311-2 Homework 2

Tina Zhang

5/4/2019

```r
knitr::opts_chunk$set(echo = TRUE)
packages <- c("xml2","rvest","tm","data.table","dplyr","tidytext","SnowballC","ggplot2","scales")
load.packages <- function(x) {
  if (!require(x, character.only = TRUE)) {
    install.packages(x, dependencies = TRUE)
    library(x, character.only = TRUE)
  }
}
lapply(packages, load.packages)
setwd("C:/Users/PIG/Desktop/311-2/R")
```

1.1. Scraping

```r
#a
wikimain <- "https://en.wikipedia.org"
webpage <- read_html("https://en.wikipedia.org/wiki/Category:Member_states_of_the_Association_of_Southea

#b
country_html <- html_nodes(webpage,".mw-category-group a")

#c
country_text <- html_text(country_html)
country_link <- paste0(wikimain , html_attr(country_html,"href"))
df <- data.frame(country_text, country_link)

#d
df$paragraphs <- c()
for (i in 1:11)
{
  df$paragraphs[i] <- read_html(country_link[i]) %>%
    html_nodes("p") %>% html_text() %>% gsub("\\n","", . ) %>%
    paste(collapse = " ")
}
```

2.1 Preprocess

```r
#a
trump <- fread("trumptweets.csv", select = 2:3, quote = "") %>%
  mutate(created_at = as.Date(substr(created_at,1,10),
                              format = "%m-%d-%Y"))
```

```
#b and c
#Note: unnest_tokens automatically removes punctuations & capitalization:
tokentext <- unnest_tokens(trump, word, text) %>%
  anti_join(stop_words) %>% mutate(word = stemDocument(word))
```

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

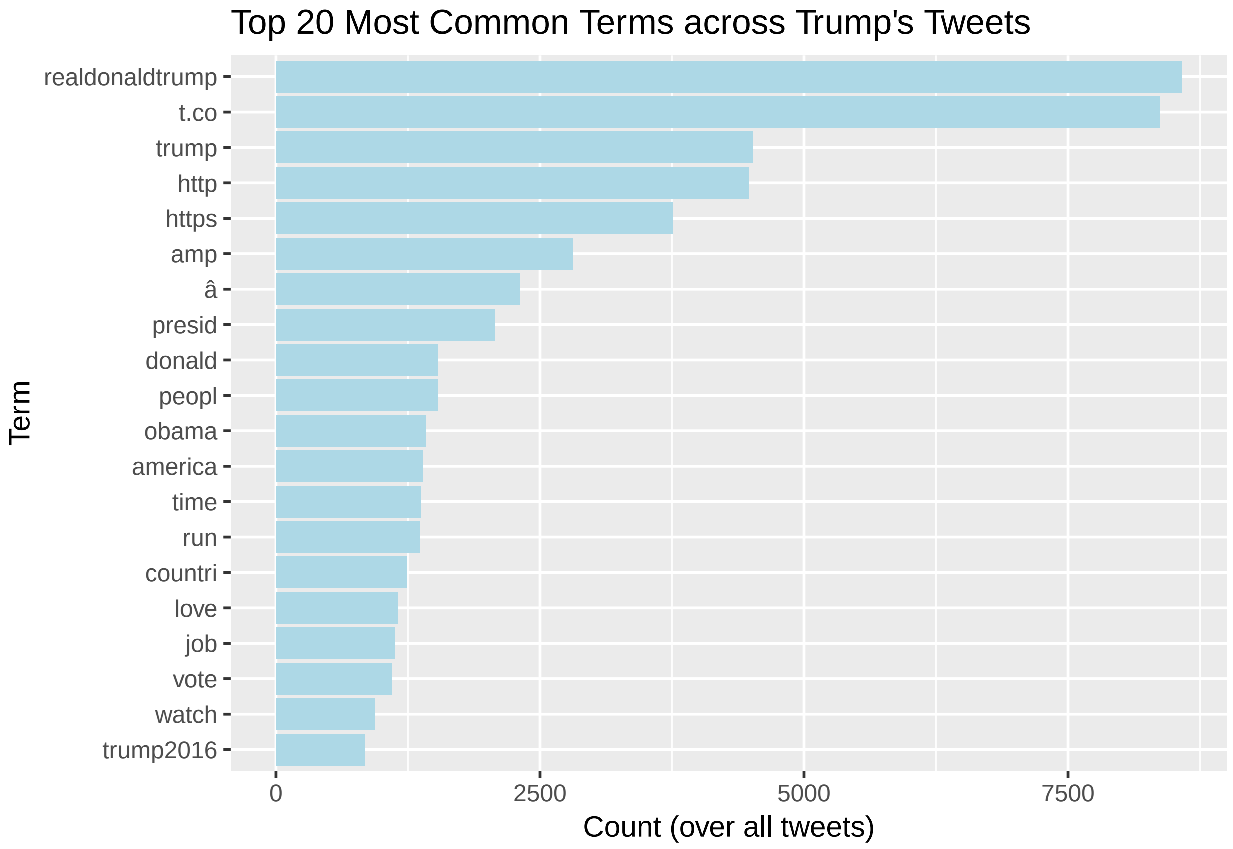
```
dtm <- tokentext %>% count(word, sort = TRUE) %>% mutate(doc = 1) %>%
  cast_dtm(doc, word, n) %>% removeSparseTerms(0.99)

#d
tidydtm <- tidy(dtm)

#e
tf_idf <- bind_tf_idf(tidydtm, term, document, count)
```

