IMDb Reviews Sentiment Case Study Rubric

DS 4002 – Ashley Nguyen

Submission format: Upload link to GitHub repository on Canvas

Individual Assignment

General Description: Submit a Canvas link to your GitHub repository for the results of your case study. Instructions for what should be included in the repository can be found below.

Why am I doing this? This case study is an opportunity to apply your data science skills to text data to extract the sentiment from various IMDb movie reviews.

What am I going to do? The GitHub repository for this assignment can be found at: https://github.com/ashleynguyen04/DS4002/tree/main/CS3. You will download two datasets as zip files. One with all contain all the IMDb movie reviews (https://ai.stanford.edu/~amaas/data/sentiment/) and the another will contain the IMDb movie metadata (https://datasets.imdbws.com/). You will then merge these into one dataset linked by their IMDb movie identifier (tt#######) and create a Multinomial Navies Bayes Model to extract their sentiment as good or bad. You will display the ratio of bad:good reviews for each genre and measure accuracy by seeing if the bad and good sentiment pulled from the model matches if the review itself was positive (>7 stars) or negative (<4 stars).

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

Spec Category	Spec Details
Formatting	• GitHub repository (submitted via link on Canvas) that contains the following ○ README.md

	○ LICENSE.md ○ Scripts folder ○ References
README.md	• Use markdown headers to divide content • Section 1: Software and platform section ○ Software used for project ○ Packages installed ○ Platform used • Section 2: Map of documentation ○ Outline or tree illustration of the hierarchy of folders and subfolders • Section 3: Instructions for reproducing results ○ Step-by-step instructions to reproduce the results
LICENSE.md	• This file explains the terms under which they may use and cite your repository. • Select an appropriate license from the GitHub options list on repository creation (recommend MIT)
Scripts folder	• This folder should contain all the source code used for your project • Include all the scripts you used to execute the dataset creation, EDA, and sentiment analysis. • Ensure all script files have proper comments so someone can easily follow your code
References	• All references should be listed in a PDF file • Use IEEE Documentation style