ISCS 540 – Big Data Analytics, Spring 2022
Group Project – Deliverable 1: Text Mining and Sentiment Analysis

For this deliverable of the group project, you will use R to perform simple text mining, sentiment analysis, and visualization on the data collected about your topic. Almost all of the tasks for this deliverable have been covered previously in in-class assignments.

You will be submitting your team's R code and all generated plots in a zip file named **TeamName_Deliverable1.zip**

Following are the requirements for this deliverable:

1. Create a new R script file with a suitable name. In the comments at the top of the file, place your team members' names and the deliverable number. Make sure you continuously save your changes as you add to this script file. Add comments indicating which step each block of code addresses.

2. Write code that will create a single data frame that contains all of the tweet data for your topic from all files. Note that even though the files have a .json extension, they are text files containing multiple JSON documents each.

3. Write code that will create a **new data frame** containing the **cleaned tweets**. Cleaning should consist of removing URLs and emojis, and reformatting the created_at field.

4. Write code that will add three additional columns to the new data frame created in step 3: day, hour, and date. These columns will be created from data in the created_at field.

5. Write code that will create a **new data frame** containing all of the **non-stop words** from all of the tweets.

6. This task contains multiple parts, please label them as 6.a, 6.b, 6.c in your script file:
   a. Write code that will find the top 20 unique words in your tweets and visualize them in a plot using the ggplot2 package. Save your plot as a .png file and a .pdf file.
   b. Write code that will create a **new data frame** by **filtering out all non-ASCII rows** from the data frame containing all non-stop words.
   c. Rerun the code from part a on the newly created data frame. Save your plot as a .png file and a .pdf file.

7. This task contains multiple parts, please label them as 7.a, 7.b, 7.c in your script file:
   a. Write code that will create and display a word cloud from the top 100 words in your data frame containing the non-stop words. Save your plot as a .png file and a .html file (web page).
   b. Rerun the code from part a on the ASCII word data frame you created in 6.b. Save your plot as a .png file and a .html file (web page).
   c. Write code that will create a new word cloud from the ASCII word data frame with the following arguments: color = "random-light", backgroundColor = "azure3". Save your plot as a .png file and a .html file (web page).

8. Write code that will use AFINN sentiment analysis to find the top 15 positive and negative words in your tweets (from the non-stop word data frame) and plot them. Save your plot as a .png file and a .pdf file.

9. Write code that will use Bing sentiment analysis to find the top 15 positive and negative words in your tweets (from the non-stop word data frame) and plot them. Save your plot as a .png file and a .pdf file.

10. Write code that will use NRC sentiment analysis to find the top 10 words for each emotion (from the non-stop word data frame) and plot them. Save your plot as a .png file and a .pdf file.

Submit your work in the form of **a ZIP file** using the following format:
TeamName_Deliverable1.zip
Your zip file should contain 1 R script file, 8 png files, 5 pdf files, and 3 html files.

Only one person per group needs to submit the zip file.