Popular Movies and Streaming

Carlos Gomes
up201603404@fe.up.pt
Eduardo Silva
up201603135@fe.up.pt
Joana Silva
up201208979@fe.up.pt
Joana Ramos
up201605017@fe.up.pt

Information Description, Storage and Retrieval
Faculty of Engineering
University of Porto
Porto, Portugal

Abstract

1 Introduction

The framework here proposed is intended to enable users to search for movie related information gathered from various sources. A pipeline will be established to aggregate this data and match their content, as well as perform cleaning operations, to finally be able to explore it properly.

2 Obtaining and Preparing the Datasets

The data that will be used is gathered from 4 different datasets. Further details about them are presented below.

2.1 IMDb Official Dataset

The IMDb official dataset [1] is a **structured dataset** in TVS format that contains all of the information regarding movies and TV series that can be found on IMDb's website. The dataset is composed of 7 different files which together add up to a **volume** of 4.6 GB, a size that increases daily as the files are updated. All of the 7 files present a very similar **structure** but contain different relevant information:

- **title.akas.tsv.gz** file - contains important information concerning titles, such as a movie's region, title and language;
- **title.basics.tsv.gz** file - has relevant information such as the release year of a movie, its run-time in minutes, the genres associated with the film and its original title;
- **title.crew.tsv.gz** file - this file contains the details about a movie's writers and directors;
- **title.principals.tsv.gz** file - contains a movie's main cast and crew details, such as the category of the job being executed, the specific job title and the characters played in case of being an actor;
- **title.ratings.tsv.gz** file - has IMDb's votes and rating information for titles, containing the number of votes and the average rating;
- **name.basics.tsv.gz** file - significant information about each person, for instance the name for which the person is most credited, birth year and death year (when applicable), its top three professions and the titles for which is most known;
- **title.episode.tsv.gz** file - this file will not be used, as it consists of information related to TV series.

In what concerns the **licence** of the dataset, IMDb, as the **source authority**, provides a limited non-commercial use of their data if all of the stipulated conditions [2] are met. Additionally, a commercial use licence can also be obtained [3] , if the need arises.

2.2 IMDb Scraped Dataset

This dataset [4] is a **structured dataset** with movie information retrieved through scraping of IMBb's website [5]. The dataset comprises all the movies with more than 100 votes as of 01/01/2020, amounting 85855 movies. The information is represented in 4 .csv (Comma Separated Values) files in UTF-8 encoding, which total 230,0 MB. This dataset is obtained from Kaggle [6] under the CC0: Public Domain **license**. Since the **authority** of this data repository might be questionable, this dataset is complementary and it's correctness can be assessed. It contains the following information:

- **IMDb movies.csv** file - contains movie related information like the title, genre, language, budget, director and actors. This file contains the movie's IMDb ID;
- **IMDb names.csv** file - has information regarding people involved in movies (directors, writers, actors etc), namely their name, bio, birth date and other personal details;
- **IMDb ratings.csv** file - a collection of ratings for each movie, including demographic information (e.g. percentage of female voters);
- **IMDb title_principals.csv** file - associations between the movies and the names files

2.3 Streaming Dataset

The Streaming Dataset [7] contains information about 16744 movies from 4 different platforms: Netflix, PrimeVideo, Hulu and Disney+. The data was last updated on May 2020. This dataset is **structured** and extracted from Kaggle [6] under the **license** CC0: Public Domain. The principal characteristics of the dataset are:

- **Title (string)**: Title of the movie;
- **Year (int)**: Year when movie was released;
- **Netflix (int)**: 1 if the movie is available in Netflix, 0 otherwise;
- **Hulu (int)**: 1 if the movie is available in Hulu, 0 otherwise;
- **PrimeVideo (int)**: 1 if the movie is available in Prime Video, 0 otherwise;
- **Disney+ (int)**: 1 if the movie is available in Disney+, 0 otherwise;

However, due to inconsistencies in certain attributes, in relation to the other datasets, and the lack of an IMDb ID, only 15531 of the available titles are used in the following stages.

