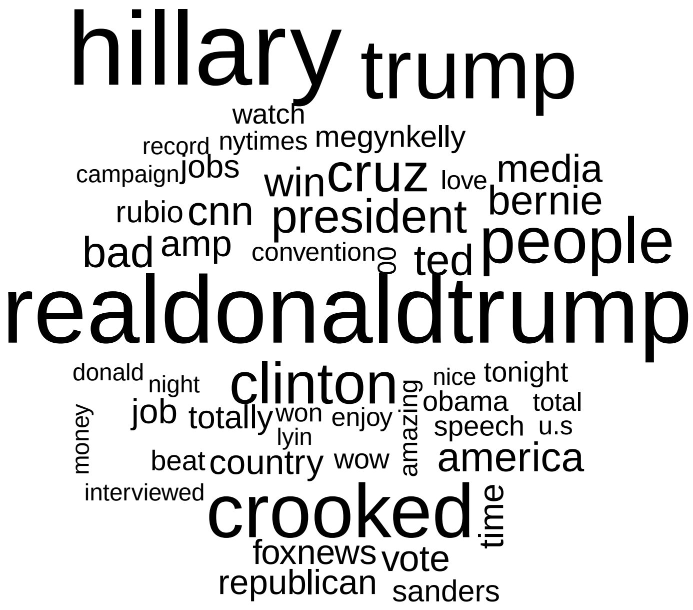
How does the language of the tweets differ by source?

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?)

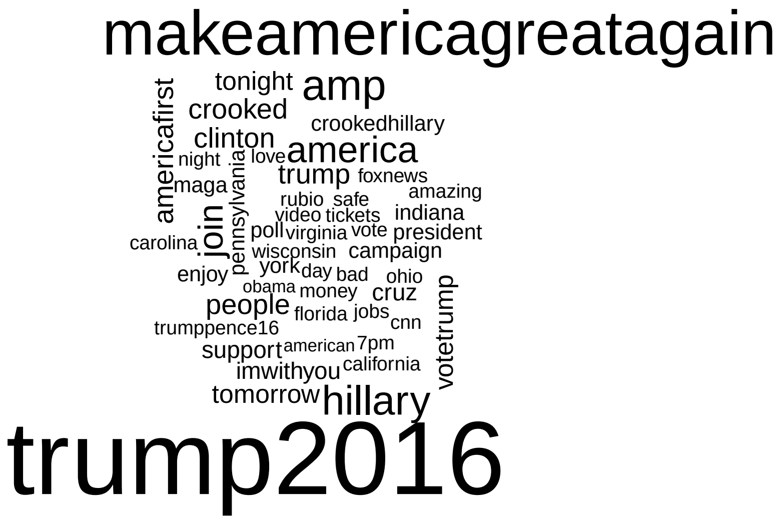
Don't forget to remove stop words before creating the word cloud. Also remove the terms "https" and "t.co".

> ANSWER: The word clouds generally have similar topics of discussion. He discusses Hilary more in the android wordcloud than the iphone wordcloud however.

```
nostopwords <- tweets2 %>%
  unnest_tokens(output = word, input = text)

nostopwords <- nostopwords %>%
  anti_join(stop_words, by = "word") %>%
  filter(word != "https") %>%
  filter(word != "t.co")

nostopwords_split <- split(nostopwords, nostopwords$source)

nostopwords_android <- nostopwords_split$Android
nostopwords_iphone <- nostopwords_split$iPhone

nostopwords_android <- nostopwords_android %>%
  count(word, sort = TRUE)

nostopwords_iphone <- nostopwords_iphone %>%
  count(word, sort = TRUE)

nostopwords_android %>%
  with(wordcloud(words = word, freq = n, max.words=50))
```



```
nostopwords_iphone %>%
  with(wordcloud(words = word, freq = n, max.words=50))
```


