CS – Amazon Book Review Sentiment Analysis Rubric

DS 4002 – Fall 2024 - Instructor: Loreto Alonzi
Due: One week after assigned
Submission format: Link to GitHub repo

Individual Assignment

General Description: Follow research project instructions to replicate this Data Science case study. Submit to Canvas a link to your case study repository.

Preparatory Assignments – Everything in the course

Why am I doing this? As a data scientist, it is important for you to have experience working through various research problems across fields you may or may not be familiar with. While data science tools are universal , it can be uncomfortable to be thrown into a problem whose context you have little prior experience with. In these situations, it is important to stay calm and not rush into producing results. Attack the problem diligently through research; turn to your peers who may be more experienced with the topic at hand; ask questions and do not be afraid to try out new things. In completing this case study and pushing yourself outside of your comfort zone, you will become a more versatile data scientist.

- Course Learning Objective: establish data sets relevant to your hypothesis/model
- Course Learning Objective: create a functioning data science pipeline
- Course Learning Objective: prepare findings for presentation to your peers

What am I going to do? First, read the one-page project outline document to familiarize yourself with the project context, problem, and overall objective. Regardless of your experience with sentiment analysis and social media text, brainstorm how you may conceptually approach the problem. Then, turn to the datasets found in the GitHub repo for this case study and familiarize yourself with the current data format. Afterwards, follow the guidelines to produce the following:

- GitHub repository containing:
    - Finalized dataset in csv format, properly cleaned and ready for analysis
    - One code file containing all code for preprocessing, modeling/analysis steps, and any other code written
    - One page PDF document evaluating sentiment analysis results
    - A README file orienting viewers of your repo and providing any additional resources used

All of this will be submitted electronically via a link to a GitHub repository.

Tips for success:

- Take your time. Read all supporting materials before starting to clean the data or run any analysis.
- Talk to your peers, the professor, and the TA. You may not have experience working with string data and sentiment analysis before – don't be afraid to ask questions.
- Confidently make decisions. Once analysis is complete, decide whether or not you think that the sentiment analysis results derive any meaningful business insights.

How will I know I have succeeded? You will meet expectations for the Case Study when you follow the criteria in the rubric below:

Spec Category	Spec Details

Formatting	• Repository – A GitHub repo containing all materials.
	  ○ Submit a link to the repo
	  ○ Everything is contained in the repo or linked to it if appropriate
	  ○ Contents:
	  ▪ Final dataset in csv format
	  ▪ Source code file
	  ▪ PDF evaluation document
	  ▪ README.md document

Dataset	• **Goal:** This csv file should contain your final dataset as used for sentiment analysis • Using the "Books" and "Reviews: files in the case study repo, create a final cleaned dataset which merges book data and review data into a combined table where each row represents a specific book review • Final dataset should include columns for *book_title, book_price, rated_helpful, rated_unhelpful, text_word_count, vader_sentiment, authors, categories,* and *ratings_count*.
Source code file	• **Goal:** This file should contain all code written for preprocessing, sentiment analysis, and predictive modeling. • Sentiment score analysis and predictive modeling can be conducted using any appropriate tests/models/plots, but supplemental documents in the case study repo

	provide guidance as to which methods may be most useful • Code may be written in preferred language, but recommendation is to use Python
PDF Evaluation document	• **Goal:** One page PDF document assessing your sentiment analysis results and conclusions, including a recommendation to your Amazonian boss based on data. • Explore sentiment scores relationship with other key book review variables. Make visualizations to explore the distributions of various variables in relation with one another. • Attempt to create a predictive model using variables that appear to be most impactful/related to the Vader sentiment score assignments. What is most indicative of low Vader scores? What is most indicative of high Vader scores? Offer potential explanations.

README.md	• **Goal:** This file provides an overview of the contents of your repo and orients visitors
	• Use markdown headers to divide content
	• Format:
	  ○ Data section
	  ▪ Data dictionary
	  ○ Source code section
	  ▪ Explain usage of code
	  ○ References section
	  ▪ Include properly formatted (IEEE style) references to any additional sources used outside of resources already provides in this case study