2.4 IMDb Movie Pages

The information regarding the synopses of the movies is directly scraped from the respective IMDb [5] pages. This data is composed of free text and is **unstructured** by nature, with the total volume being limited by that of the other datasets.

2.5 Data Pipeline Process

The data pipeline is illustrated in figure 2. The matching between all the datasets is done through a common IMDb IDs. All of the datasets posses this attribute except for the streaming one. In this case, a new column `imdb_id` is added with this value. The ID's are obtained through a match between titles, year of release and the IMDb classification (`short`, `movie`, `tvmovie`, `tvshort` or `video`). It is also worth noting that due to the smaller volume of this particular dataset, it serves as a bottleneck for the other 3.

After all the data is matched and aggregated, cleaning operations are performed, namely duplicate value removal and title and date normalization.

Finally, the data is then stored in SQL relational database, ready for further exploration.

3 Datasets Characterization

The number of movies per streaming platform is represented in figure 1. Clearly, Prime Video dominates the dataset, with Netflix coming in second while Hulu and Disney+ are more or less equivalent.

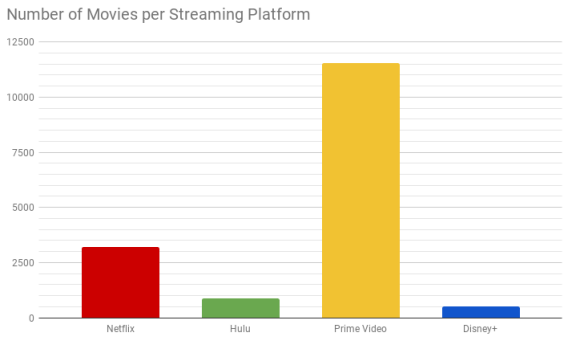

Figure 1: Number of Movies per Streaming Platform

Figure 3 illustrates the distribution of movies on streaming platforms, with relation to the movie's release year. It's interesting to notice that while Prime Video [8] has a bigger movie collection overall, Netflix [9] hosts more recently released movies. In spite of this, we should keep in mind that this is a sample, and might not be representative of the whole set.

The average movie rating by year can be seen in figure 4, with the ratings from IMDb [5] and Rotten Tomatoes [10] represented. We can verify that in Rotten Tomatoes the rating is almost always lower.

Regarding the textual information, the average word count of synopses and biographies can be seen in figure 5. It seems that biographies tend to be bigger than synopses.

4 Conceptual Model of the Domain

The conceptual model of the domain is illustrated in figure 6 and presents the **main entities** of the domain, which are movies and people involved in them, such as actors, directors, writers, among others.

A **movie** can have as attributes its title, release date, run-time in minutes and synopsis. Moreover, a movie can also be from a certain country and have many different languages associated with it, as well as ratings from different sources like IMDb [5], Rotten Tomatoes [10] and even Metacritic [11] (metascore). Furthermore, a movie can also be in different streaming platforms such as Hulu [12], Disney+ [13], Netflix [9] and PrimeVideo [8].

A **person** has as attributes its name, date of birth, gender and biography. In addition, a person can also have one or more job titles in a movie, being for example an actor, a writer or director, and, in the case of being an actor, the character or list of characters that it plays in the movie.

5 Data Retrieval Tasks

The focus of the data retrieval tasks are movies and the people involved in them. Some possible queries are:

1. Retrieve all movies in which an actor appears;
2. Retrieve all movies by director;
3. Retrieve high rated movies for each genre;
4. Retrieve high rated movies for each year;
5. Retrieve high rated movies for each language;
6. Retrieve movies that fit into a text description (e.g. "second world war movies");
7. Retrieve people that fit into a text description (e.g. "young puerto rican actor").

Some of these can already be performed in other tools or applications, however, currently, there is no single solution that offers all of the possibilities. For example, queries 1, 2, 3 and 4 can be performed in the IMDb official website [5], but with only their own rating being considered, and not those of other platforms. For query 6, similar results can be obtained in WhatIsMyMovie [14] through the use of Machine Learning.

References

[1] IMDb Official Dataset
https://datasets.imdbws.com.
Accessed in October, 2020.

