CS 210 – Introduction to Data Science

Movie Data Analysis Project

İpek Akkuş – 30800

19.01.2024

Motivation

The motivation behind this project is to gain a deeper understanding of personal movie-watching preferences, explore patterns in the data, and utilize web scraping techniques to enrich the dataset with additional features. By exporting information from the Letterboxd platform, the goal is to uncover insights into movie genres, themes, top actors, and directors, providing a comprehensive overview of movie-watching habits. This project intended to developed more to built a user-friendly search engine matching with the personal taste of watching movies.

Data Source

The movie dataset used in this project is sourced from a personal collection and is stored in a CSV file (`watched.csv`). The data includes information about watched movies, such as the title, date of viewing, and a link to the corresponding Letterboxd page. This link is used to enhance the dataset, meaning that language, top 5 actor, top 5 genres, top 5 themes, director etc. of that movie was not available in downloaded watched.csv document and added by web scraping. We can say that it was a home-made API thanks to the Letterboxd for not guaranteed use of its own API :)

Data Analysis

Techniques Used

- Web scraping with BeautifulSoup for extracting information from Letterboxd.

- Multi-threading using ThreadPoolExecutor to enhance data extraction efficiency.

- Descriptive statistics, bar plots, and histograms for exploratory data analysis.

Stages of Analysis

1. Data Extraction: Utilized web scraping techniques to extract information from Letterboxd, including top actors, directors, genres, themes, language, and statistics for each movie.

2. Data Cleaning: Handled missing values and applied specific cleaning steps, such as filling missing language values with the mode.

3. Exploratory Data Analysis (EDA): Conducted analysis on the enriched dataset to explore distribution patterns of genres, themes, top actors, and directors. Examined the time difference between movie release years and watch dates.

4. Visualization: Generated plots to visually represent key findings.

Findings

What I Learned About Myself

- Movie Preferences: Identified top genres, themes, and favorite actors/directors, providing insights into personal movie-watching preferences. Note that my favorite director is not Nolan, it is Tarkovsky who I watched all his movies, but as he has less movies than Nolan, Nolan seems to be the winner :(

- Temporal Patterns: Explored trends in the time difference between movie release years and watch dates, understanding how much I follow contemporary film releases. This was different than my assumptions of myself, so it was good to see.

Limitations and Future Work

Limitations

- Data Availability: The analysis heavily relies on the availability and accuracy of data from Letterboxd. Incomplete or inaccurate information on the platform may affect the results.

- Biases: Movies released before my birth year are excluded to avoid biases, but this may limit the analysis.

Future Work

- Improved Data Collection: Explore options to enhance data collection, such as integrating with more movie databases or using APIs for richer information.

- Advanced Analysis: Implement advanced analysis techniques, including sentiment analysis on reviews, collaborative filtering for movie recommendations, or applying machine learning for predictive modeling.

- Make the user entered searched movie, in a more proper way to search. There may be another key components that is related with my movie habits and we do not consider. So, it would be good to think on them and explore more interesting relations. Also, this search might be useful for me to pick the upcoming movie that I would watch from the movie pool. Considering that I am really an undecisive person in selection, it would be really helpful for me :)

Future Plans

I plan to continually update and refine the project, incorporating additional features and insights. Future iterations may include the integration of more data sources, implementing user reviews sentiment analysis, and enhancing the visualizations for a more interactive experience.

Feel free to contribute, provide feedback, or suggest improvements to make this project more comprehensive and insightful.

Let’s try to understand what my functions does.

Enhance_dataset.py

This Python code focuses on extracting and enriching movie data from the Letterboxd platform. It begins by loading a movie dataset into a Pandas DataFrame, converting date columns, and then defines functions for web scraping various details from Letterboxd, such as top actors, director, language, genres, themes, and ratings. The use of a ThreadPoolExecutor enables parallelized extraction for efficiency. After dynamically adding columns with extracted information to the DataFrame, the code performs data cleaning by filling missing language values with the mode. Finally, the processed DataFrame is printed to the console, and the results are exported to an Excel file named 'extracted1.xlsx'. The code showcases an effective approach to automate the extraction and analysis of movie-related information from Letterboxd.

You may see some parts of the output created by this code in here.

Index	Date	Name	Year	Letterboxd URI	Rating	Top Actors	Director	Genres	Themes	Language
350	2024-01-13 00:00:00	Vagabond	1985	https://boxd.it/T0u	nan	['Sandrine Bonnaire', 'Macha Méril…	Agnès Varda	['Drama', 'Humanity and the worl…	['Humanity and the world around …	English
349	2024-01-12 00:00:00	Inside	2012	https://boxd.it/3tPi	nan	['Engin Günaydın', 'Nergis Öztürk'…	Zeki Demirkubuz	['Drama']	None	English
348	2024-01-12 00:00:00	Benedetta	2021	https://boxd.it/gdNQ	nan	['Virginie Efira', 'Charlotte Ramp…	Paul Verhoeven	['Romance', 'Drama', 'History', …	['Faith and religion', 'Sins, f…	English
347	2024-01-10 00:00:00	Beginning	2020	https://boxd.it/qEyS	nan	['Ia Sukhitashvili', 'Rati Oneli',…	Dea Kulumbegashvili	['Drama']	None	English
346	2024-01-09 00:00:00	Anatomy of a Fall	2023	https://boxd.it/yuDE	nan	['Sandra Hüller', 'Swann Arlaud', …	Justine Triet	['Mystery', 'Drama', 'Thrillers …	['Thrillers and murder mysteries…	English
345	2024-01-06 00:00:00	How to Have Sex	2023	https://boxd.it/EFIM	nan	['Mia McKenna-Bruce', 'Lara Peake'…	Molly Manning Walker	['Drama']	None	English
344	2023-12-27 00:00:00	Life	2023	https://boxd.it/qAeO	nan	['Miray Daner', 'Burak Dakak', 'Ce…	Zeki Demirkubuz	['Drama']	None	English
343	2023-12-07 00:00:00	Afire	2023	https://boxd.it/y06K	nan	['Thomas Schubert', 'Paula Beer', …	Christian Petzold	['Romance', 'Drama']	None	English
342	2023-12-05 00:00:00	The Confession	2001	https://boxd.it/2KtQ	nan	['Taner Birsel', 'Başak Köklükaya'…	Zeki Demirkubuz	['Drama']	None	English
341	2023-11-25 00:00:00	Napoleon	2023	https://boxd.it/shJM	nan	['Joaquin Phoenix', 'Vanessa Kirby…	Ridley Scott	['Drama', 'History', 'War', 'Epi…	['Epic history and literature', …	English
340	2023-08-30 00:00:00	A Tale of Winter	1992	https://boxd.it/1WJO	nan	['Charlotte Véry', 'Frédéric van d…	Éric Rohmer	['Romance', 'Drama', 'Moving rel…	['Moving relationship stories', …	English
339	2023-08-30 00:00:00	A Tale of Autumn	1998	https://boxd.it/1WJY	nan	['Marie Rivière', 'Béatrice Romand…	Éric Rohmer	['Drama', 'Romance', 'Relationsh…	['Relationship comedy', 'Moving …	English
338	2023-08-26 00:00:00	All Quiet on the Western Front	2022	https://boxd.it/H08	nan	['Felix Kammerer', 'Albrecht Schuc…	Edward Berger	['Drama', 'War', 'War and histor…	['War and historical adventure',…	English
337	2023-08-24 00:00:00	A Summer's Tale	1996	https://boxd.it/183e	nan	['Melvil Poupaud', 'Amanda Langlet…	Éric Rohmer	['Comedy', 'Romance', 'Drama', '…	['Relationship comedy', 'Moving …	English
336	2023-08-24 00:00:00	The Broken Circle Breakdown	2012	https://boxd.it/4pHm	nan	['Veerle Baetens', 'Johan Heldenbe…	Felix van Groeningen	['Drama', 'Faith and religion', …	['Faith and religion', 'Moving r…	English
335	2023-08-22 00:00:00	Frances Ha	2012	https://boxd.it/41Xg	nan	['Greta Gerwig', 'Mickey Sumner', …	Noah Baumbach	['Drama', 'Comedy', 'Relationshi…	['Relationship comedy', 'Song an…	English
334	2023-08-19 00:00:00	Primal Fear	1996	https://boxd.it/28ws	nan	['Richard Gere', 'Laura Linney', '…	Gregory Hoblit	['Drama', 'Crime', 'Thriller', '…	['Thrillers and murder mysteries…	English
333	2023-08-19 00:00:00	John Wick: Chapter 4	2023	https://boxd.it/mh48	nan	['Keanu Reeves', 'Donnie Yen', 'Bi…	Chad Stahelski	['Action', 'Thriller', 'Crime', …	['Epic heroes', 'High speed and …	English
332	2023-08-18 00:00:00	The Nice Guys	2016	https://boxd.it/94Hg	nan	['Russell Crowe', 'Ryan Gosling', …	Shane Black	['Action', 'Comedy', 'Crime', 'C…	['Crude humor and satire', 'Thri…	English
331	2023-08-03 00:00:00	Wings of Desire	1987	https://boxd.it/2b26	nan	['Bruno Ganz', 'Solveig Dommartin'…	Wim Wenders	['Drama', 'Romance', 'Fantasy', …	['Humanity and the world around …	English
330	2023-08-03 00:00:00	Biutiful	2010	https://boxd.it/O1I	nan	['Javier Bardem', 'Maricel Álvarez…	Alejandro González Iñárritu	['Drama', 'Moving relationship s…	['Moving relationship stories', …	English
329	2023-08-02 00:00:00	Chocolat	2000	https://boxd.it/2aue	nan	['Juliette Binoche', 'Alfred Molin…	Lasse Hallström	['Comedy', 'Drama', 'Romance', '…	['Faith and religion', 'Moving r…	English
328	2023-07-31 00:00:00	Memories of Murder	2003	https://boxd.it/1TSo	nan	['Song Kang-ho', 'Kim Sang-kyung',…	Bong Joon-ho	['Thriller', 'Drama', 'Crime', '…	['Thrillers and murder mysteries…	English

Note that you can access xlsx files in the folder as well.

List_to_be_compared.py

Working mechanism is quite similar to the **Enhance_dataset.py.** Please see the explanation above. This function creates a searching movie pool that I may ask for correlation with my watched data, indicating my movie taste. You can see some parts of the output below, as well as the .xlsx file.

Date	Name	Tags	Letterboxd URI	escripti	Top Actors	Director	Genres	Themes	Language
1	Poor Things	2023	https://boxd.it/tNWU	nan	['Emma Stone', 'Mark Ruffalo', 'Will…	Yorgos Lanthimos	['Romance', 'Science Fic…	['Humanity and th…	English
2	World War III	2022	https://boxd.it/BZBU	nan	['Mohsen Tanabande', 'Mahsa Hejazi',…	Houman Seyyedi	['Drama']	None	English
3	The Boy and the Heron	2023	https://boxd.it/ipeM	nan	['Soma Santoki', 'Masaki Suda', 'Ko …	Hayao Miyazaki	['Animation', 'Adventure…	['Surreal and tho…	English
4	About Dry Grasses	2023	https://boxd.it/oNbO	nan	['Deniz Celiloğlu', 'Merve Dizdar', …	Nuri Bilge Ceylan	['Drama']	None	English
5	All of Us Strangers	2023	https://boxd.it/Bz3C	nan	['Andrew Scott', 'Paul Mescal', 'Jam…	Andrew Haigh	['Romance', 'Drama', 'Fantasy']	None	English
6	The Promised Land	2023	https://boxd.it/B1hA	nan	['Mads Mikkelsen', 'Amanda Collin', …	Nikolaj Arcel	['History', 'Drama']	None	English
7	Saltburn	2023	https://boxd.it/z4eg	nan	['Barry Keoghan', 'Jacob Elordi', 'R…	Emerald Fennell	['Comedy', 'Drama', 'Thr…	['Intense violenc…	English
8	Past Lives	2023	https://boxd.it/oNB8	nan	['Greta Lee', 'Teo Yoo', 'John Magar…	Celine Song	['Drama', 'Romance', 'Mo…	['Moving relation…	English
9	Heroic	2023	https://boxd.it/AafQ	nan	['Santiago Sandoval', 'Fernando Cuau…	David Zonana	['Drama', 'Thriller']	None	English
10	The Teachers' Lounge	2023	https://boxd.it/BINY	nan	['Leonie Benesch', 'Leonard Stettnis…	İlker Çatak	['Drama']	None	English
11	Infinity Pool	2023	https://boxd.it/oPSK	nan	['Alexander Skarsgård', 'Mia Goth', …	Brandon Cronenberg	['Science Fiction', 'Hor…	['Horror, the und…	English
12	Fools	2022	https://boxd.it/1enu	nan	['Dorota Kolak', 'Łukasz Simlat', 'T…	Tomasz Wasilewski	['Drama']	None	English
13	Green Border	2023	https://boxd.it/GqGq	nan	['Jalal Altawil', 'Maja Ostaszewska'…	Agnieszka Holland	['Drama']	None	English
14	Fingernails	2023	https://boxd.it/tJ82	nan	['Jessie Buckley', 'Riz Ahmed', 'Jer…	Christos Nikou	['Science Fiction', 'Rom…	['Relationship co…	English
15	Monster	2023	https://boxd.it/DJEM	nan	['Sakura Ando', 'Eita Nagayama', 'So…	Hirokazu Kore-eda	['Thriller', 'Drama', 'M…	['Moving relation…	English
16	Critical Zone	2023	https://boxd.it/HnYy	nan	['Amir Pousti', 'Shirin Abedinirad',…	Ali Ahmadzadeh	['Drama']	None	English
17	The Holdovers	2023	https://boxd.it/vHza	nan	['Paul Giamatti', 'Dominic Sessa', "…	Alexander Payne	['Comedy', 'Drama', 'Und…	['Underdogs and c…	English
18	May December	2023	https://boxd.it/vEE2	nan	['Natalie Portman', 'Julianne Moore'…	Todd Haynes	['Drama', 'Comedy', 'Mov…	['Moving relation…	English
19	Fair Play	2023	https://boxd.it/yirc	nan	['Phoebe Dynevor', 'Alden Ehrenreich…	Chloe Domont	['Thriller', 'Drama', 'I…	['Intense violenc…	English
20	The Delinquents	2023	https://boxd.it/qLyA	nan	['Daniel Elías', 'Esteban Bigliardi'…	Rodrigo Moreno	['Comedy', 'Drama']	None	English
21	Inside the Yellow Cocoon Shell	2023	https://boxd.it/Gama	nan	['Le Phong Vu', 'Nguyen Thinh', 'Ngu…	Pham Thien An	['Drama']	None	English
22	Dream Scenario	2023	https://boxd.it/v2h0	nan	['Nicolas Cage', 'Julianne Nicholson…	Kristoffer Borgli	['Comedy', 'Fantasy', 'R…	['Relationship co…	English
23	Subtraction	2022	https://boxd.it/C69o	nan	['Navid Mohammadzadeh', 'Taraneh Ali…	Mani Haghighi	['Drama', 'Thriller']	None	English

Read_extracted.py

This Python code conducts an exploratory data analysis (EDA) on a movie dataset extracted from Letterboxd. It utilizes Pandas, Matplotlib, and Seaborn for data manipulation and visualization. The EDA starts by displaying the first few rows, general information, descriptive statistics, data types, and missing values of the DataFrame. Subsequently, it delves into analyzing the top actors and directors by visualizing their occurrence counts in bar plots. The distribution of movie genres and themes is explored similarly, showcasing the most frequent genres and themes through bar plots. Finally, the analysis examines the time difference between the release year and the watched year for movies released after 2015, shedding light on potential biases in movie preferences based on release years. The code effectively combines data exploration and visualization techniques to gain insights into the user's movie-watching patterns and preferences. You may see the outputs on “EDA and Data Visualization Outputs” folder.

All of them may be interpreted easily but it is good to explain plot_time_differences a bit. It is created to show how much I am following the trends of the new releases of the year. The bar plot illustrates the average time difference between movie release years and the years they were watched, focusing on films released after 2015. Positive bars indicate the user tends to watch movies a certain number of years after their release. Small error bars suggest consistent behavior for movies released in a given year, while larger ones imply more variability. The analysis helps uncover trends and potential biases in the user's movie preferences based on release years. See figure 1 to clearly understand.

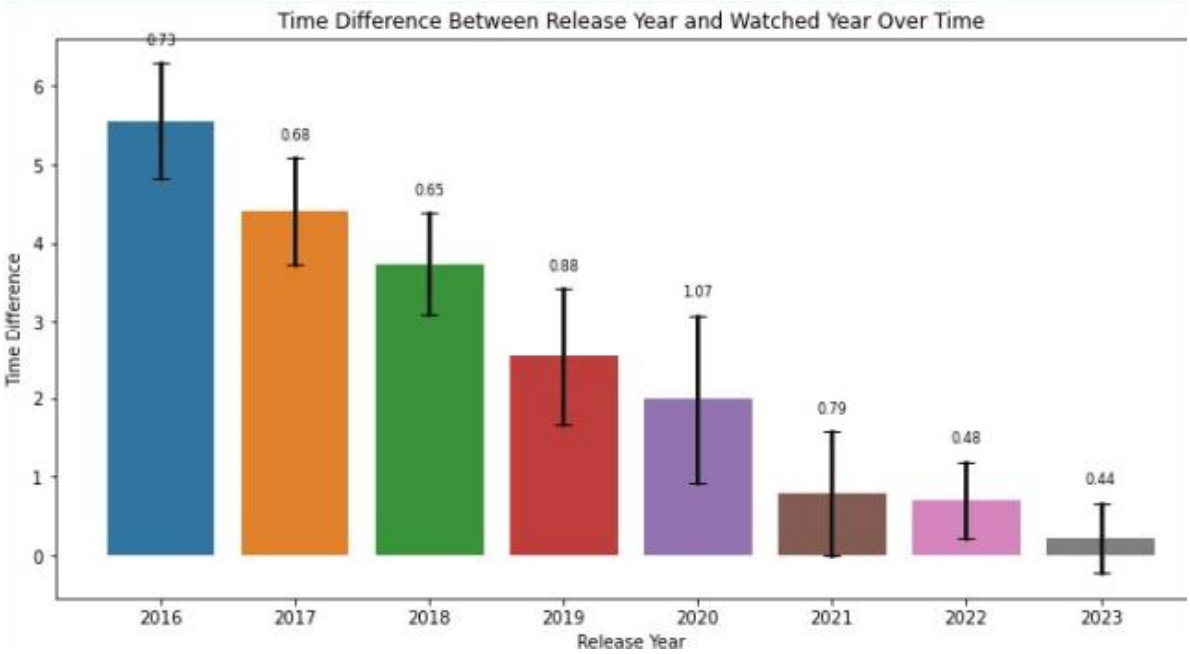

Figure 1. Time Difference Between Release Year and Watched Year Over Time

Please refer to the visuals which have names in the form of plot_top_NUMBER_VARIABLE.png in the outputs folder.

Explore_data.py

In this set of codes, I'm working on a content-based movie recommendation system. I start by preprocessing the dataset, converting string representations of lists, exploding columns with multiple values, and handling missing data. I then identify the top actors, directors, genres, and themes based on movie counts. To create a numerical representation, I manually assign scores to these top values and apply these mappings to encode the dataset. After that, for a user-entered movie, I calculate similarity scores by summing the encoded values for actors, director, genre, and theme. The dataset is then sorted based on these similarity scores, and I display the top similar movies. This process allows me to offer personalized movie recommendations based on the user's watched movies and preferences.

User may search for a movie from the pool and these codes calculates correlated movies from the pool that I have already watched and display similars. See sample search.

index	Date	Name	Tags	Letterboxd URI	Description	Top Actors	Director	Genres	Themes
0	1	Poor Things	2023	https://boxd.it/tNWU	nan	['Emma Stone', 'Mark Ruffalo', 'Willem Dafoe', 'Ramy Youssef', 'Jerrod Carmichael']	Yorgos Lanthimos	['Romance', 'Science Fiction', 'Comedy', 'Huma…	['Humanity and the world around us', 'Relat

As I explained above, this program is open to develop more and more. Anyways, hope you enjoyed so far!