Final Project Progress Report

Scraping and Ranking RottenTomatoes

UIUC: Fall'21 CS 410 - Text Information Systems

Topic: Scraping and Ranking RottenTomatoes

Theme: Intelligent Browsing

Team Members: Jeremy Wisuthseriwong (jrw7), Munesh Bandaru (muneshb2), Supriya Puri (puri6)

Team Captain: Munesh Bandaru (muneshb2)

Overall Tasks Status

Tasks	Completion Percentage
Data Scraping	100%
Data refining	100%
Modelling & Evaluation	60%
Web Application for interaction	70%
End to end testing	0%
Project Report and Presentation	25%

Progress Made

1. Data Scraping:
   - Built Python script to scrape the urls for the main top 100 movies page
   - Built Python script to scrape the content and reviews from each movie page
2. Data refining:
   - Refined the content dataset to include information such as movie title, synopsis, rating, genre, cast, and critic reviews

3. Modelling & Evaluation
    - Performed initial modelling using BM25
    - Calculated ranking results for each sample query for the top10 movies
    - Performed initial evaluation by calculating average precision for each sample query and mean average precision for all the sample queries
4. Web Application to display results
    - Built a prototype of an interface for user query interaction
    - Include the title and url to be displayed in the webapp for the top 10 movies ranked according to the query

Remaining Tasks

1. Modelling & Evaluation
    - Continue modelling using BM25 parameters and other ranker algorithms
    - Continue relevance testing and evaluation using other algorithms such as NDCG@10
2. Web Application
    - Create a fully functional web app displaying the results for the input query
3. End to end testing
    - To verify the results are in sync with the query judgements that have been created by manually checking each movie for matching the input query.
4. Drafting Presentation and Project report

Challenges/Issues

- Debugging reasons for low precision results
- Domain research to improve the query matching by identifying appropriate stopwords.
- Challenges in manually ranking query judgements for evaluating our model
- Deciding on the most appropriate ranker algorithm