[2] IMDb Software Integration Help Page
https://help.imdb.com/article/imdb/general-information/can-i-use-imdb-data-in-my-software/G5JTRESSHJBBHTGX#
Accessed in October, 2020.

[3] IMDb Developer
https://developer.imdb.com/?ref_=helpms_ih_gi_developer
Accessed in October, 2020.

[4] Stefano Leone
IMDb Scraped Dataset
https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset
Accessed in October, 2020.

[5] IMDb's Official Website
https://www.imdb.com/.
Accessed in October, 2020.

[6] Kaggle
www.kaggle.com
Accessed in October, 2020.

[7] Ruchi Bhatia
Streaming Dataset
https://www.kaggle.com/ruchi798/movies-on-netflix-prime-video-hulu-and-disney
Accessed in October, 2020.

[8] Prime Video
https://www.primevideo.com/
Accessed in October, 2020.

[9] Netflix
https://www.netflix.com/
Accessed in October, 2020.

[10] Rotten Tomatoes
https://www.rottentomatoes.com/
Accessed in October, 2020.

[11] Metacritic
https://www.metacritic.com/
Accessed in October, 2020.

[12] Hulu
https://www.hulu.com
Accessed in October, 2020.

[13] Disney+
https://www.disneyplus.com
Accessed in October, 2020.

[14] What is My Movie
https://www.whatismymovie.com/
Accessed in October, 2020.

