CPSC 5305 Assignment 1

This assignment is intended to be completed individually, and any submitted work should be your original work. Please refrain from copying, pasting, or sharing code or solutions with others. While collaboration can be enjoyable and helpful, it is acceptable only to discuss the general approach to a problem at a high level. Under no circumstances should you directly copy or use someone else's written code without first attempting to solve the question independently.

If you have engaged in any form of collaboration or high-level discussion regarding the assignment, it is essential to list the names of all collaborators, including those with whom you discussed the problem at a general level.

Maintaining academic integrity and ensuring that your work reflects your own efforts is of utmost importance in this assignment.

What to submit on Canvas?

- Jupyter notebook **(pdf and .ipynb)** with answers to the questions listed below. You can upload a zip file containing both the pdf and .ipynb file
- Use proper code formatting.
- Use markdown cells to write questions, answer descriptive questions, and provide explanations. Use code cells to insert your code, run your code, and show output.
- The topmost cell of the Jupyter Notebook should mention the libraries that one needs to install to run the code
- Your code should run without any errors

Data: You are provided a dataset “trump_20200530.csv” that contains President Trump’s tweets from the moment he took office on January 20, 2017, to Mar 30, 2020. Download the data from Canvas to solve this assignment.

Note: If your laptop does not have enough computing resources to work on the entire dataset, please work on a subset of data. And clearly mention that in a markdown cell in the Jupyter Notebook.

Question 1: It is an important skill to look at the data and come up with questions that you can answer. What are some compelling questions that you can ask with the provided dataset (list at least 2 questions)? **(5 points)**

Question 2: Inspect & Data Cleaning

- **Inspect:** Write code to inspect the data. What do you observe? Along with the code, write your observation in the markdown cell. **(2.5 points)**

- **Clean:** Write code to clean the data. Use at least 5 methods. For each method, along with the code, you need to write the rationale behind the cleaning process. For this question, you can assume that you are solving one of the questions that you wrote for the first question. **(5 points)**
- **Tokenize:** Write code to tokenize your entire dataset. Use at least 2 different types of tokenizers. Compare their results and write your observations. **(2.5 points)**

Question 3: Learn how to use new Python packages or online APIs (10 points)
Pick an existing package a library or an API to determine the sentiment (positive, negative, neutral) for each of the tweets in the dataset. You can also use open-source code provided on GitHub repos. DO NOT WRITE THE SENTIMENT ANALYSIS CODE FROM SCRATCH

For example: Use VADER sentiment analysis from NLTK (link)

Question 4: Analyze Data over time (10 points)
How does the sentiment of your corpus change over time? Answer this question by showing plots (at least 2 graphs). Be creative!

Question 5: Drawing inference from the plots (5 points)
What can you infer from the plots?