

# Case Study: IMDB Movie Review Text Analysis

DS 1000 - Spring 2023 – Alexa Owen

Due: May 10, noon

Submission format: GitHub repository (submitted by link to canvas),

## Individual Assignment

**General Description:** Submit to Canvas assignments a link to your GitHub repository for this project.

### Preparatory Assignments/Classes

- Previous classes discussing the case study, classes discussing text analysis

**Why am I doing this?** Learning to be a data scientist is more than being able to code or understand statistics. As a data scientist, it will be your job to ask questions such as: Is this analysis ethical? What do I want this analysis to show me? Am I biased in my analysis to aim for better results? Why am I being asked to do this analysis? What is the end goal of the analysis? Questions such as these do not often come to students heads when given an assignment that explicitly outlines what the student needs to accomplish. Therefore, the goal of this assignment is to help you learn how to start thinking like a data scientist. Furthermore, another goal of this project is being able to take the proposed topic in the case study and come up with a way to accomplish your end goal.

- Assignment Learning Objective: Begin to learn how to think like a data scientist
- Assignment Learning Objective: Be able to apply the skills learned in class and from other resources
- Assignment Learning Objective: Be able to synthesize and explain your project

**What am I going to do?** For this project, you will first begin by reading the one-page hook document. This document will outline the topic of the case study being presented and what will be expected from your analysis. After reading through the hook document use the supplementary materials provided to learn more about the topic and get ideas for your analysis. The documents provide background information and possible analysis choices, but you can choose to complete the project in whatever way you feel completes the requirements. After deciding on the type of analysis you are conducting create a plan to help keep you on track and outline the steps you need to take to complete the project. From there, you will think of a research question and hypothesis to test your analysis against. Then, you will complete your analysis and upload your code, data, and figures to a GitHub repository. Finally, you will create an approximately 10-slide presentation that outlines your project from start to finish.

### Tips for success:

- Don't be afraid to step out of your comfort zone. This project is meant to test you and will likely include some aspects that you have never done before.

- Breathe. The point of this project is not to stress you out, but to help you gain a better understanding of data science and what being a data scientist means.
- Be creative. All you are provided are the basic requirements and a one-page hook document, the project is meant to seem a little vague, giving you creative freedom.
- Don't be afraid to ask for help. Part of learning and growing comes from asking for help. You aren't expected to know everything, and you are allowed to ask for a second opinion.

**How will I know I have Succeeded?** You will meet expectations on the Case Study Assignment when you follow the criteria in the rubric below.

Spec Category	Spec Details
Formatting	<ul style="list-style-type: none"> <li>• One Github Repository (submitted via link on canvas)</li> <li>• The main repository page should contain               <ul style="list-style-type: none"> <li>○ A README.md file</li> <li>○ A LICENSE.md file</li> <li>○ An ANALYSIS folder</li> <li>○ A DATA folder</li> <li>○ A FIGURES folder</li> </ul> </li> <li>• One approximately 10-slide presentation (submitted via PDF on canvas)</li> </ul>
README.md	<ul style="list-style-type: none"> <li>• Goal: This file is meant to give a brief overview of your project from start to finish</li> <li>• Look up markdown documentation to create headers, tables, ect.</li> <li>• Use headers to divide each section</li> <li>• Context section               <ul style="list-style-type: none"> <li>○ Write one paragraph outlining the background of your project</li> </ul> </li> <li>• ANALYSIS section               <ul style="list-style-type: none"> <li>○ Provide the packages used in your analysis and how the user should install said packages</li> <li>○ Explain how the code should be run</li> </ul> </li> <li>• DATA section               <ul style="list-style-type: none"> <li>○ Provide a link to where the data was originally downloaded from</li> <li>○ Create a table that outlines each variable (you can choose to just include the columns you are using for your analysis, or you can explain all of the columns in your dataset)</li> </ul> </li> <li>• FIGURES section               <ul style="list-style-type: none"> <li>○ Create a table with the figure name and a brief explanation of the figure</li> <li>○ No images actually have to be uploaded to this part</li> </ul> </li> <li>• REFERENCES section               <ul style="list-style-type: none"> <li>○ List all references in IEEE</li> <li>○ Include references for the original supplementary materials</li> </ul> </li> </ul>
LICENSE.md	<ul style="list-style-type: none"> <li>• Goal: Tells others what they are allowed to do with your work.</li> <li>• Recommended: MIT license</li> </ul>

ANALYSIS folder	<ul style="list-style-type: none"> <li>• Goal: This folder contains the code and any other appropriate documentation for your analysis</li> <li>• Upload all the code files you created</li> <li>• Upload any necessary supplemental materials or documentation</li> </ul>
DATA folder	<ul style="list-style-type: none"> <li>• Goal: This folder contains all of the data used in your project</li> <li>• Upload the data file to this folder</li> <li>• If the file cannot be uploaded, create a document that explains how to access your data</li> </ul>
FIGURES folder	<ul style="list-style-type: none"> <li>• Goal: This folder contains the images you created during EDA and with your analysis results</li> <li>• Upload every image/figure you created for this project to this folder</li> </ul>
References	<ul style="list-style-type: none"> <li>• Any additional references used that were not provided should be cited</li> <li>• Use IEEE Documentation style (<a href="#">link</a>)</li> </ul>
Presentation	<ul style="list-style-type: none"> <li>• Goal: This presentation should provide a complete summary of your project start to finish</li> <li>• Background/Context (1 slide) <ul style="list-style-type: none"> <li>○ Briefly introduce the topic of your project (1 slide)</li> </ul> </li> <li>• Research Question/Hypothesis (1 slide) <ul style="list-style-type: none"> <li>○ State your research question and hypothesis (1 slide)</li> </ul> </li> <li>• Analysis (2-3 slides) <ul style="list-style-type: none"> <li>○ Discuss your analysis plan, why you chose what form of analysis, and how you accomplished the analysis (2-3 slides)</li> </ul> </li> <li>• Results and Future Steps (2 slides) <ul style="list-style-type: none"> <li>○ Discuss your findings in relation to your research question and hypothesis (1 slide)</li> <li>○ Discuss possible additional questions that could be asked (1 slide)</li> </ul> </li> <li>• References/Questions (2 slides) <ul style="list-style-type: none"> <li>○ Put your references in IEEE format on the slide (1 slide)</li> <li>○ Add a slide for questions (1 slide)</li> </ul> </li> </ul>

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