6 Annex

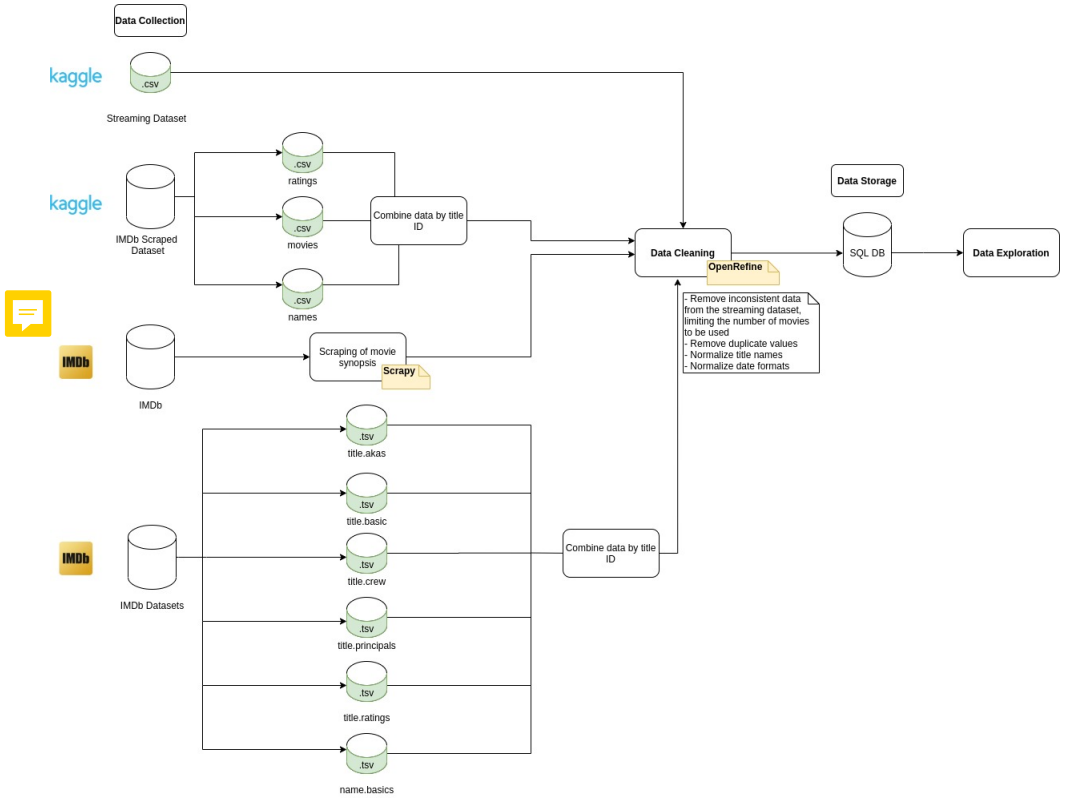

Figure 2: Data Pipeline Diagram

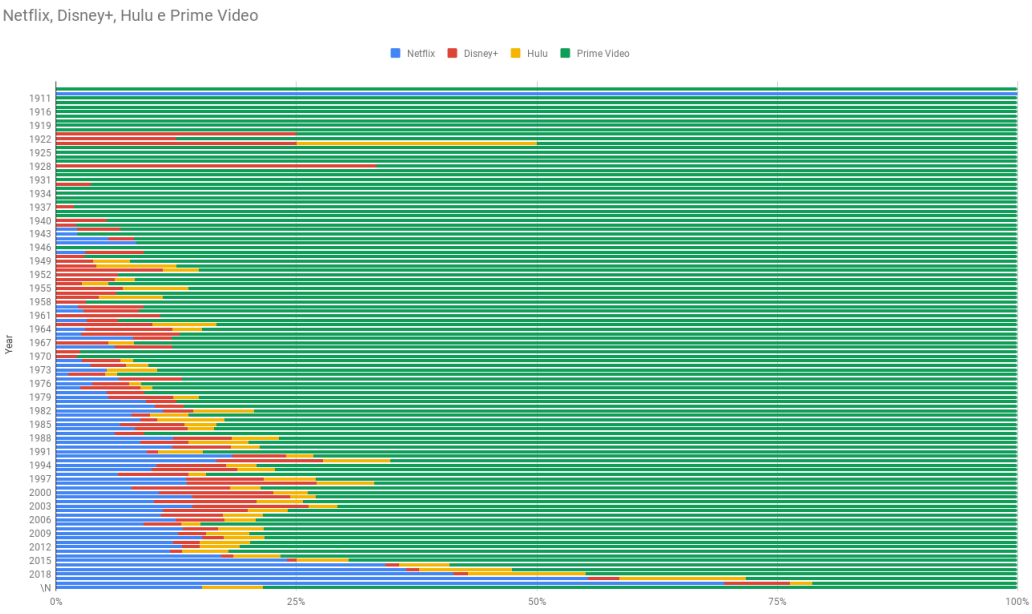

Figure 3: Distribution of movies on streaming platforms

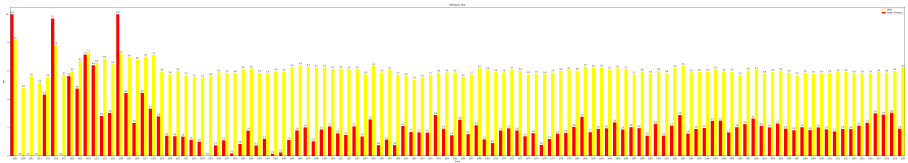

