STAT 231: Problem Set 6B

Sean Wei

due by 5 PM on Friday, October 9

This homework assignment is designed to help you futher 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.

```
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`

```
# Confirm that realDonaldTrump is the only screen name in the dataset
unique(trump_tweets_df$screenName)
```

```
## [1] "realDonaldTrump"
```

```
# New dataset with only text, created, and statusSource
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 are 5 unique tweet sources: Instagram, Twitter for Android, Twitter for iPad, Twitter for iPhone, and Twitter Web Client. There were 762 tweets from Twitter for Android, 628 tweets from Twitter for iPhone, 120 tweets from Twitter Web Client, and 1 tweet from both Instagram and Twitter for iPad.

```
# Isolate the text in the anchor tags
tweet_sources <- tweets %>%
  mutate(
    statusSource = gsub("</a>", "", statusSource),
    statusSource = gsub(".*>", "", statusSource)
  )
glimpse(tweet_sources)
```

```
## 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> "Twitter for Android", "Twitter for iPhone", "Twitter ...
```

```
# Put each observation in statusSource into one big string separated by commas
source_count <- paste(tweet_sources$statusSource, collapse = ",")

# Turn large string into a character array
source_count <- strsplit(source_count, ",")

# Create frequency table for the character array
table(source_count)
```

```
## source_count
##           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. (Note that "regex" stands for "regular expression".)

> ANSWER: The `extract` function is given a regular expression to find groups and then makes a new column based on them. Below, the `extract` function looks for instances that begin with "Twitter for" and then makes a new column with values of what comes after it and before the <. If the statusSource doesn't match this regex, the value of the new column is N/A.

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

```
## Rows: 1,390
## Columns: 4
## $ 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=\...
## $ source       <chr> "Android", "iPhone", "iPhone", "Android", "Android", "...
```

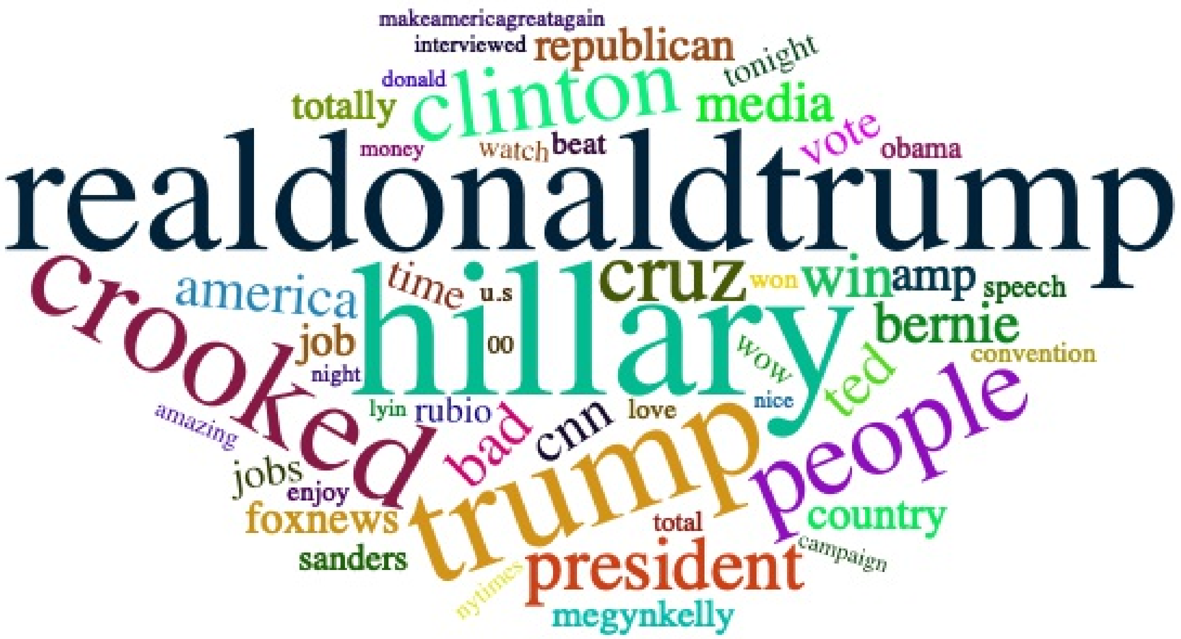
How does the language of the tweets differ by source?

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?)

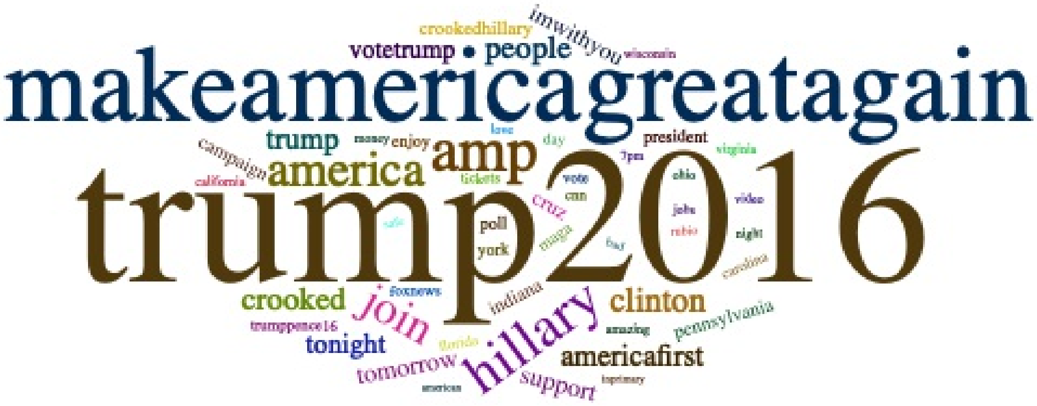
Don't forget to remove stop words before creating the word cloud. Also remove the terms "https" and "t.co".

> ANSWER: The wordcloud generated by the Android tweets seem to be more focused on his opponents. The most common word being "Hillary," and other accompanying words such as "crooked" and "bad" imply that these tweets were created to negatively portray the other candidates. On the other hand, the iPhone tweets seemed to be more geared to progressing his campaign. While "Hillary" is also very common, common words in the iPhone tweets are "trump2016," "makeamericagreatagain," and "votetrump." Although there are many common words shared in both the Android and iPhone tweets, it appears that the motivation behind the tweets differed.

