Project Proposal – CS410, Fall 2022

Name: Chenfeng Chu; NetID: chu61

Team Information

I will be doing this project by myself considering my full-time job responsibilities and additional classes enrolled this semester. Even though the scope of work I could do might be limited, it gives me good opportunities to understand the end-to-end process to build it thoroughly. Thanks for the understanding.

Project Topic: Favorite Movie Recommender

I would like to pursue a free topic project which will recommend similar movie and TV shows based on user's favorite movie. I plan to build an application that asks a user to enter the title of their favorite movie and TV show and it will fetch the movie description for it and then recommend the top 3 movie and TV shows whose description has the highest similarity with the one chosen with matching scores.

Recommender systems are everywhere these days and I'm very interested in develop one myself to understand what is going on behind the scene and learn the logic behind all the recommendation we are receiving from many applications. We also discussed content similarity push during the first half of the semester, and I think this is a very relevant topic for me to pursue as a project for the course.

I plan to use Python for the project. Specially, I plan to use Dash to build the front-end user interface which takes user input and return recommended movie based on similarity with movie descriptions (initially thinking about BM25 but will explore a few other approaches later on). I plan to use the Netflix Movie and TV show dataset for my project, which can be found at this link: https://www.kaggle.com/datasets/shivamb/netflix-shows?resource=download. It contains about 8,800 movie and TV show records that include basic information for a movie or TV show such as title, cast, year, length and short descriptions.

Evaluation involves user feedback as for majority of the NLP use cases. For this project scope, we will assume the top ranked recommendation would be relevant for the users (i.e., pseudo relevance feedback). The output of my project can be easily expanded in the future to include feedback input functionality that will seek user feedback on whether the recommended show is useful or not. It can also be expanded to incorporate additional features other than description to make it more robust.

The project will for sure take me more than 20 hours to complete given my proficiency on Python and tools that will be used; but I'm excited to get it started. Below is a rough estimation for each step:

- Ingest, understand and pre-process dataset – 5 hours
- Develop effective algorithm to calculate similarity score between two movies – 10 hours
- Build front end application layout and config application logic to take in user input and return model output – 10 hours
- Conduct test and debugging – 5 hours