CS108 Project – Movie Mania

Avinash Chaudhari

April 28, 2024

Contents

1 Introduction

Moview is a movie information and rating platform featuring web scraping technology.

2 Project Setup

2.1 Images folder

Note:
The images for the website are not present in the project folder. Please download the images before running the website.
Link: https://drive.google.com/drive/folders/1-AyXc-4A4iuMBJQpCHjsUXB5iTMZrn9S?usp=drive_link
Download the `images.zip` file from the link and place the unzipped `images` folder in the project folder. Name the folder as `images`

2.2 Server

[1] To run the website on the server first install Node.js and run the following command in the terminal:

```
node server.js
```

This will start the server and thus sign in, sign up, rate and suggestions will work. Check if `.vscode` folder contains `settings.json` file and in that file `liveServer.settings.ignoreFiles` has `"node-server/**"` written in it. The `settings.json` file should look something like this:

Figure 1: Live Server settings

Now you can open the website with VS Code's Live Server extension.

3 The Web Scraper

A web scraper is a tool or program that automates the extraction of data from websites. It can gather information like text, images, links, and more, usually by parsing through the HTML code of web pages. Web scraping is often used for tasks like gathering market data, monitoring competitors, or aggregating content. All the information required for the site Moview has been scraped from IMDB [3] website

[1] As I did not know much about Express.js, I have used codeium[2] for making the server while wrote the rest of the javascript code

3.1 Applications

Some of the applications of a web scraper have been described in brief below:

- **Competitor Monitoring:** Web scrapers can help businesses track their competitors' pricing strategies, product offerings, promotions, and customer reviews. This data can inform competitive pricing, marketing campaigns, and product development.
- **Market Research:** Companies use web scraping to collect data on market trends, consumer preferences, and industry news from various sources like e-commerce websites, social media platforms, and news portals. This information aids in making informed business decisions and staying updated with market dynamics.
- **Lead Generation:** Web scrapers extract contact information such as email addresses, phone numbers, and social media profiles from websites, directories, or social platforms. This data is valuable for building marketing databases, generating sales leads, and reaching out to potential customers.

3.2 Requirements

- **Python:**
  - Python 3.8.10
  - `sudo apt-get install python3`
- **Selenium:**
  - Selenium 4.6.0
  - `pip3 install selenium`
- **Chromedriver:**
  - chromedriver 4.1.2
  - `pip3 install chromedriver`
- **Requests:**
  - requests 2.28.1
  - `pip3 install requests`
- **webdriver_manager:**
  - webdriver_manager 3.8.3
  - `pip3 install webdriver_manager`
- **urllib3:**
  - urllib3 1.26.13
  - `pip3 install urllib3`

3.3 Basic Idea of the Code

The webdriver is given the link to the website to be scraped i.e. the page that contains a list of movies. It then finds the link to each movie page and stores it in a list (max 100 movies). It then goes through each link and scrapes the required data from the web page:

- Movie name
- Year
- Advisory category
- Duration
- Rating
- Cast
- Directors
- Writers
- Storyline
- Genre
- Languages
- Images
- Trailer

3.4 Working of the Code

The Python code uses Selenium and Chromedriver to scrape the data from the website.The applications of these libraries are explained in the code using comments and also become clear as they get used later in the code. The use of different variables including lists, dictionaries etc also has been explained, moreover desciptive names have been used for the ease of undestanding. The driver takes in the url and opens it. The web page is then parsed and the information is stored in a dictionary. The dictionary of all the movies is then stored in a `movies.json` file while the downloaded images are stored in the images folder. The web scraper files can work in two modes: Headless and GUI mode. In headless mode the web browser is opened in the background while the web scraping is happening. In GUI mode the web browser is opened and one can see how the web scraping is happening (It feels quite amazing seeing it automated like that).

3.5 Usage

The usage of the files is simple. Open a terminal on your computer. Change the directory to the location where the python files are present (`webscraper` folder). Ensure that you have all the required libraries installed: selenium, webdriver_manager, chromedriver and requests. Now run the following command on the terminal:

```
python3 <filename>
```

Now the web scraping will start. It will append all the data in a `movies.json` file. This can be done for all the files in the folder. If you are confident enough on your internet and cpu speed then you can try opening multiple terminals and running many files in parallel to speed up the process.
The `movies.json` file will be used by the website to display all the movies.