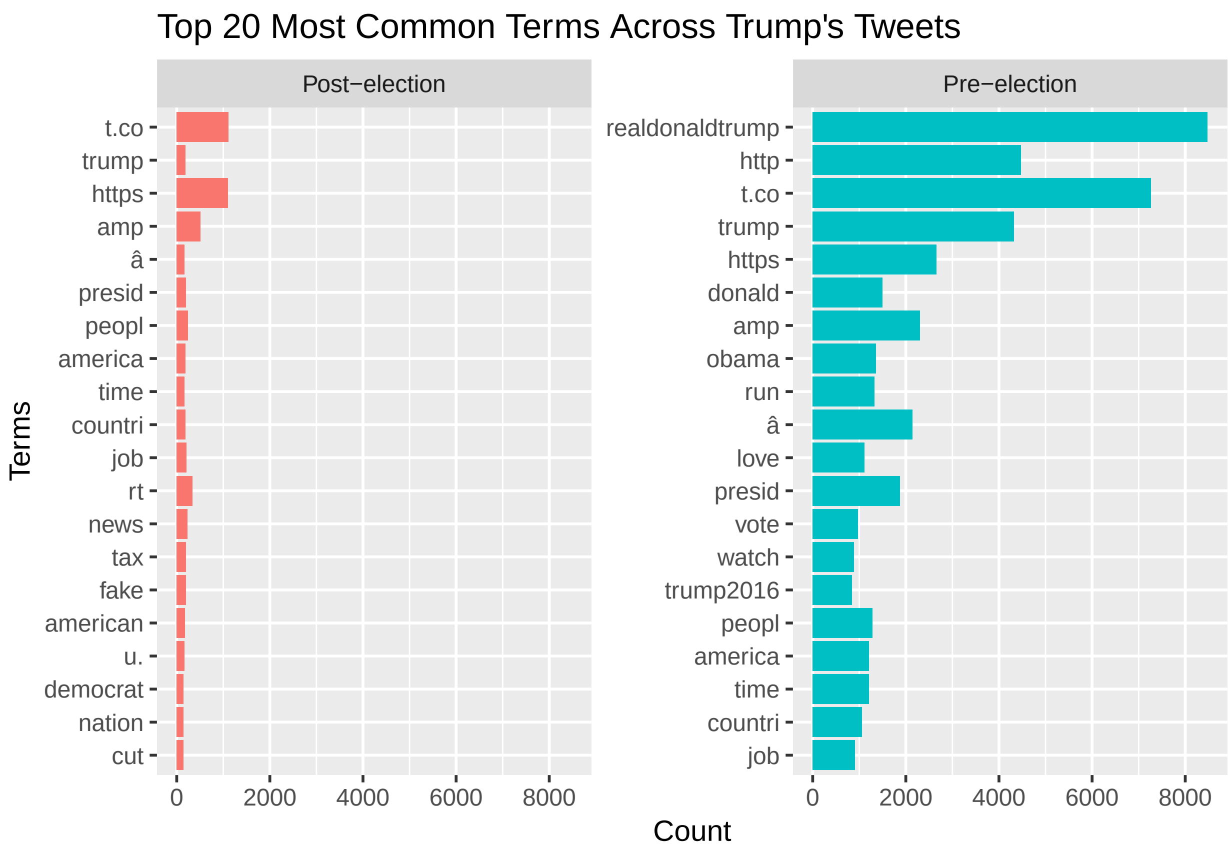
2.2 Word Frequency/Dictionary Methods

```
#a
tidydtm[1:20, ] %>% mutate(term = reorder(term, count)) %>%
  ggplot(aes(term, count)) + geom_col(fill = "lightblue") +
  labs(x = "Term", y =  "Count (over all tweets)",
       title = "Top 20 Most Common Terms across Trump's Tweets") +
  coord_flip()
```



```
#b
tokentext$doc <- ifelse(tokentext$created_at > as.Date("2016-11-08"),
                        "Post-election", "Pre-election")
dtm2 <- tokentext %>% count(word, doc) %>%
  cast_dtm(doc, word, n) %>% removeSparseTerms(0.99)
tidydtm2 <- tidy(dtm2)
tidydtm2 %>% group_by(document) %>%
  top_n(20) %>%
  ungroup() %>%
  mutate(term = reorder(term, count)) %>%
  ggplot(aes(term, count, fill = document)) +
  geom_col() + facet_wrap(~document, scales = "free_y")+
  labs(y = "Count", x = "Terms",
       title = "Top 20 Most Common Terms Across Trump's Tweets") +
  coord_flip() + theme(legend.position = "none")
```

```
## Selecting by count
```