```r
data("stop_words")

# Create wordcloud for Android
android <- tweets2 %>%
  filter(source == "Android") %>%
  rename(tweet = text) # can't replicate text column

android <- tibble(text = android$tweet) %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  filter(!word %in% c("https", "t.co")) %>%
  count(word, sort = TRUE) %>%
  head(50)

wordcloud2::wordcloud2(data=android, size = 0.5, color='random-dark')
```



```r
# Create wordcloud for iPhone
iphone <- tweets2 %>%
  filter(source == "iPhone") %>%
  rename(tweet = text) # can't replicate text column

iphone <- tibble(text = iphone$tweet) %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  filter(!word %in% c("https", "t.co")) %>%
  count(word, sort = TRUE) %>%
  head(50)

wordcloud2::wordcloud2(data=iphone, size = 0.5, color='random-dark')
```


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

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:

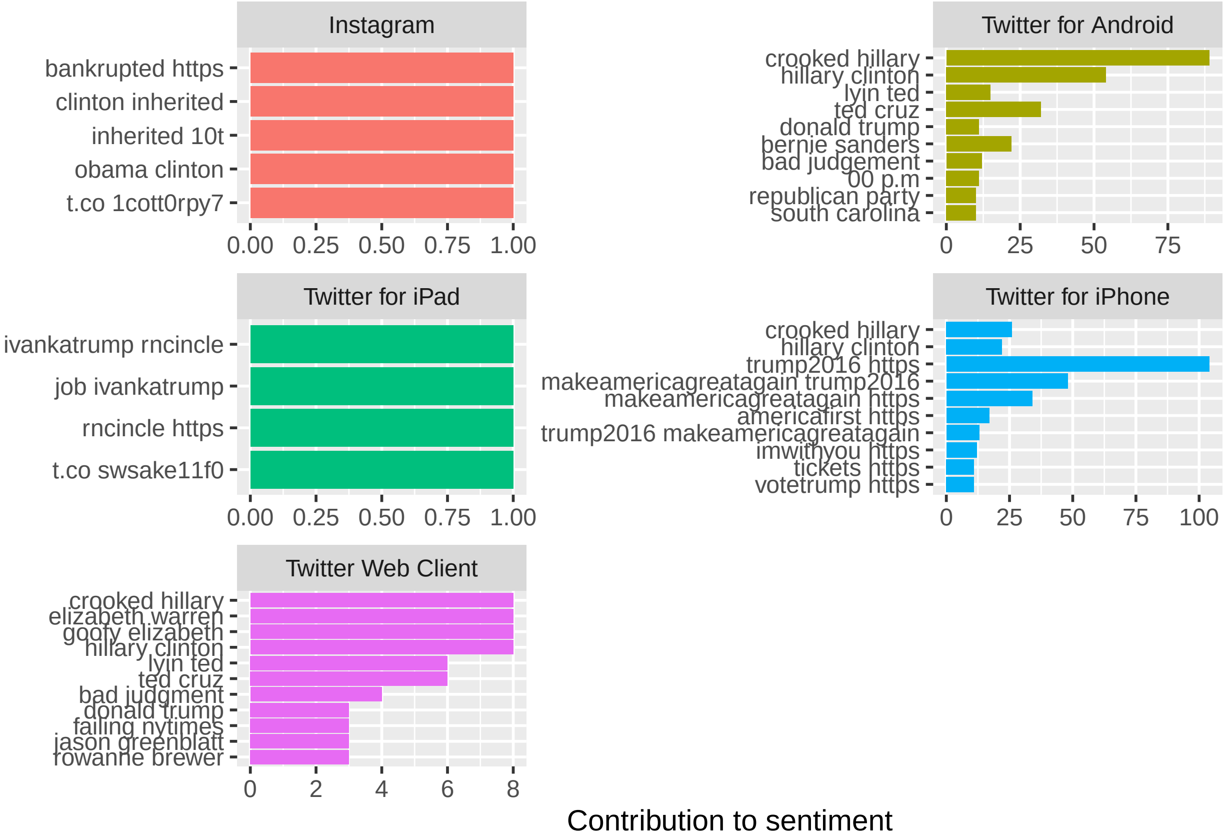
```
library(tidytext)

# Using tweet_sources data from 1b
trump_bigrams <- tweet_sources %>%
  unnest_tokens(bigram, text, token = "ngrams", n = 2) %>%
  group_by(statusSource) %>%
  count(bigram, sort = TRUE)

# Remove stop words
bigrams_separated <- trump_bigrams %>%
  separate(bigram, c("word1", "word2"), sep = " ")

bigrams_filtered <- bigrams_separated %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word)

trump_bigrams_final <- bigrams_filtered %>%
  unite(bigram, word1, word2, sep = " ") %>%
  filter(bigram != "https t.co") %>%
  arrange(desc(statusSource, n)) %>%
  top_n(10)
```

```
## Selecting by n
```

```
head(trump_bigrams_final)
```

```
