CS410 Fa21 Project Proposal

Team Members

Team name: Chris Nolan's Fans Club
Below is the list of the team members in this project.

Name	**NetID**
Anasthasia Amelia Sugiharso*	aas13
Aravind Pillai	pillai5
Mike Zhou	mikez2

** team captain*

Topic

Our team chooses the Free Topics theme for the CS410 Course Project. The topic is scraping the IMDb website to capture the types of films individual actors act in and allow searching for actors by arbitrary topics and showing a list of movies matching those topics. For example, searching for the phrase "time travel" should return a list of actors, directors and producers who have been in movies related to the topic of "time travel". It is an interesting task because the existing IMDb website does not support the capability of searching for actors, directors, or producers based on the types of movies they have been involved in.

The planned approach is to use the IMDb basic names dataset that we can get from https://www.imdb.com/interfaces/. Each entry is an actor, director or producer name and a list of the titles they have been involved in. We will use the titles to scrape the IMDb plot summary website for that title and build an inverted index for each title. Then the person's inverted index will be the sum of the indexes for the titles they have contributed to. We will use a search algorithm to power a search for actors based on user queries. The service will be hosted on a website and the inverted indexes stored in a database.

The expected outcome is to create a website where users can search for actors, directors, or producers based on the content of the movies they have been involved in. The programming languages we are planning to use are including but not limited to Python and JavaScript. The work will be evaluated by relevance judgement by the group members.

Work Estimation

Below are the breakdown of tasks that are needed to be completed in this project.

Tasks	**Time (hours)**
Set up environment	3
Download and understand IMDB data sets	5
Scrape plot summary for each movie from IMDB web page and create document list.	15
Associate each actor to a list of documents (movie plots)	5
Create inverted index for the list of documents	10
Use BM25 search algorithm and rank the documents based on user search	15
Build the website user interface to display the results	10
Test the website	10
Report and Documentation	10
Total workload estimate	***83***