Figure 4: Average movie rating by year

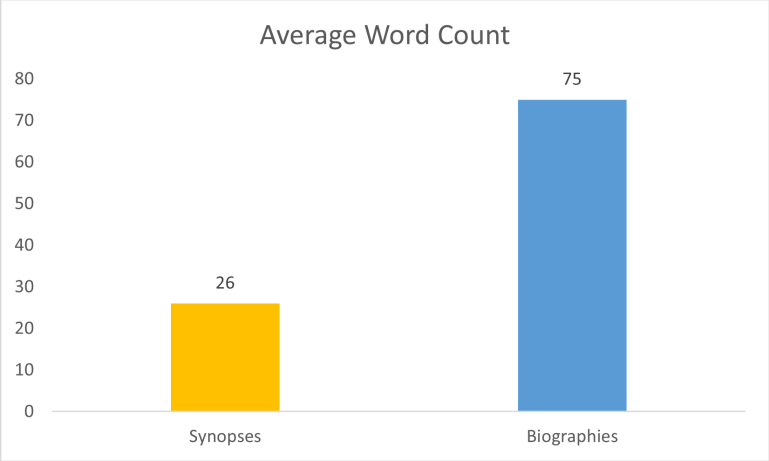

Figure 5: Average word count of synopses and biographies

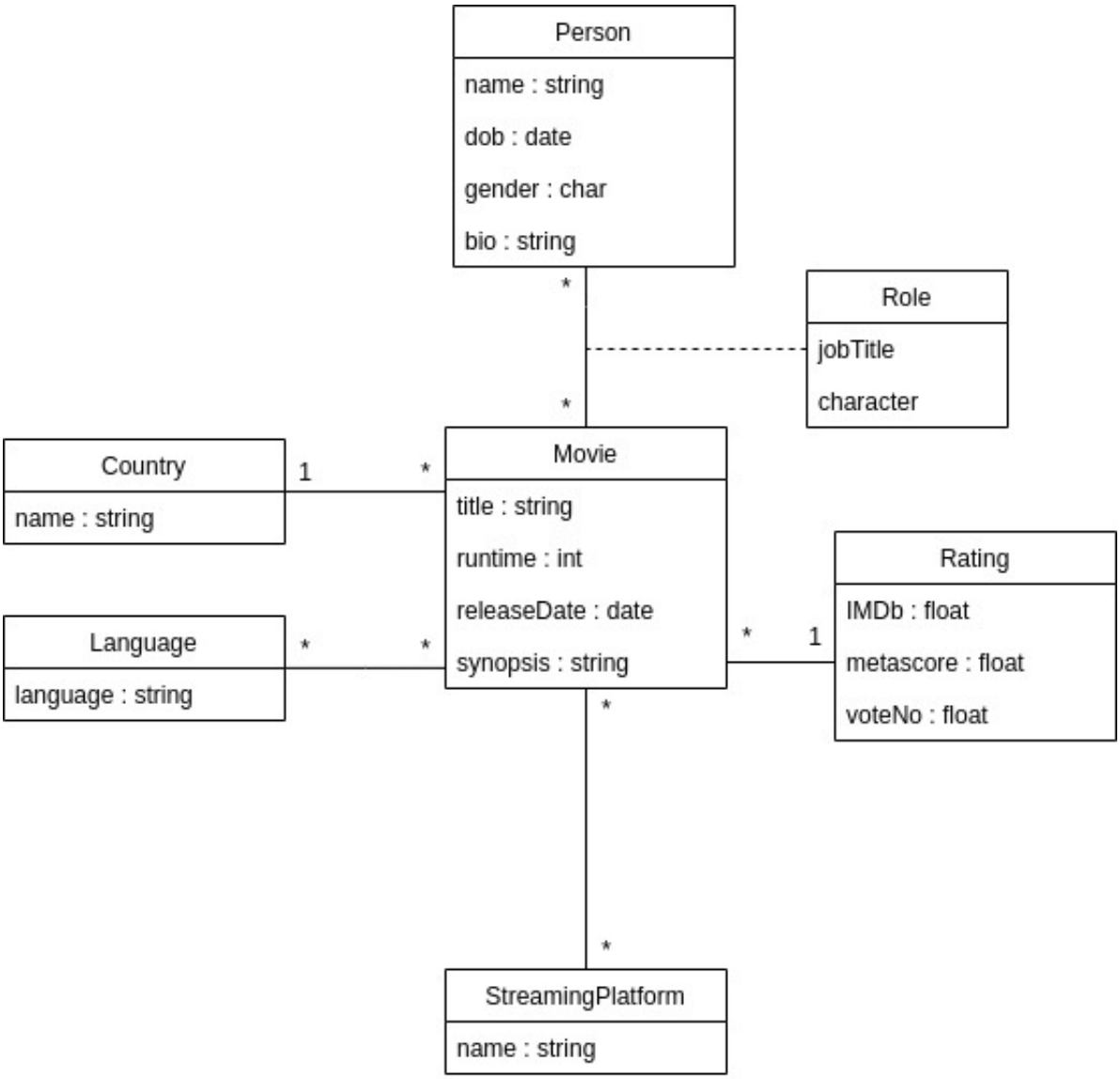

Figure 6: Conceptual Model of the Domain