A main difference is that Trump used the term "realdonaldtrump" a lot pre-election, but he no longer used the term frequently post-election. Some other terms he used frequently pre-election but not post-election include "run", "obama", "love", "vote", "watch", and "trump2016" which probably related to his presidency campaign. Some terms he used frequently post-election but not pre-election include "new", "tax", and "cut", which seem related to the policies he's currently working on.

Note: in part c, I set the token as "tweets" because otherwise the tokenization sometimes separates #word into # and word. I use gsub to remove all characters aside from #, space, letters, and numbers. This is technically not necessary because when token is set to "tweets", unnest_tokens automatically removes all punctuation beside # and @, but I wrote it just in case you wanted to see it.

```
#c
tokentext2 <- trump %>% mutate(text = gsub("[^[:alnum:][:space:]#]", "",text)) %>%
  unnest_tokens(word, token = "tweets",text, strip_url = TRUE) %>%
  anti_join(stop_words) %>% mutate(word = stemDocument(word))
```

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

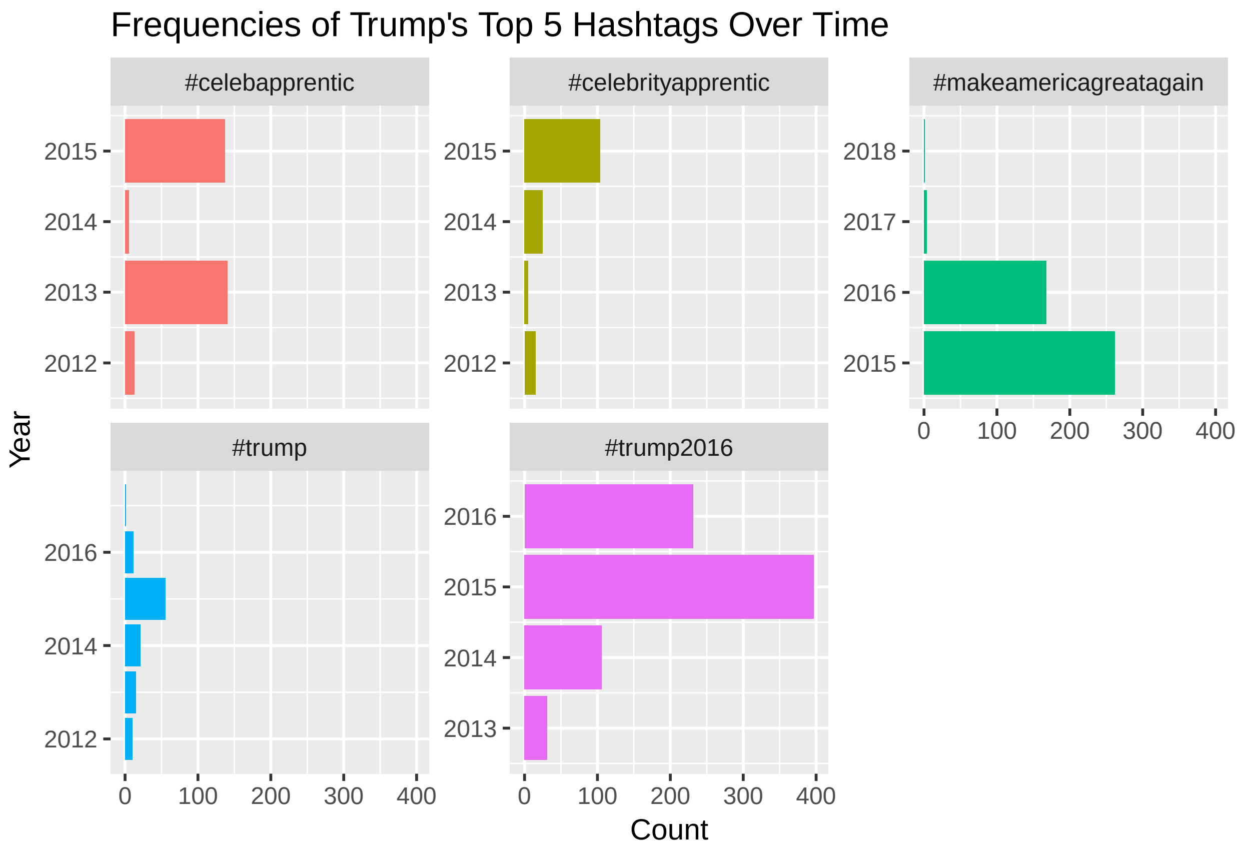
```
dtm2 <- tokentext2 %>% count(word) %>% mutate(doc = 1) %>%
  cast_dtm(doc, word, n) %>% removeSparseTerms(0.99)
tidydtm2 <- tidy(dtm2)

#d
top5 <- tidydtm2[startsWith(tidydtm2$term, "#"),2:3] %>%
  top_n(5, count) %>% mutate(term = reorder(term, count))
top5hts <- as.character(top5$term)
print(top5hts)
```

```
## [1] "#celebapprentic"         "#celebrityapprentic"
## [3] "#makeamericagreatagain" "#trump"
## [5] "#trump2016"
```

ˆ These are Trump's top 5 most-used hashtags over all time.

```
#e
top5overtime <- tokentext2[tokentext2$word %in% top5hts, ]
top5overtime$year <- as.Date(cut(top5overtime$created_at,breaks = "year"))
plot <- top5overtime %>% count(word, year)
ggplot(plot, aes(year, n, fill = word)) + geom_col() +
  facet_wrap(~word, scales = "free_y")+
labs(y = "Count", x = "Year", title = "Frequencies of Trump's Top 5 Hashtags Over Time") +
  coord_flip() + theme(legend.position = "none")
```

