Project Proposal for Information Retrieval System Project

Team Member: Jiacheng Sun, Zhihao Wan

Movie Preference Recommendations Based on COVID-19 Era Rottentomatoes Review

Project Proposal

Jiacheng Sun
Js211@rice.edu
S01401314

Zhihao Wan
wanz@rice.edu
S01400505

Problem Statement

During the COVID-19 pandemic, many countries and regions have implemented strict government-mandated quarantine policies, which have had a negative impact on traditional in-person theaters and cinemas. However, the demand for film and television productions has remained high, and people were spending more time at home watching movies or TV series compared to the pre-pandemic era. Thus, producers may concern about what kind of films or tv shows that people enjoy during the pandemic. It requires the analysis of trend about the people's preferences for films or tv shows genres during the pandemic. We plan to collect review data and other related data from Rotten tomatoes, analyze the sentiment classification then return several data visualization graphs and conclusion about genre trends for producers.

Data

The data are from the rottentomatoes.com, which are as follows:

- <2019 Best Movie>
  https://editorial.rottentomatoes.com/guide/best-movies-of-2019/
- <2020 Best Movies>
  https://editorial.rottentomatoes.com/guide/the-best-movies-of-2020/
- <2021 Best Movies>
  https://editorial.rottentomatoes.com/guide/2021-best-movies/
- <2022 Best Movies>
  https://editorial.rottentomatoes.com/guide/best-2022-movies/

Those data we crawled from the above sources are stored in the GitHub link: https://github.com/JiachengSun0520/comp631-info-retrieval. Data are in for csv format with terms movie name, reviews, genre, language, release date, rating, etc.

Project Proposal for Information Retrieval System Project

Team Member: Jiacheng Sun, Zhihao Wan

Plan

In the Task 1, the main objective is to implement several packages to web crawl data from taget website. Collect all information including movies/ TV shows name, rating, genre, reviews, and related data based on the top movies/ TV shows from 2019 to 2023(Covid period). Save it to csv files to prepare for the following task.

In the Task 2 the main objective is to use search server Solr to process query. After the document parsing, we plan to upload the localhost csv data to Solr to run the indexing and searching process.

In the Task 3, the main objective is to create a user-friendly front-end interface and render the query result on the front-end interface. It includes the data graphs and the conclusion that we draw. The interface should include data graphs and a conclusion drawn from the results.

Algorithm and Techniques

Our application aims to check and process the data we did web crawl from the rottentomatoes by using html scrape information package like Beautifulsoup and automates browsers package Selenium, and associate with csv output documents. Besides, we use the Solr to search and return closest query document and render them to the front-end interface. Our preliminary decision for front-end is React, the backend is node.js or Django depends on the real development situation, and the MongoDB as the database.

Evaluation and Test

Since this project is designed to film producers who concern about the production direction or audience who curious about the film trends during covid period. We are supposed to collect the feedback from those people as a critical task. Also, we need to consider about the general precision and recall measuring our information align with users' queries.

Expectation

We expect to create a website application with whole front-end user-friendly interface to display the data plots and the backend for storing and aggregating data that we crawl from the Rotten Tomatoes, which supports to output the data plots based on queries. The result includes sentiment analysis results based on audiences' reviews, popular genre analysis based on top movie lists, top movie director analysis, and more. They are output as dot plots, bar graphs, scatter plots and so on. According to those quantified results, we will draw serval conclusion for producers as suggestions.