Final Project Progress Report
Titi Lee
DSCI-510
11/13/25

Project Scope Update

When I started working on the project, my vision was to move through mountains of data and identify patterns through data visualization. I underestimated how much more time it would take to build the foundation of the database, from the database schema to the webscraper functions. Now that I've started, I have a clearer path to reaching the goal I set in mind, but I have adjusted my expectations for the project. At first I thought I'd be spending maybe 10-20% of the time collecting data, and 80-90% of the time analyzing and visualizing the data. However, I've since realized it's closer to 60-70% building the foundational program and functions to collect, organize, update, and store the data, before getting to the analytics and data visualization.

One thing that I hadn't considered deeply when proposing the project was how to present the database itself through a user-friendly command line interface. I visualized jumping straight to the GUI and design, when in reality, building the basic interface through python commands has helped me give this database more structure and organization, which I'm grateful for. For example, by creating a menu in the "main.py" program, I've been able to break down the functions I need to write based on what a user will likely want to have the capability to do. For example, "add title", "update title", or "search title".

Rather than focus on analyzing a lot of data, my goal is more focused on creating a functional, modifiable database with a more limited dataset to start with (instead of thousands of titles, starting with hundreds, to make sure the structure is sound) before scaling up.

Finally, prior to starting this project, we hadn't covered SQL schemas, but now that we have, I've pivoted my tables to using SQL, and it's made the organization of key data so much more efficient.

Data Sources

I used Beautiful Soup to scrape IMDB, and it's taking me longer to figure out than I expected. Unfortunately I have not been able to get to the API data source yet – though I plan to use YouTubeAPI to scrape comments from videos.

Issues/Difficulties

It took me longer than I expected to figure out how to parse through data and extract it from IMDB for my first data source. It involved a lot of troubleshooting and inspecting elements on the page, since the website is set up differently than the assignments we've had in class (with Wikipedia etc), and required a little more finessing when it came to identifying html tags, class/types of objects. In fact, while I've managed to get the code to scrape the page, I haven't yet managed to get the details all correct – as you'll see, the project is a work in progress.

However, now that I've created the skeleton and structure for how this project will go, it will be easier for me to build it. Many of the later steps will mirror errors I figure out early on, so I'll be able to properly collect the data I need and analyze it once I get the tags and details right for web scraping.