3.6 Challanges and Solutions related to web scraping

- **Headless mode:**
  During the headless mode, the website gives a **403 forbidden** error. This error is caused by the web browser not being able to connect to the website. To solve this error, the web browser needs a user agent.

- **Error handling:**
  Sometimes the web scraper might fail to find the required elements. This can be due to the web browser not being able to load the website. This might be due to slow or unstable internet connection or due to content delivery network (CDN) issues. This is handled by using try-except statements in the code. So if some element is not found in the webpage then it will be given the value of `None`.

3.7 Precautions

If you have modified the web scraper files to add more information to be scraped, and commented or removed the `options.add_argument('--headless')`, then the web browser will open in GUI mode so you can see how the web scraping is happening. But beware, don't click any buttons on the website or open any links as this may interfere with the web scraping.

4 The Website

The website is made using HTML, CSS JavaScript and Node.js.

4.1 The Homepage

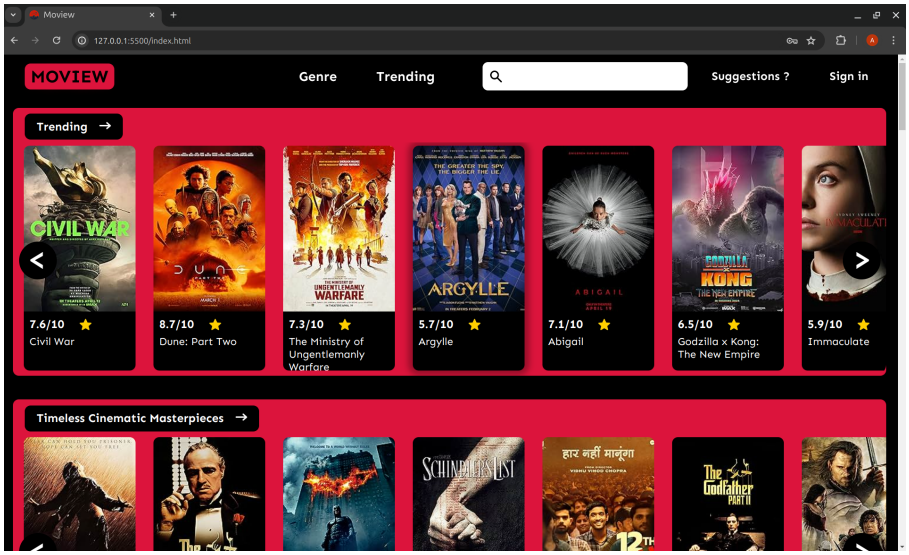

Figure 2: Moview Homepage

On the homepage you can browse the movies according to the genre while also search for a particular movie. To get more details about a particular movie, click on the movie poster. This will take you to the `movie.html` page.

4.2 Movie Details Page

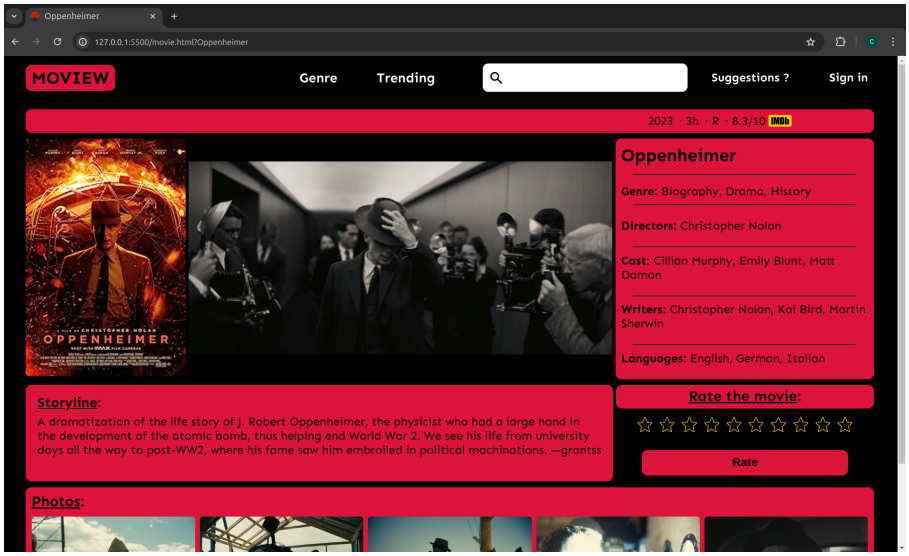

Figure 3: Movie Details Page

This page contains all the information about the movie. This includes:

- **Movie Name**
- **Movie Poster**
- **Movie Trailer**
- **Year of Release**
- **Duration**
- **Advisory Category**
- **IMDB Rating**
- **Genre**
- **Director**
- **Cast**
- **Writers**
- **Languages**
- **Storyline**
- **Rate the Movie**
- **Photos**

All this information has been arranged in a user-friendly and intuitive manner. The user can rate the movie through the **Rate the movie!** option. But this requires the user to be logged in. This can be done through the **Sign in** button present in the top right.

4.3 Sign in Page

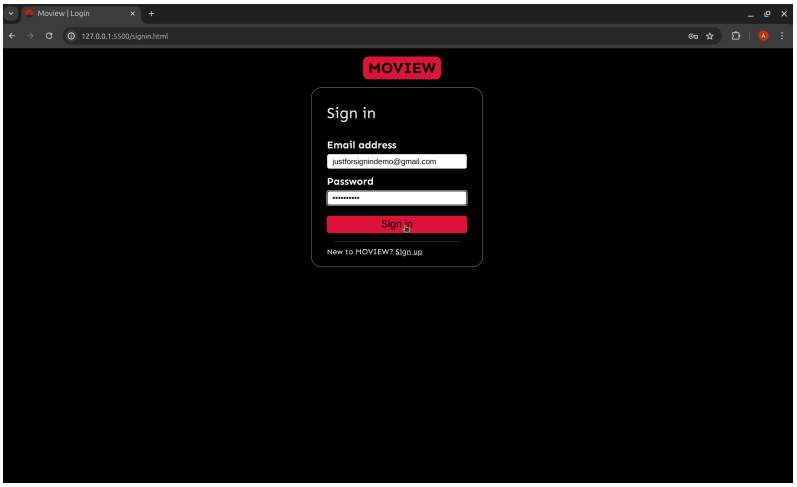

Figure 4: Sign in Page

To sign in, enter your email and password. Then click on the **Sign in** button. This will take you to the `index.html` page. If you don't have an account click on the **Sign up** button.

4.4 Sign up Page

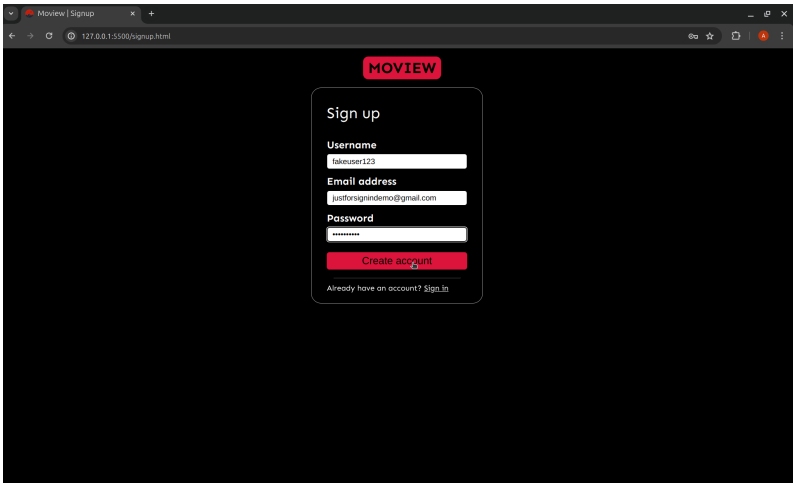

Figure 5: Sign up Page

You can create your account by entering your username, email and password.Then click on the **Create account** button. This will take you to the **index.html** page.
Both sign up and sign in require the server to be running. If you don't have a server running, then you can start the server by running the `node server.js` command.
You can see your profile by clicking on the profile photo in the top right. You can sign out by clicking the **Sign out** button.

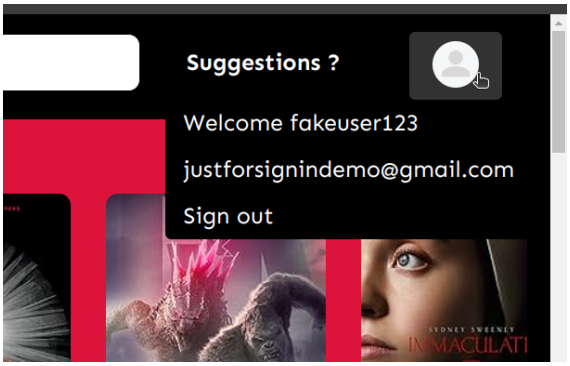

Figure 6: Profile

4.5 Search-bar

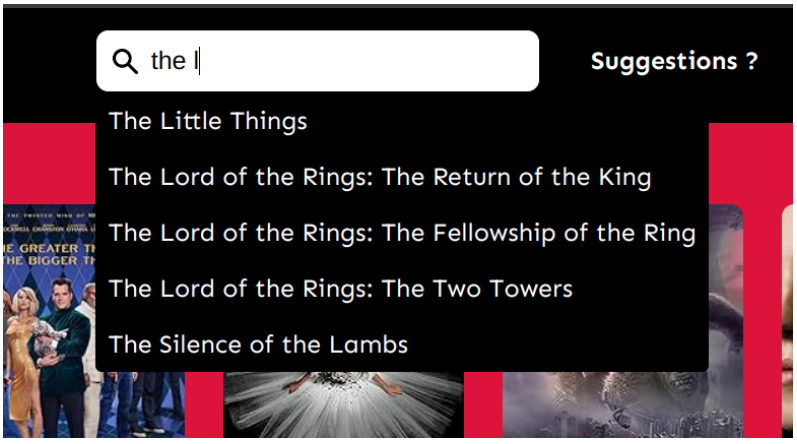

Figure 7: Search-bar

You can type in the search-bar to search for a particular movie. The search-bar will try to predict which movie you are looking for. The search will be case insensitive. By pressing **Enter** the search will take you to the first movie in the list. If there is no match it will take you to **404.html** page.

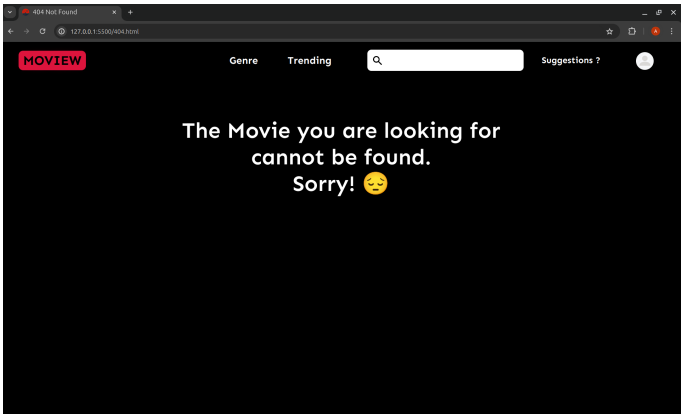

Figure 8: 404 Page

4.6 Search Algorithm

The search algorithm uses **Levenshtein Distance** [1] to find the closest match. The distance is calculated using the Levenshtein algorithm. This algorithm calculates the minimum number of edits (insertions, deletions, substitutions) required to transform one string into another. It is used to find the closest match to a search term in a list of movies by comparing the Levenshtein distance between the search term and each movie title. The closest match is determined by sorting the results based on the distance. But first the movies titles that contain the search term are showed first. This is to ensure that the user can quickly find the movie they are looking for. This does not take into account the spelling mistakes that may occur when only searching few terms of a movie title.

4.7 Suggestions Page

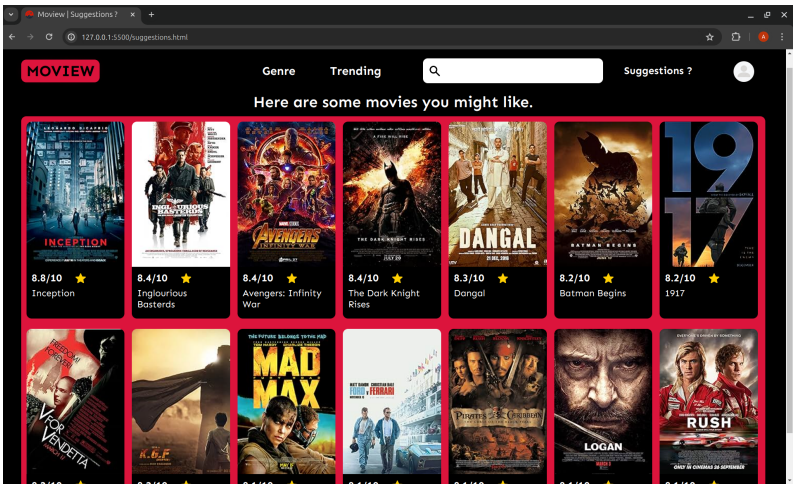

Figure 9: Suggestions Page

The suggestions page will only function if the user has rated at least five movies.

4.8 Suggestion Algorithm

The algorithm for this is written in `suggest.py`. The suggestions will be based on the users past rating. It takes the rating, genres, year and languages as its parameter for coming up with the suggestions. It calulates the top 2 genres, years and languages and then combines them to come up with the final suggestions.

Example:

Input Ratings (out of 10):

- *The Dark Knight - 9 (2008) (Action/Crime/Drama/Thriller) (English/Mandarin)*
- *The Lord of the Rings: The Return of the King - 9 (2003) (Action/Adventure/Drama/Fantasy) (English/Quenya/Old English/Sindarin)*
- *Top Gun: Maverick - 9 (2022) (Action/Drama) (English)*
- *Shichinin No Samurai - 3 (1954) (Action/Drama) (Japanese)*
- *Psycho - 3 (1960) (N/A) (English)*

- *3 Idiots - 9 (2009) (Comedy/Drama) (Hindi/English)*
- *Interstellar - 10 (2014) (Adventure/Drama/Sci-Fi) (English)*
- *Avengers-Endgame - 9 (2019) (Action/Adventure/Drama/Sci-Fi) (English/Japanese/Xhosa/German)*
- *Spider-man: Across the Spider-verse - 9 (2023) (Animation/Action/Adventure/Fantasy/Sci-Fi) (English)*

Suggestions:

Up to 14 movies are suggested. The better ones are:

- *The Dark Knight Rises (2012) (Action/Drama/Thriller) (English/Arabic)*
- *Batman Begins (2005) (Action/Crime/Drama) (English/Mandarin)*
- *Avengers: Infinity Wars (2018) (Action/Adventure/Sci-fi) (English)*
- *Inception (2010) (Action/Adventure/Sci-Fi/Thriller) (English)*
- *Pirates of the Caribbean: The Curse of the Black Pearl (2003) (Action/Adventure/Fantasy) (English)*

5 Customisations

- **Homepage:**
  The homepage shows a list of movies arranged according to their genre. The movies are clickable and take you to the `movie.html` page. Also the buttons in the nav-bar for genre and trending are clickable.
- **Trailer:**
  Trailer has been added to the movie details page. The trailer, set to mute, automatically starts playing once the page has finished loading.
- **Photos:**
  A maximum of 5 photos have been added to the movie details page.
- **Sign in and Sign up:** [4] [5]
  The Sign in and Sign up pages are working as expected. While the user is signed in, the user can also sign out.
- **Rate the Movie:**
  The user can rate the movie through the **Rate the movie!** option. The rating is stored in a file named `<email of user>.json`.
- **Search-bar Predictions:** [7]
  The search-bar tries to predict the movie you are looking for. The search will be case insensitive. It has been explained in the *Search-bar* [4.6] section.

6 Bugs

- **Search predictions:**
  The search predictions that are given below the search-bar do not disappear when the search-bar is out of focus.

References

[1] ChatGPT. Levenshtein algorithm. https://chat.openai.com/, April 2024.

[2] Codeium. server.js. https://codeium.com/, April 2024.

[3] IMDB Website. https://www.imdb.com/.