**Assignment 2: Text Analytics for National Institutes of Health (100 points)**

**Student Name:**

**Purpose:** To perform text analytics including creating word clouds, perform sentiment analysis and topic modeling

**Description:** The data for this assignment has been collected from psychcentral.com. This website offers an online forum for posting questions and answers related to mental health. Please visit <https://forums.psychcentral.com> for more information. Our objective is to perform text analytics to discover useful information related to mental health.

**Instructions:** Please follow these steps:

1. In Canvas, navigate to Assignments and then Assignment2
2. Download and save the data set psychcentral\_data.csv
3. Read the file: data <- fread("psychcentral\_data.csv", sep=",", header=T, strip.white = T, na.strings = c("NA","NaN","","?"))
   1. **(1 point)** What are the column names in the data?
   2. **(1 point)** How many rows does this data have?
4. Use libraries “dplyr” and “tidytext” to tokenize column q\_content. Then remove the stop-words. The count the number of tokens.
   1. **(2 points)** What are the top five tokens returned?
   2. **(2 points)** Use library “ggplot2” to create a visualization that shows the frequency of the tokens that appeared for at least 2000 times. (Hint: Change n in argument filter to 2000). Paste the visualization below:
   3. **(2 points)** Based on the results in 4.2., would you suggest stemming on this text? Why? Bring one example from the visualization above that shows stemming should be done on this text?
   4. Install “SnowballC” package using install.packages("SnowballC", repos = "https://cran.r-project.org"). Use library “SnowballC” to stem q\_content using the code below:

library(SnowballC)

tidy\_text <- data %>%

unnest\_tokens(word, q\_content) %>%

mutate(word = wordStem(word))

* + 1. **(2 points)** Then remove the stop-words. Now what are the top five tokens after stemming?
    2. **(2 points)** Use library “ggplot2” to create a visualization that shows the frequency of the tokens that appeared for at least 4000 times. (Hint: Change n in argument filter to 4000). Paste the visualization below:
    3. **(3 points)** Use library “wordcloud” to create a word cloud with the 200 most used tokens. Paste the visualization below:
    4. **(5 points)** Create a color-coded word cloud based on sentiment. Use the most frequent 100 tokens for positive and negative words. Paste the word cloud in the space below:

1. Repeat all the steps in question 4 but this time for column answers.
   1. **(2 points)** What are the top five tokens returned?
   2. **(2 points)** Use library “ggplot2” to create a visualization that shows the frequency of the tokens that appeared for at least 4000 times. (Hint: Change n in argument filter to 4000). Paste the visualization below:
   3. Install “SnowballC” package using install.packages("SnowballC", repos = "https://cran.r-project.org"). Use library “SnowballC” to stem answers using the code below:

library(SnowballC)

tidy\_text <- data %>%

unnest\_tokens(word, answers) %>%

mutate(word = wordStem(word))

* + 1. **(2 points)** Then remove the stop-words. Now what are the top five tokens after stemming?
    2. **(2 points)** Use library “ggplot2” to create a visualization that shows the frequency of the tokens that appeared for at least 6000 times. (Hint: Change n in argument filter to 6000). Paste the visualization below:
    3. **(6 points)** Use library “wordcloud” to create a word cloud with the 200 most used tokens. Paste the visualization below:
    4. **(6 points)** Create a color-coded word cloud based on sentiment. Use the most frequent 100 tokens for positive and negative words. Paste the word cloud in the space below:

1. Use the following code to perform topic-modeling on q\_content:

library(RTextTools)

library(tm)

library(wordcloud)

library(topicmodels)

library(slam)

data <- data[1:1000,] # We perform LDA on the rows 1 through 1000 in the data.

corpus <- Corpus(VectorSource(data$q\_content), readerControl=list(language="en"))

dtm <- DocumentTermMatrix(corpus, control = list(stopwords = TRUE, minWordLength = 2, removeNumbers = TRUE, removePunctuation = TRUE, stemDocument = TRUE))

rowTotals <- apply(dtm , 1, sum) #Find the sum of words in each Document

dtm.new <- dtm[rowTotals> 0, ] #remove all docs without words

lda <- LDA(dtm.new, k = 5) # k is the number of topics to be found.

* 1. **(5 points)** The code above will create the beta scores for each document per topic (k = 5). Then create bar plots (similar to what we created in class) for each topic for 10 tokens (top\_n(10, beta)). Paste the visualization below.
  2. **(5 points)** Based on the visualization in 6.1., what can you say about k? Would you try a larger k or a smaller k?
  3. **(10 points)** Repeat 6.1. with the following ks:
     1. K = 2. Paste your visualization in the space below:
     2. K = 3. Paste your visualization in the space below:
     3. K = 4. Paste your visualization in the space below:
     4. K = 10. Paste your visualization in the space below:
     5. Based on the results recommend the number of topics that would be appropriate for this corpus.

1. Use the following code to perform topic-modeling on answers:

library(RTextTools)

library(tm)

library(wordcloud)

library(topicmodels)

library(slam)

data <- data[1:1000,] # We perform LDA on the rows 1 through 1000 in the data.

corpus <- Corpus(VectorSource(data$answers), readerControl=list(language="en"))

dtm <- DocumentTermMatrix(corpus, control = list(stopwords = TRUE, minWordLength = 2, removeNumbers = TRUE, removePunctuation = TRUE, stemDocument = TRUE))

rowTotals <- apply(dtm , 1, sum) #Find the sum of words in each Document

dtm.new <- dtm[rowTotals> 0, ] #remove all docs without words

lda <- LDA(dtm.new, k = 10) # k is the number of topics to be found.

* 1. **(5 points)** The code above will create the beta scores for each document per topic (k = 10). Then create bar plots (similar to what we created in class) for each topic for 10 tokens (top\_n(10, beta)). Paste the visualization below.
  2. **(5 points)** Based on the visualization in 6.1., are the tokens in all topics similar? Then what can you say about k? Would you try a larger k or a smaller k?
  3. **(10 points)** Repeat 6.1. with the following ks:
     1. K = 2. Paste your visualization in the space below:
     2. K = 8. Paste your visualization in the space below:
     3. K = 11. Paste your visualization in the space below:
     4. K = 14. Paste your visualization in the space below:
     5. Based on the results recommend the number of topics that would be appropriate for this corpus.

1. **(20 points)** Suppose that you are a researcher who works for National Institutes of Health (NIH). You are working on a project that aims to identify the most important reasons for mental disorders. Based on your analysis above, can we propose any hypothesis about the reasons for mental disorders in the society? Please explain.