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

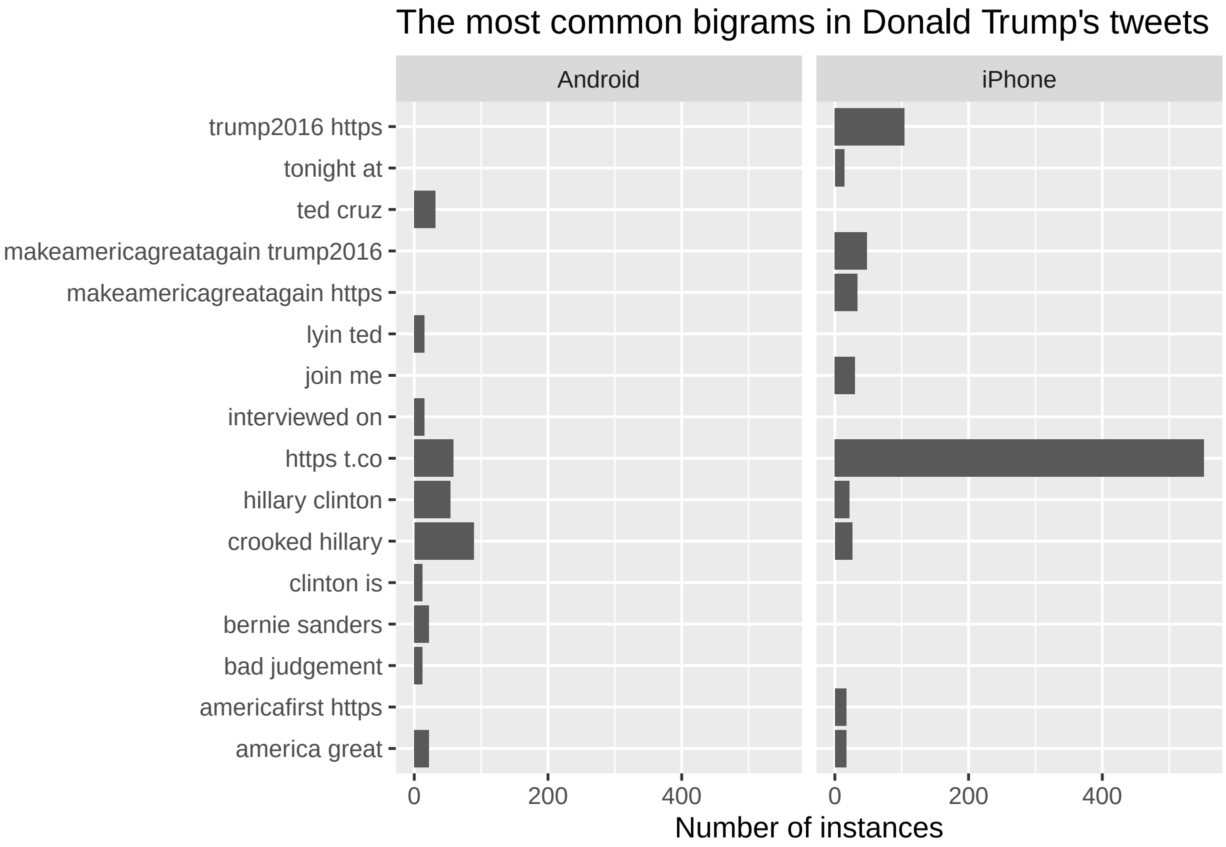
2b. Create a visualization that compares the top 10 *bigrams* appearing in tweets by each source (that is, facet by source). After creating a dataset with one row per bigram, you should remove any rows that contain a stop word within the bigram.

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

> ANSWER: The bigrams are generally similar. Again, he refers to Hilary more from the android and general election stuff from the iPhone.

```
tweets_bigrams <- tweets2 %>%
  unnest_tokens(output = bigram, input = text, token = "ngrams", n = 2) %>%
  separate(bigram, into = c("one", "two"), sep = " ", remove = FALSE) %>%
  anti_join(stop_words, by = c("one" = "word")) %>%
  anti_join(stop_words, by = c("one" = "word")) %>%
  group_by(source) %>%
  count(bigram, sort = TRUE)
```

```
tweets_bigrams %>%
  slice(1:10) %>%
  ggplot(aes(x = bigram, y = n)) +
  geom_col() +
  facet_wrap(~source) +
  xlab(NULL) +
  coord_flip() +
  labs(y = "Number of instances"
       , title="The most common bigrams in Donald Trump's tweets") +
  guides(color = "none", fill = "none")
```


The most common bigrams in Donald Trump's tweets
Android
iPhone
trump2016 https
tonight at
ted cruz
makeamericagreatagain trump2016
makeamericagreatagain https
lyin ted
join me
interviewed on
https t.co
hillary clinton
crooked hillary
clinton is
bernie sanders
bad judgement
americafirst https
america great
0
200
400
0
200
400
Number of instances

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: There are surprisingly an equal number of “anger” and “joy” words, but more negative words than positive words.

```
nrc_lexicon <- get_sentiments("nrc")
nrcplot <- nostopwords %>%
  inner_join(nrc_lexicon, by = "word") %>%
  filter(sentiment %in% c("anger", "joy"))
count(nrcplot, sentiment = "anger")
```

```
## # A tibble: 1 x 2
##   sentiment     n
## * <chr>     <int>
## 1 anger       963
```

```
count(nrcplot, sentiment = "joy")
```

```
## # A tibble: 1 x 2
##   sentiment     n
## * <chr>     <int>
## 1 joy         963
```

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: I don't find that convincing. There is enough similarity between the android and iPhone tweets, in my opinion.