Yelp Sentiment Analysis Case Study Rubric

DS 4002- Fall 2024 - Aniyah McWilliams
Due: Monday, December 9th
Submission format: Upload link to GitHub repository to Canvas

Individual Assignment

General Description: Submit to Canvas a link to your work for the case study. This includes all the necessary components for someone to understand the dataset, your visualizations, the sentiment analysis, and the predictive model.

Why am I doing this? This case study allows you to leverage your data science knowledge, while simultaneously introducing you to the new concepts of sentiment analysis. Combining previous knowledge and new ideas will allow you to analyze and predict sentiment in restaurant Yelp reviews. As you work through this assignment, you will be exposed to how data analysis can be utilized for real-world applications.

- Course Learning Objective: Explore and graph key variables in the dataset
- Course Learning Objective: Analyze the sentiment of the restaurant reviews
- Course Learning Objective: Build and evaluate a predictive model

What am I going to do? The GitHub repository for this case study can be found at https://github.com/aniyahlater/DS4002-CS2. The dataset that you will be using for this case study can be found at the following link. After you arrive at the Kaggle page, download the dataset named "top 240 restaurants recommended in Los Angeles 2.csv". After downloading the dataset, begin by performing Exploratory Data Analysis (EDA) and cleaning the data to the best of your ability. For reference, a cleaned version of the dataset is available in the "Materials" folder and is named "Cleaned Dataset for Reference". Once you have cleaned the dataset, generate relevant EDA visualizations to uncover key patterns and insights. Next, use the provided code to guide you through installing VADER and applying it for sentiment analysis of the text. Create informative visualizations that highlight important findings from the text analysis. Finally, use the provided code to build a Naive Bayes model that predicts the sentiment of future Yelp reviews. Afterward, evaluate the model's performance using key metrics. All guidance codes for the EDA, VADER sentiment analysis, and predictive model is available in the "Materials" section under the file name "Guiding Code for Case Study".

Tips for success:

- Do not approach sentiment analysis with a binary mindset. Understanding sentiment requires nuance so attempt to look at keywords beyond the very fluid categories
- Make your file names (i.e., the maps) and your variable names easily interpretable.
- When writing attempt to write it as simply and concise as possible

How will I know I have Succeeded? You will meet expectations on the Yelp review text analysis case study when you follow the criteria in the rubric below.

Spec Category	**Spec Details**
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 ● Brief summary of your process for the case study ○ Should include the conclusion and results ○ Be concise and simply ○ Be comprehensible ● Brief outline or tree illustrating the hierarchy of folders and subfolders and what is listed in what ● An individual should be able to understand the structure and case study process based on this document
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. ● Include all the scripts you used. Try to name each script according to the order it needs to be executed to reproduce the results ● Well documented Jupyter Notebook file that contains the code

	used to execute the EDA, text analysis and predictive modeling ○ It could be smart to organize your code into these three sections ● This folder must include: ○ EDA code ○ Text analysis using VADER code ■ Some graphs displaying the variation ○ Naive Bayes model to predict the sentiment of a text revie ■ Include some evaluative metrics ■ Include header comments that detail the purpose of the code; should have copious comments explainin various chunks of code
References	● All references should be listed at the end of the document ● Use IEEE Documentation style (link)

Acknowledgements: Special thanks to Jess Taggart from UVA CTE for coaching on making this rubric. This structure is pulled from Streifer & Palmer (2020).