Task1 Report for Information Retrieval System Project
Team Member: Jiacheng Sun, Zhihao Wan

Movie Preference Recommendations Based on COVID-19 Era Rottentomatoes Review

Task 1: Data crawling

Jiacheng Sun
Js211@rice.edu
S01401314

Zhihao Wan
wanz@rice.edu
S01400505

Introduction

In this task1 report, we mainly focus on the get the documents with web crawling technologies. Those documents and data are from rotten tomatoes, a review-aggregation website. We concerned about the data during pandemic period (2019-2022), since the Covid-19 impacted the movie industry, and we cared about audience's reaction about movies in that period compared with movies not in that period. We aim to conduct data analysis on audiences' preferences, attitude, etc., based on audience reviews and classify data on other fields.

During our data crawling process, we found that we cannot fully crawling data by tracing the urls. We will introduce selenium in next sections to deal with the problems we meet on dynamic pages. And, overall, to scrape the data, we use the web parser, beautifulsoup4.

Implementation

To do the data crawling at Rotten Tomato, we chose to use the web parser to retrieve needed data from the html file and use the browser to inspect html elements we need. Here are the packages we used:

- Beautifulsoup4 4.11.2. This package contains methods to read and retrieve html elements including <a href>, <p>, etc.
- Selenium 4.8.3. This package helps us deal with dynamic changes on the live page. For example, the pagination on the review page of Rotten Tomato does not involve the url change, so we use this package to simulate the mouse click to retrieve all the reviews.
- Pandas. This package is used to save the data we retrieved into a csv file.

To do the complete data crawling for our dataset, we have a 3-level step.

At the first step, we manually found root urls and put them into a list. ["https://editorial.rottentomatoes.com/guide/best-movies-of-2019/", "https://editorial.rottentomatoes.com/guide/the-best-movies-of-2020/", "https://editorial.rottentomatoes.com/guide/2021-best-movies/", "https://editorial.rottentomatoes.com/guide/best-2022-movies/"]. After we have all the root urls,

Task1 Report for Information Retrieval System Project

Team Member: Jiacheng Sun, Zhihao Wan

we construct a loop to use beautifulsoup4 to open each html file and extract all urls <a href> tags and only save urls starting with "https://www.rottentomatoes.com/m/". Therefore we have all movie sites with url in the form of https://www.rottentomatoes.com/m/ + "movie_name", and then we add "/reviews" to each movie site url to get https://www.rottentomatoes.com/m/movie_name/reviews which give us the review site for each movie. Also, we saved movie names in this step.

At the second step, we have a list of movie-site urls. We again use beautifulsoup4 to open urls iteratively and save movie information by extracting the tags in the <li class= "meta-row clearfix" data-qa= "movie-info-item> tags. We only saved Genre, language, Date and Distributor information by giving the fixed order.

At the final step, we used the list of movie-review-site urls. Different from steps above, at this step we use selenium to iterate the urls. First, we pass the url into the webdriver, which is a library in the selenium. Then the beautifulsoup4 can read the dynamic page information from the driver. Next, same as what we did above, we extract all text from <p class= "review text" data-qa = "review-quote"> to get and save the review by using beautifulsoup4. After we scrape the current page, we use selenium to simulate the click event on the <rt-button class="js-prev-next-paging-next" theme="light" data-direction="next" data-qa="next-btn"> and repeat scraping until there is no button points to the next page.

During scraping webs, we keep writing to csv files using pandas.

Conclusion

We stored documents as 4 csv files according to the year. The format is in terms of movie name, reviews, genre, release date, language, etc. Our data base contains about 100k documents. Here is an example of the data structure. And all csv files have been uploaded to Github: https://github.com/JiachengSun0520/comp631-info-retrieval

web-scraper-order	web-scraper-start-url	name	name-href	all critic	all critic-href	review	Rating	Genre	language	Release Date (Theaters)	Release Date (Streaming)	Distributor
1676867407-1145	https://editorial.rottentomatoes.com/guide/best-movies-of-2019/	They Shall Not Grow Old	https://www.rottentomatoes.com/m/they_shall_not_grow_old				R (Disturbing War Images)					
1676867407-1146	https://editorial.rottentomatoes.com/guide/best-movies-of-2019/	They Shall Not Grow Old	https://www.rottentomatoes.com/m/they_shall_not_grow_old					Documentary, History, Drama, War				
1676867407-1147	https://editorial.rottentomatoes.com/guide/best-movies-of-2019/	They Shall Not Grow Old	https://www.rottentomatoes.com/m/they_shall_not_grow_old						English (United Kingdom)			
1676867407-1148	https://editorial.rottentomatoes.com/guide/best-movies-of-2019/	They Shall Not Grow Old	https://www.rottentomatoes.com/m/they_shall_not_grow_old							Feb 1, 2019		
1676867407-1149	https://editorial.rottentomatoes.com/guide/best-movies-of-2019/	They Shall Not Grow Old	https://www.rottentomatoes.com/m/they_shall_not_grow_old								Mar 12, 2019	
1676867407-1150	https://editorial.rottentomatoes.com/guide/best-movies-of-2019/	They Shall Not Grow Old	https://www.rottentomatoes.com/m/they_shall_not_grow_old									Warner Bros. Pictures
1676867427-1151	https://editorial.rottentomatoes.com/guide/best-movies-of-2019/	They Shall Not Grow Old	https://www.rottentomatoes.com/m/they_shall_not_grow_old	View All Critic Reviews (156)	https://www.rottentomatoes.com/m/they_shall_not_grow_old/reviews	[A] moving tribute to the men who served.						
1676867427-1152	https://editorial.rottentomatoes.com/guide/best-movies-of-2019/	They Shall Not Grow Old	https://www.rottentomatoes.com/m/they_shall_not_grow_old	View All Critic Reviews (156)	https://www.rottentomatoes.com/m/they_shall_not_grow_old/reviews	They Shall Not Grow Old isn't just a compelling motion picture; it's an important slice of cinema.						
1676867427-1153	https://editorial.rottentomatoes.com/guide/best-movies-of-2019/	They Shall Not Grow Old	https://www.rottentomatoes.com/m/they_shall_not_grow_old	View All Critic Reviews (156)	https://www.rottentomatoes.com/m/they_shall_not_grow_old/reviews	By itself, the footage shown in They Shall Not Grow Old is not all that grand, but Jackson has made it so with 21st century technology, bringing to life a time						

In the following developing process, we will mainly process the content of reviews to analyze audiences' attitudes and their preference towards movie, including genre, language, rating, etc.