Jamie Kemman & Holly White

CIS 388

April 1st, 2025

**Data Queeries Project Proposal**

We will be creating a movie recommendation search engine based on vector databases that relate to public datasets with information about a large selection of movies. The data collected will consist of the release year, title, origin, director, plot, genre, wiki page URL, and long form description of the plot. We plan on simplifying the creation of the search engine by training it on the vectorized data. The user will be able to input a movie that they enjoy, and the search engine will suggest three movies that are similar and provide their respective Wikipedia links so that the user might learn more information about them. If the user provides a title that the search engine cannot find data on, then it will ask the user to describe to input 3-5 key words describing the movie to offer a match. In this instance, the search engine will also give the user a warning that the results may be less accurate, and it will suggest that the user input key words found specifically on the movie’s Wikipedia page, in order to best match the data within the search engine’s database. Our group members or Jamie Kemman and Holly White, Jamie Kemman will be our team leader.

This project is an appropriate choice for CIS 388 because it will allow us opportunities to gain experience with a variety of skills. We will gain experience with training search engines and machine learning which are ever relevant within the current state of technological advancements. We will also gain experience with processing and vectorizing large amounts of data. We will also have a better understanding of rust, the program language we have chosen to complete the project in. Learning this language will be advantageous because it is growing rapidly in popularity and usage, and it rivals even python due to its memory stability. Finally, a tool such as a movie recommendation system is relevant today due to the large amounts of streaming services that are offered. This tool could not only be useful to individuals deciding on their next movie, but also to streaming services that are trying to determine what titles to offer based on titles that are popular.

We will be conducting this experiment using datasets found online through public dataset sources such as Hugging Face and Kaggle. The main dataset that we will be using was found on Kaggle, (<https://www.kaggle.com/datasets/jrobischon/wikipedia-movie-plots>) and has data for ~35,000 movies. This will not be a relational database because we will be creating a vector database that will rely on vector embedding. We will be using rust to complete this project. We will be using a combination of both Doom Emacs and VSCode to implement this project. We will also be using a variety of data science libraries such as Linfa, SmartCore, RustLearn, and RustyMachine. All of these libraries are rust libraries that are analogous to libraries such as Scikit-Learn in Python. We will communicate through both text and email, and we plan on meeting once a week on campus to work collaboratively and we will both work independently on the project throughout the week.

In terms of division of workload, Jamie will focus on researching the proper syntax for completing this project in Rust, determining the appropriate model for the project, and setting up a vector database. Holly will focus on preprocessing and processing the data and debugging and troubleshooting any issues that may arise. Holly will also take the lead on any reports or presentations necessary of the project. Both members feel as though the work is divided evenly and that the roles are well suited to our individual strengths.

In summary, we will be creating a movie recommendation program that will accept a movie title or 3-5 key words and return a short list of movie suggestions. We will be using Rust, a variety of data science libraries, and a combination of Doom Emacs and VSCode to complete this project. The data used will be taken from public databases of movies and their associated Wikipedia pages. Both members of this team, Jamie and Holly, will have separate responsibilities on this project, but will collaborate and support the team throughout each step of the process.

