Current Progress

-Web scraper and backend database

We have completed our scraper for getting essential information from IMDB and TMDB websites and constructed a schema for . The extracted data is pipelined into the local mongoDB.

-Web frontend and host

A temporary frontend web page is constructed for a web hosting test. We rented a personal serverer and completed Nginx settings to host our react frontend project.

-Text corpus and tags

We have built up a text corpus to indicate the relations between the plot overview of various movies. We have started to tag all the movie data into several rough clusters and obtain all the required values.

Remaining tasks

-Main similarity algorithm

After tagging all the movie data, we will construct a likelihood algorithm to obtain the similarity between movies based on their plots. We will train the algorithm based on some sample data.

-Backend API server

The algorithm will be hosted on a backend server. The user requests of the website will be parsed to the backend where the results will be pushed back to the frontend website.

Challenges

-We found out that IMDB websites are not very easy to scrape and managed to generate groups of urls and scrape the search result page of constructed queries.

-Sample data for the likelihood algorithm is hard to obtain. We decided to use text structural similarity combined with sample data to train our algorithm.