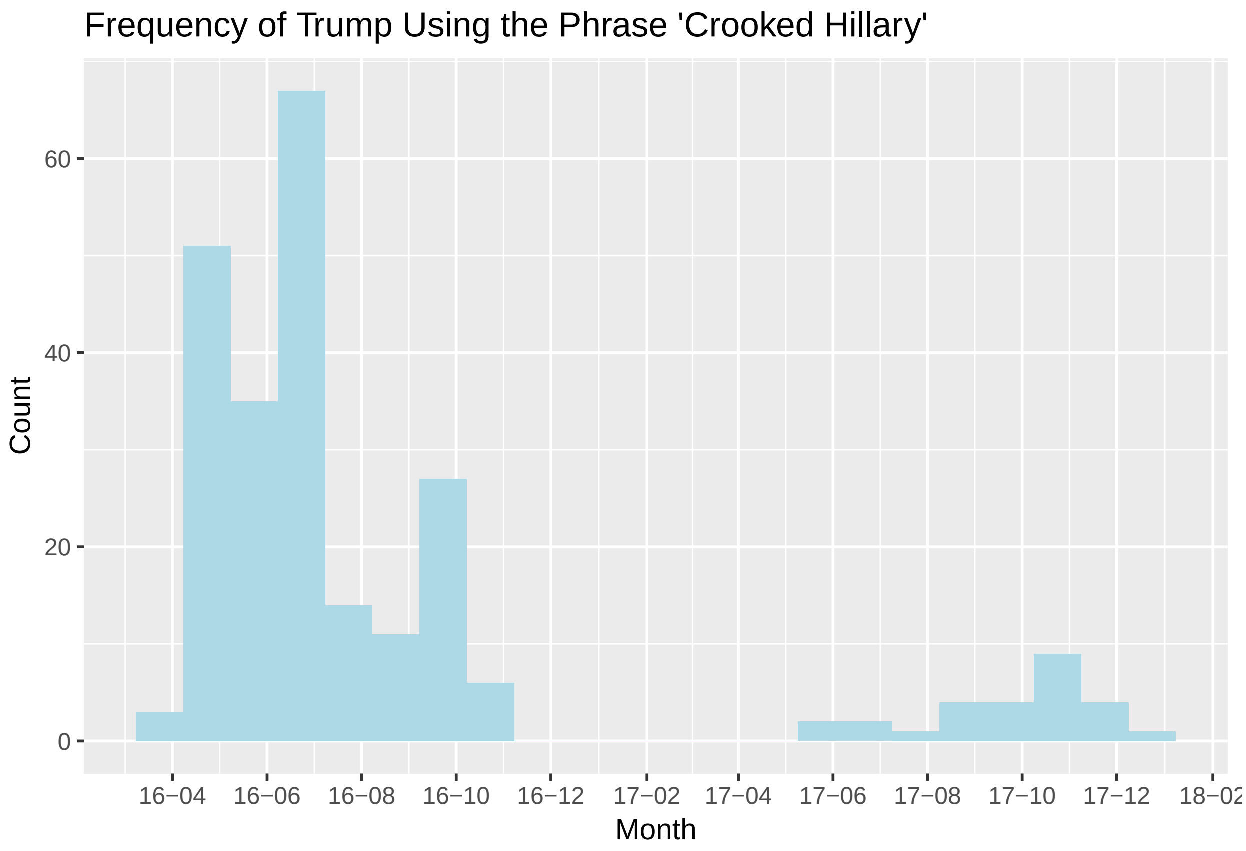


Note: for part f, I used grepl to search for all tokens that contain "crooked hillar", since the stemming sometimes removes the "y" in hillary. Also, I'm using grepl rather than searching for exact matches because I thought that I should also catch mentions of "crooked hillaryclinton".

```
#f
bitoken <- unnest_tokens(trump, word, text, token="ngrams", n=2) %>%
  anti_join(stop_words) %>% mutate(word = wordStem(word))
```

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

```r
chovertime <- bitoken[grepl("crooked hillar",bitoken$word), ]
chovertime$month <- as.Date(cut(chovertime$created_at,breaks = "month"))
ggplot(chovertime, aes(x = month)) +
  geom_histogram(bins = 22, fill="lightblue") +
  scale_x_date(labels = date_format("%y-%m"), breaks = "2 month") +
  labs(y = "Count", x = "Month", title = "Frequency of Trump Using the Phrase 'Crooked Hillary'")
```



Frequency of Trump Using the Phrase 'Crooked Hillary'