Analyzing Text Data Rubric

DS 4002 - Spring 2024 - Channing Pitts
Due: May 10, 2024
Submission format: GitHub repository (submitted by link to Canvas)

Individual Assignment

Why am I doing this?
The goal of this assignment is to expose you to scraping information from the internet. Social media provides real world and real time information from any of its users so learning how to grab this data will allow you to acquire an up and coming skill. This assignment requires you to scrape information, clean text data from users, and perform an analysis that will be presented to your peers.

What am I going to do?
You will use the materials and data provided to you for this analysis (an occasional Google search to understand code can be helpful)! All of this information can be found here https://github.com/channingpitts/CS3_DS4002. Follow the README.md for more directions and guidance on the Github repository.

Tips for success:

- Take some time to read the 'Materials' Section on the Github repository.
- Do not be afraid to take a few minutes Google searching for more information to help you better understand the code and how it works!
- Document your code with comments so in the future, you can come back later and remember what you were doing.

How will I know I have succeeded?
You will meet expectations of this project when you follow the criteria in the rubric below.

Formatting	● One Github Repository (submitted via link on canvas) ● To ensure **reproducibility**, the repository will adapt parts of the TIER Protocol 4.0. In a nutshell, the top level page of the repository should contain: ○ A README.md file (which auto displays) ○ A LICENSE.md file (use MIT as default) ○ A SCRIPTS folder ○ A DATA folder ○ AN OUTPUT folder
README.md	● Goal: This file serves as an orientation to everyone who comes to your repository, it should enable them to get their bearings.

	● Section 1: Intro and Guidance o In this section, you should write a summary paragraph on the topic of this project. This helps future viewers to get a baseline understanding of what they will be looking at. In addition, please briefly explain the steps taken to produce your results and include any aspects that caused problems. ● Section 2: Software and platform section ○ The type(s) of software you used for the project. ○ The names of any add-on packages that need to be installed with the software. ○ The platform (e.g., Windows, Mac, or Linux) you used. ● Section 3: A Map of your documentation. o In this section, you should provide an outline or tree illustrating the hierarchy of folders and subfolders contained in your Project Folder, and listing the files stored in each folder or subfolder.
LICENSE.md	● Goal: This file explains to a visitor the terms under which they may use and cite your repository. ● Select an appropriate license from the GitHub options list on repository creation. ● Usually, the MIT license is appropriate.
SCRIPTS folder	● Goal: This folder contains all the source code for your project. ● Try to name each script according to the order it needs to be executed to reproduce the results. ● Throughout all your scripts, you should include copious comments explaining what each command or sequence of commands accomplishes and what the purpose is.
DATA folder	● Goal: This folder contains all of the data for this project. ● Include original and cleaned datasets for Simone Biles and Jonathan Owens. ● If your data fits in github, place all of it here. ● If your data does not fit in GitHub use a single file explaining the process to obtain the dataset.
OUTPUT folder	● Goal: This folder contains all of the output generated by your project, e.g. figures, tables, word clouds, etc. ● Use informative names for your files.
References	● Link any additional references used (you do not need to reference the documents I provided) ● Use IEEE Documentation style (link)