## # A tibble: 6 x 3
## # Groups:   statusSource [1]
##   statusSource       bigram                n
##   <chr>              <chr>             <int>
## 1 Twitter Web Client crooked hillary      8
## 2 Twitter Web Client elizabeth warren     8
## 3 Twitter Web Client goofy elizabeth      8
## 4 Twitter Web Client hillary clinton      8
## 5 Twitter Web Client lyin ted             6
## 6 Twitter Web Client ted cruz             6
```

```
# Visualization of top 10 bigrams across the different sources
trump_bigrams_final %>%
  ungroup %>%
  mutate(word = factor(bigram, levels = rev(unique(bigram)))) %>%
  ggplot(aes(word, n, fill = statusSource)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~statusSource, ncol = 2, scales = "free") +
  labs(y = "Contribution to sentiment", x = NULL) +
  coord_flip()
```


Instagram
bankrupted https
clinton inherited
inherited 10t
obama clinton
t.co 1cott0rpy7
0.00 0.25 0.50 0.75 1.00
Twitter for Android
crooked hillary
hillary clinton
lyin ted
ted cruz
donald trump
bernie sanders
bad judgement
00 p.m
republican party
south carolina
0 25 50 75
Twitter for iPad
ivankatrump rncincle
job ivankatrump
rncincle https
t.co swsake11f0
0.00 0.25 0.50 0.75 1.00
Twitter for iPhone
crooked hillary
hillary clinton
trump2016 https
makeamericagreatagain trump2016
makeamericagreatagain https
americafirst https
trump2016 makeamericagreatagain
imwithyou https
tickets https
votetrump https
0 25 50 75 100
Twitter Web Client
crooked hillary
elizabeth warren
goofy elizabeth
hillary clinton
lyin ted
ted cruz
bad judgment
donald trump
failing nytimes
jason greenblatt
rowanne brewer
0 2 4 6 8
Contribution to sentiment

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:

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: