Netflix Titles and IMDb Reviews: processing, retrieval and analysis

Ricardo Fontão
M. EIC
FEUP
Porto, Portugal
up201806317@up.pt

Telmo Baptista
M. EIC
FEUP
Porto, Portugal
up201806554@up.pt

Tiago Silva
M. EIC
FEUP
Porto, Portugal
up201806516@up.pt

***Abstract*—In the current age people start to spend more and more time watching movies and TV shows in their free time, utilising a multitude of services to provide their favorite shows. One of the most popular streaming service is Netflix, counting with an ever increasing number of monthly users and shows. This project joined Netflix titles and IMDb user review data to later develop a search engine.**

***Index Terms*—data collection, data preparation, data processing, data analysis, movies, TV shows, Netflix**

I. Introduction

Netflix has been steadily growing in subscribers over the years, constantly adding new TV Shows and Movies. To better understand people's opinion on these shows, in this first step of the project we combine a Netflix dataset with a dataset of reviews from the IMDb platform. In this report we will detail all the steps we took from downloading the data in the Data Collection section, processing it in the Data Preparation section and building the pipeline to do it automatically in the Pipeline section. We will also make an exploratory data analysis to understand the characteristics of the data and how it is distributed.

II. Data Collection

We required both show information as well as their respective user reviews. As such, our search for datasets started in the Kaggle platform, chosen thanks to the various diverse and high quality datasets it provides, most of which also offer an in-depth analysis of the data contained therein.

Our search for a comprehensive Netflix dataset yielded a continuously updated dataset [1] that contained the entirety of Netflix's library. This dataset contains the following 12 columns: the id of the show, title, type of show (TV Show, Movie), director(s), cast, country(ies) of origin, date added to Netflix, release year, parental rating, duration (minutes for movies and seasons for TV Shows), the genres of the show and its description. It contains 8807 rows and takes up approximately 3.4MB. To download the data we decided to use Kaggle's built-in public API due to it's free use, good download speeds and simple integration into the pipeline. To use this API, Kaggle provides a command-line utility that can be installed from pip (Python's package manager).

After having the Netflix titles, we required their respective reviews. Our first approach to solve this issue revolved around using IMDb's built-in API to download the necessary reviews. This idea was quickly discarded due to the limitations imposed by the API's free tier, which allows only two queries per second, making it unfeasible to download all the required data within a reasonable time frame.

The next step was to fall back to the Kaggle platform, where we found a dataset [2] containing all of IMDb's user reviews as of January 11th 2021. This dataset contains approximately 5.5 million reviews and is split into 6 different JSON files, each of which sized at about 1.3GB totalling about 7.78 GB. Kaggle's public API was also used to retrieve this dataset, which downloaded a single zip file with 2.9GB. The data contained in each row is the summary of the review and the detailed review, the user who posted it, the rating given by the user, the date of the review, a flag to indicate it contains spoilers and the number of people who found the review useful from the total number of votes.

III. Data Preparation

To ensure an effective use of the data collected, it was filtered, cleaned and transformed from the sources mentioned above.

To better comprehend the problem domain, an initial study on the structure of the information was conducted. In this brief analysis, we focused on the data's overall format and structure, with a careful analysis of each attribute available and its pattern.

A. Filtering data

The reviews dataset contained information relating to every entry on IMDb, which is a superset of the titles present in Netflix. As this data was not pertinent to the domain of our problem, we decided to filter it out. To perform this filtering, we used the title of the movie or show to match data between both datasets. This reduced the number of reviews from 5.5M (7.8GB) to approximately 900k (955MB).

A method to drop shows created before a certain year was implemented, as an option to reduce the number of reviews on our dataset, which may have an impact on later stages of the project.

B. Cleaning and Transformation of data

We cleaned each dataset individually, starting by the Netflix data. We started by looking for null values on the dataset. Attributes like *director*, *cast*, and *country* had plenty of null values, but those are used to indicate that those attributes were unknown. However, the amount of null values on the attributes *date_added*, *rating* and *duration* were of small size so we decided to analyse them further. Through that analysis we discovered the null values on the attribute *duration* were caused by faulty data, having its value on the *rating* attribute instead. To treat this case, we swapped the values on those rows and as the amount of rows was small, we searched manually on Netflix for the maturity rating and filled the correct value on those rows.

Afterwards, we analysed the rows which had the attribute *date_added* with null value. Through research on the movies and TV shows that expounded those values, we found out that those movies and TV shows were no longer available in Netflix, and thus appearing with the attribute *date_added* as null. We decided to keep these rows as they still contain pertinent data.

For the null values present on the *rating* attribute we filled with the correct values by manually searching the correct rating again as the amount of rows that presented the issue was small enough to handle it manually.

Other than null values, there was another issue that needed to be treated on the dataset. The attribute *country* which contains a list of countries in which the show was produced is a comma separated list in string format, however the string in majority of the rows started with a comma and a space, which we promptly removed to ease the handling of the list later on.

After cleaning the Netflix dataset, we move on to the IMDb review dataset. Searching for null values, we found 3 in the *review_summary* column which we decided to delete as they were not relevant and the source of the value was unknown. The second column that contained null values was the *rating*. These values come from a time where a score was not needed to publish a review so we let them stay as they are as it is valid data.

After the null values search was concluded we decided to change the *helpful* column to make it easier to work with later. This column was split into three new attributes: *helpful*, *unhelpful* and *total*. These features are, respectively, the amount of votes considering the review helpful, amount of votes considering the review unhelpful and total amount of votes.

IV. Pipeline

To streamline the collection of data described in the Data Collection section and the processing of data in the Data Preparation section, we created a pipeline. Its purpose is to collect all the necessary datasets, process them and then save them to relational storage at a later stage. This pipeline can be triggered with a single command by using a Makefile. A visual representation of this pipeline can be found at Fig. 1

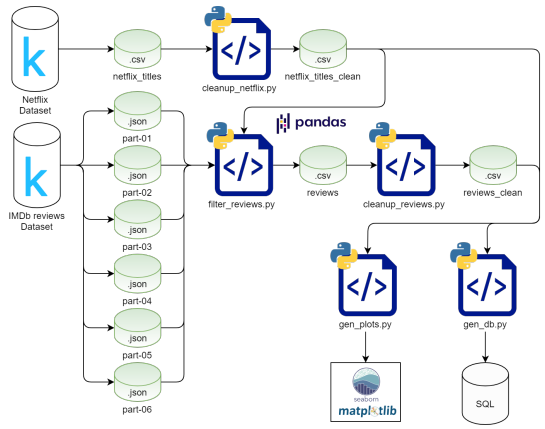

Fig. 1. Data Pipeline

V. Conceptual Model

After all the data processing steps a conceptual model is ready to be created. It can be found in Fig. 2.

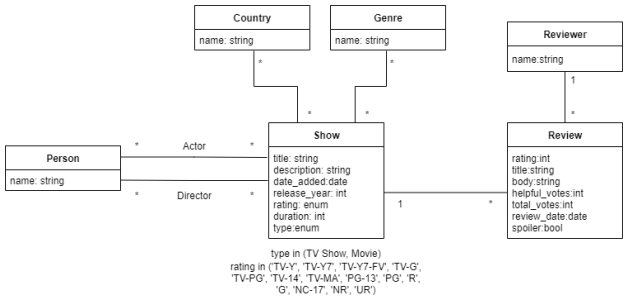

Fig. 2. Conceptual Model

The two main classes consist of *Show* and *Review*. *Show* refers to a Netflix show with its attributes brought from the clean dataset. It has associations with the *Person*, *Country* and *Genre* as these are repeated a lot though the entire dataset. The *Person* class represents either an actor or a director as there are examples of elements belonging to both at the same time. As for the *Review* class it associates with the *Reviewer* class which also has a lot of repeated elements across the dataset.

VI. Dataset Characterization

In order to better understand the data, we conducted an exploratory analysis by creating a set of plots and visual representations of the data. This helped us better understand certain aspects of the data that weren't apparent at first glance. As said before the cleaning, the reviews dataset contains about 900k reviews and the Netflix dataset contains 8807 unique shows. From those 8807 unique shows, 2676 are TV Shows and 6131 are movies.

A. Maturity Rating Distribution

To understand the maturity rating distribution of both Movies and TV Shows we plotted a bar plot that can be found on Fig. 3

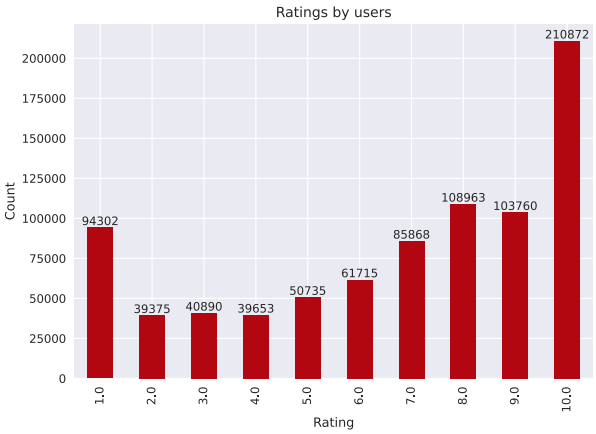

Fig. 3. Amount of shows by Maturity Rating

By analysing this plot it can be concluded that most of the shows present on the Netflix platform are catered to a more mature and adult demographic.

B. Duration Distribution

To analyze TV Show's duration we decided to again use a bar plot that can be found in Fig. 4. By analyzing it we arrive at the conclusion that most of the TV Shows released never get a second season. From the 2676 TV Shows on Netflix, only 883 have 2 seasons or more and only 17 shows have 10 seasons or more.

Regarding Movie durations, we decided a histogram was the best way to represent them because it's less discrete than the TV Show's duration. It can be found at Fig. 4. By analysing the plot, it can be asserted that most movies have durations in the 90 minutes to 110 minutes range.

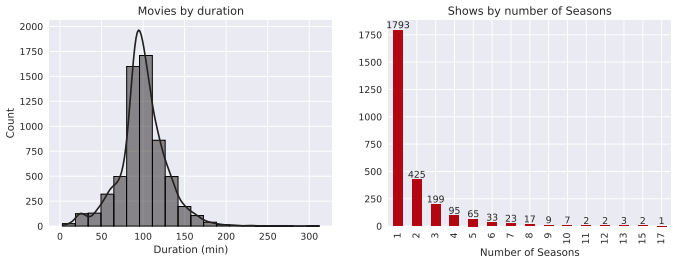

Fig. 4. Duration

C. Show scores

From the rating scores present in the reviews dataset it is trivial to calculate the score of each title. To analyze this feature we plotted a histogram that can be found in Fig. 5

By analyzing the plot we can conclude that most shows are seen in a positive light, since most show have a score between 6 and 8. On top of that scores between 0 and 4 seem less prevalent than scores between 8 and 10.

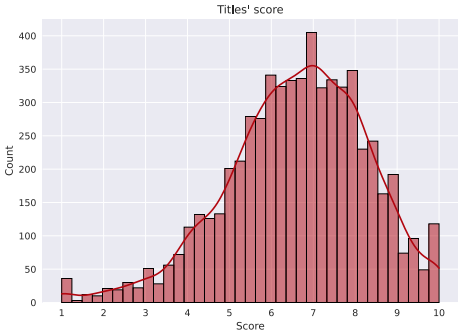

Fig. 5. Title score distribution

D. User Review Distribution

In a platform like IMDb there's always a small group of people who have the most presence on rating and reviewing the movies and TV shows that are released every year. Taking this in mind we decided to analyse the amount of reviews made by each reviewer. From the almost 479 thousand reviewers, on average each one made 8 reviews in its lifetime, however this statistic in itself doesn't accurately represent the distribution. By further analysing the distribution we found out that the third quartile is equal to 1, this is, 75% of the reviewers had only made a single review on their lifetime. By looking for outliers, the most accusatory value was the reviewer that presented over 1500 reviews, and, among others, caused the huge discrepancy on the distribution.

E. Review Length

To understand better how a review length is affected by other features in the dataset, two plots were created.

The first one can be found at Fig. 6 and presents the distribution of the review lengths across the dataset. From this plot we conclude that most reviews are of very short length. Even though this doesn't mean the review is of low quality, it probably means the review is a low effort one. There are , however, a considerable amount of reviews in the 200 to 400 words which represent higher effort reviews.

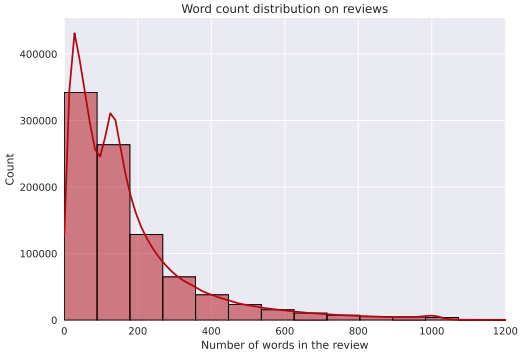

Fig. 6. Amount of Reviews by word count

The second plot is a box plot and can be found at Fig. 7 and presents the distribution of the review lengths based on its rating. This is an interesting plot that shows that extreme reviews (1 and 10) tend to have less words and therefore be of less effort than the more moderate ones (5 to 8).

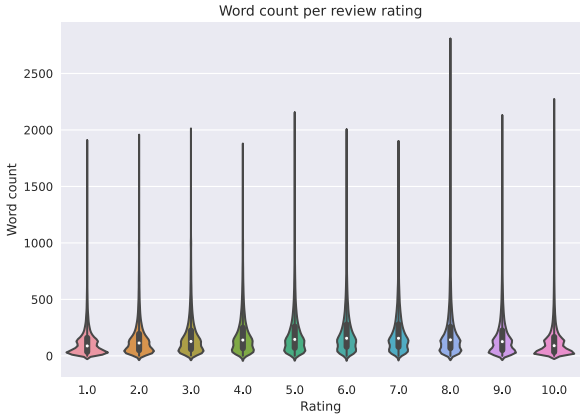

Fig. 7. Word Count by Review Rating

F. Countries

Studying the distribution of the countries of production of Netflix movies and TV shows that can be found in Fig. 8, we found, as expected, that United States lead with the most produced TV shows and movies. After that, the country with most produced TV shows varies from the one with most movies produced. United Kingdom, Japan and South Korea lead the TV show production after United States, with practically all the other countries having significantly lower amount of TV shows produced. Meanwhile on the movie production, India leads after United States followed by United Kingdom with approximately half the amount of movies produced by India. Like on the TV shows, the rest of the countries have significantly lower movies produced.

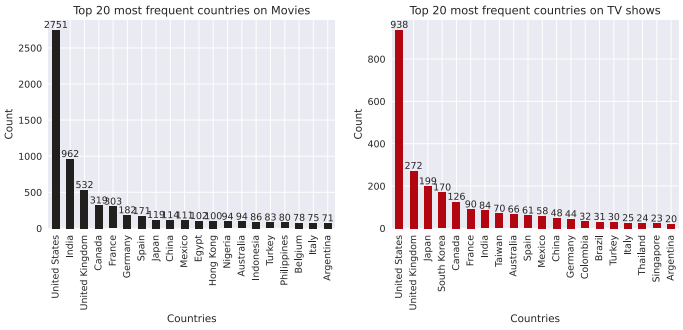

Fig. 8. Top 20 most frequent countries on Netflix titles

G. Genre

In search engines that operate through media like ours, it's also important to know how the genre of the movies and TV shows are distributed, as they can be used as keywords to search for content of similar type, as such we plotted the genres of TV shows and movies to analyse their distribution. From the plot in Fig. 9 we can see dramas and comedies are the most occurring type of movies and TV shows on Netflix. All other genres have significantly lower occurrence save for crime, kids targeted, romances and documentary type of TV shows that are relatively close to the most occurring.

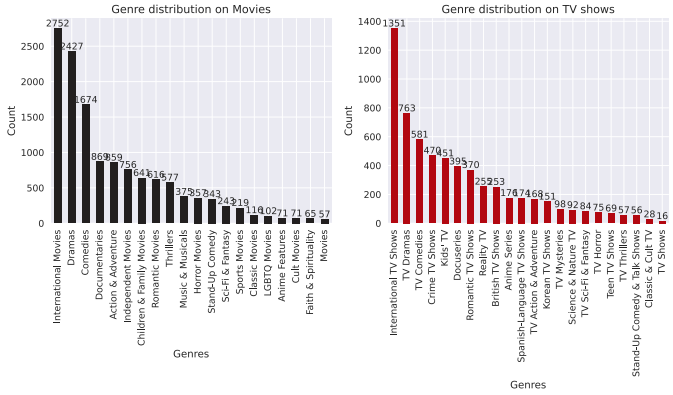

Fig. 9. Genre distribution on Netflix titles

VII. Search Tasks

Given the data available, the following example queries can be made to the database originated from the combined data:

- Search for highest rated titles of a specific actor or director
- Search for movies and/or TV shows where an actor is playing a certain role (e.g. actor X is a criminal in Tokyo)
- Search for movies that contain actors of certain nationality (e.g. Spanish actors movies)
- Search the top animation TV shows that involve different world
- Search by keywords that aren't explicit on movie genres or description but still describe the type of movie or characters in it (e.g. searching "Jotaro" should give me results such as "JoJo's Bizarre Adventure")

VIII. Conclusion

Throughout this report, it was explained how the datasets were cleaned, processed and analysed, as well as the conceptualisation of the database model. Through our analysis of the data, we were able to achieve a thorough understanding of the problem domain, allowing us to accurately utilise the data available to develop the search engine on later milestones. All the objectives for this milestone were concluded, the entirety of the actions performed on the dataset can be seen on the pipeline created. For future work in the next milestones, the conversion to a proper database will be performed as well as the start of the search engine.

REFERENCES

[1] Bansal, S., 2021. Netflix Movies and TV Shows. [online] Kaggle.com. Available at: <https://www.kaggle.com/shivamb/netflix-shows>[Accessed 13 November 2021].

[2] Biswas, E., 2021. IMDb Largest Review Dataset. [online] Kaggle.com. Available at: <https://www.kaggle.com/ebiswas/IMDb-review-dataset>[Accessed 13 November 2021].