CASE STUDY RUBRIC

DS 4002 - Fall 2024
Due: Dec 16
Submission format: PDF

Individual Assignment

General Description: Upload pdf and link to GitHub repo to Canvas

Why am I doing this? This case study is an opportunity to apply your data science knowledge to analyze and interpret real-world data. You will strengthen your technical skills in exploratory data analysis (EDA), topic modeling, and visualization, while also developing your ability to derive meaningful insights and test hypotheses.

What am I going to do? In this project, you will analyze text reviews from two University of Waterloo courses, ECON 101 and CS 115, to uncover key topics that define student experiences. The goal is to test the hypothesis that "course difficulty" is one of the most frequently discussed topics. This involves cleaning and preprocessing the dataset to handle missing values and remove irrelevant stop words, performing exploratory data analysis (EDA) to visualize patterns in the data, and applying Non-Negative Matrix Factorization (NMF) to identify and interpret topics from positive and negative reviews for each course. You will compare insights between ECON 101 and CS 115, evaluating whether the hypothesis is supported and highlighting similarities and differences in topics discussed. Deliverables include:

- A GitHub repository - containing your code, data, and visualizations.
- A PDF report - summarizing your analysis, findings, conclusions, and references.

Tips for Success

- Carefully clean and preprocess your data to ensure meaningful results.
- Use clear, labeled visualizations to communicate your findings effectively.
- Document your code thoroughly, making it easy to understand and reproduce.
- Test your hypothesis, ensuring all conclusions are well-supported.

How will I know I have succeeded? You will know you have succeeded when you meet the following criteria

Spec Category	Spec Details
Formatting	● Submit a GitHub repository containing ○ A README.md with:

	■ Project purpose, research question, and hypothesis. ■ Setup instructions and references. ■ Dataset explanation and links to resources. ○ LICENSE.md specifying project reuse terms (e.g., MIT or Apache). ○ Source code in a well-documented Jupyter Notebook or Python script. ○ An organized folder structure with the following directories: ■ DATA: Contains the dataset and related files (e.g., stop words list). ■ SCRIPTS: Includes the source code for analysis and modeling. ■ OUTPUT: Stores generated visualizations and model outputs (saved as PNG or PDF files).
Source Code	● Include a Jupyter Notebook or Python script with clearly structured sections for: ○ Exploratory Data Analysis (EDA): ■ Load the dataset and handle missing values appropriately. ■ Remove irrelevant stop words using a custom stop words list. ■ Generate and save at least three visualizations: ● A word cloud showcasing the most common words in the reviews. ● Histograms illustrating the distribution of "useful," "easy," and "liked" ratings. ○ Topic Modeling: ■ Use Non-Negative Matrix Factorization (NMF) to extract topics for the following subsets of the dataset: ● Positive reviews of ECON 101. ● Negative reviews of ECON 101. ● Positive reviews of CS 115. ● Negative reviews of CS 115. ■ Name each topic based on top-ranked terms and provide bar charts for the top descriptors. ■ Save all topic-related visualizations in the OUTPUT folder. ● Ensure the code is modular, well-documented, and easy to reproduce.
PDF Report	● Submit a PDF report that includes: ○ A concise introduction and summary of the research objectives and findings. ○ Visualizations, such as word clouds, histograms, and bar charts, with clear labels and explanations. ○ Analysis and interpretation of topic modeling results. ○ A comparison of findings between ECON 101 and CS 115, addressing the hypothesis. ○ Conclusions drawn from the analysis.
References	● Include citations in IEEE format for: ○ Datasets (e.g., Kaggle dataset used). ○ External tools or libraries referenced. ○ Any additional resources or papers used in the analysis.