Topic: IMDb Actors Search Based on Movie Topics
Team: Chris Nolan's Fans Club

Progress

Below is the progress made in this project:

- [Done] Download and understand IMDb data sets
  We use the following datasets for this project:
  https://datasets.imdbws.com/name.basics.tsv.gz
  The dataset contains information that we will use throughout the project such as: `nconst`, `primaryName`, `birthYear`, `primaryProfession`, and `knownForTitles`.

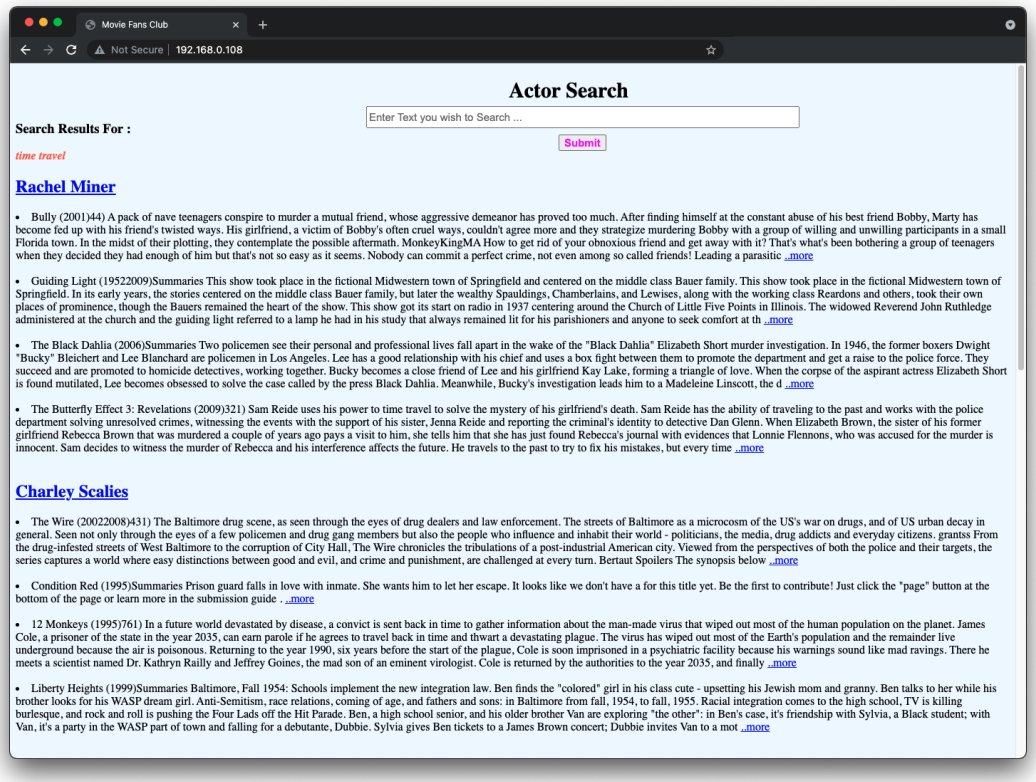
- [Done] Data scraping
  Currently we filter out the datasets with certain parameters (e.g `birthYear` > 1960, `primaryProfession` is actor/actress, and `knownForTitles` count > 3).
  This results in 771,701 actors. For the first step, we only used 2,500 actors' data (exported to file `actors.csv`). Based on this actors data, we scrape each actor's movie plot summaries from the IMDb website. It took around 2 hours to scrape 2500 actors' movie plot summaries.

- [Done] Create document list
  Each scraped plot summary is written to a file that becomes the document. The document list consists of files of plot summaries. These documents can be found in the `movieFile/` folder.

- [Done] Associate each actor to a list of documents (movie plots)
  Create a single document for each actor that contains a list of the actor's movies' plot summaries. This document will be used to show a list of movies of the corresponding actor in the search results. These documents can be found in the `actorFile/` folder.

- [In progress] Implement/tweak BM25 search algorithm
  We currently use BM25 search algorithm from https://github.com/dorianbrown/rank_bm25

- [In progress] Web interfaces
  We have displayed the working solution in a simple web interface.



Remaining Tasks

- Improve the search algorithm. The current search functionality does not include stop-words removal and stemming. This will be added to improve its performance.
- Improve the website performance. Pre-prepare display fields. Explore options to use cache.
- Scrape more actors data (probably total of 10,000 actors) to have a bigger corpus.
- Store the query frequency to a file.
- Testing and evaluation by relevance judgement.
- Project demo and documentation.

Challenges/Issues

- Takes time to scrape the data.
- Plot summary missing for some of the movies
- Web page performance issues.