#Class: Week 11
#Course: Big Data and Social Analysis
#Semester: Spring 2021
#Lesson: Text Mining
#Instructor: Chung-pei Pien
#Organization: ICI, NCCU

Student Information --------

#Chinese Name: 辛鐘成
#English First Name: Jongsung Shin
#UID: 110ZU1038
#E-mail: sjongsung97@gmail.com

Questions --------

#In the midterm, we used the data file fake_tweets_election.xlsx to analyze
#misinformation tweets collected during and after the first 2020 US presidential debate.

#Please read fake_tweets_election.xlsx and use the following code to delete
#the observation on the date before September 27, 2020.

```
library(dplyr) # bring all the library we might use for data cleaning
library(ggplot2)
library(readxl)
library(tidyr)
library(tidyverse)
library(zoo)
library(tidytext) # an efficient tool for text mining in R, merging with dplyr package
library(tm)
library(wordcloud2)
library(widyr)


setwd("C:/Users/sung/Desktop/Big data with R/BD EXCEL FILE")

fake_tweets <- read_xlsx("fake_tweets_election.xlsx")
fake_tweets <- fake_tweets %>%
  filter(date >= "2020-09-27")
```

#The column type records the text is a tweet or a retweet.
#In the midterm, we didn't delete retweet because we want to analyze misinformation's spread and size.

#Question 1: (2 points)

```r
#In this week's howework, we attempt to apply text mining techniques to do word frequency, word association and
#sentiment analysis.

#Do you think you need to delete retweets? Please tell me your answer and provide me your reasons.

# Yes, I think we need to delete retweets since it shows the tweets already used.
# It will increase the word frequency and association and sentiment analysis,
# which disturbs us to get an accurate analysis.

#If your anawer is to delete retweets, please use filter() to delete them.
fake_tweets <- fake_tweets %>%
  filter(!type %in% ('retweet'))

unique(fake_tweets$type) # check if 'retweet' has been removed or not

#Question 2: (10 points)

#The column text records the content of tweets.
#Please remove words and symbols that we do not need for word frequency, and word association, and sentiment
analysis.
#Remember, the cleaning process may do many times when you find the results of word frequency, and word association,
and sentiment analysis involve many terms needed to eliminate.

# I want to remove words and symbols that do not have any impact on the meaning to it
text <- fake_tweets$text
text[1:10] # I will check some text each time to see the change

# Set the text to lowercase
text <- tolower(text)
text[1:10]

# gsub(pattern, replacement, string) => replace all matches
# Remove urls, emojis, etc.
text <- gsub("https?://.+", "", text)
text[1:10]

# \d is a digit (a character in the range 0-9), and + means 1 or more times. So, \d+ is 1 or more digits.
# ^[\w*]$ will match a string consisting of a single character, where that character is alphanumeric (letters, numbers) an
underscore ( _ ) or an asterisk ( * ).
# Details: The " \w " means "any word character" which usually means alphanumeric (letters, numbers, regardless of
case) plus underscore (_)
#  \d matches any decimal digit. The signification of a "decimal digit" depends on the options of the regex: Without
RegexOptions.
```

```
text <- gsub("\\d+\\w*\\d*", "", text)
text[1:10]

text <- gsub("[^\x01-\x7F]", "", text) # this is for emoji
text[1:10]

# Remove references to other twitter users and hash tags
text <- gsub("@\\w+", "", text)
text[1:10]

# Remove hash tages
text <- gsub("#\\w+", "", text)

# Remove number and punctuation
text <- gsub("[[:digit:]]", "", text)
text <- gsub("[[:punct:]]", " ", text)
text[1:10]

# Remove spaces and newlines
text <- gsub("amp", "", text)
text <- gsub("\n", " ", text)
text[1:10]

# There are spaces where the digits were, we need to remove it
text <- gsub("^\\s+", "", text)
text <- gsub("\\s+$", "", text)
text <- gsub("[ |\t]+", " ", text)
text[1:10]

# Remove single alphabet
text <- gsub("\\W[a-zA-Z]\\W", "", text)
text[1:10]

fake_tweets$new_text <- text

colnames(fake_tweets)

fake_tweets_temp <- fake_tweets %>%
  select(content_id, new_text) # show me these columns only
```

```
#Question 3: (3 points)

#Please tokenize the tweets
fake_tokens <- fake_tweets_temp %>%
  unnest_tokens(word, new_text) %>% # Tokenize fake_tweet_temp with stop word
  anti_join(stop_words) %>%
  filter(!word %in% stopwords('ENGLISH'))



#Question 4: (5 points)

#Please count tokenized terms’ frequency

fake_frq <- fake_tokens %>% # Count word frequency
  count(word, sort = TRUE)

#Question 5: (5 points)

#Please plot a word cloud with 200 top terms.
wordcloud2(fake_frq[1:201, ], shape = 'circle') # Use 201 to get 200 top terms

#Question 6: (5 points)

#Please use word association methods to tell me the top 5 high association terms with Biden.

# Create tokenized tables so that we can easily calculate the word association.
fake_cors <- fake_tokens %>%
  group_by(word) %>%
  pairwise_cor(word, content_id, sort = TRUE)

# Select a word I want to analyze
fake_biden <- fake_cors %>%
  filter(item1 == 'biden') # biden is a criteria
# top 5 high association terms with Biden: 1.debate / 2.joe / 3. wallace / 4. trump / 5. chris
```