PRI - Information Processing and Retrieval, 2022/2023

Bernardo Ferreira
up201806581@up.pt

Pedro Pereira
up201905508@up.pt

Telmo Botelho
up201806821@up.pt

***Abstract*—This paper describes the work that was done in the first milestone of the group project for the Information Processing and Retrieval course.**

1. Introduction

In this paper we discuss the work done during the first milestone of the Information Processing and Retrieval course. The objective of this milestone was to collect, prepare and characterize the data that we choose to work with. First, we talk about the thematic of our data. Next, we talk about how we prepared, cleaned and joined that data. Finally, we characterize our information with several graphs. We also show and explain the conceptual data model.

mds
october 08, 2022

2. Data Preparation

In this first milestone we focus on data preparation and characterization. In this first phase we first choose the thematic and the data sets that we wanted to use. We search those data sets from the kaggle website. We focused on data sets with a lot of information and rich in text. After that we started to explore the data to understand more clearly what we had. After that we proceed to clean and join all the data sets that we have to end up with the final data set. All of this is shown more clearly in the following pipeline.

2.1. Pipeline

After a thorough analysis we designed our final data pipeline.

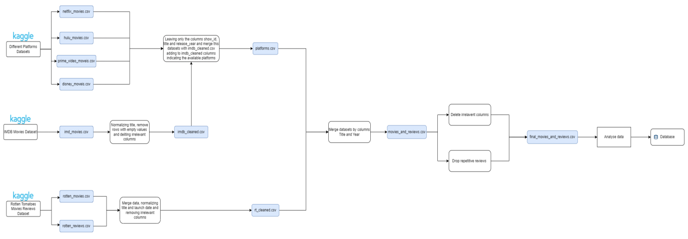
kaggle
Different Platforms Datasets
netflix_movies.csv
hulu_movies.csv
prime_video_moveis.csv
disney_moveis.csv
Leaving only the columns show_id, title and release_year and merge this datasets with imdb_cleaned.csv adding to imdb_cleaned columns indicating the available platforms
platforms.csv
kaggle
IMDB Movies Dataset
imd_movies.csv
Normalizing title, remove rows with empty values and deliting irrelevant columns
imdb_cleaned.csv
kaggle
Rotten Tomatoes Movies Reviews Dataset
rotten_movies.csv
rotten_reviews.csv
Merge data, normalizing title and launch date and removing irrelevant columns
rt_cleaned.csv
Merge datasets by columns Title and Year
movies_and_reviews.csv
Delete irrelavent columns
Drop repetitive reviews
final_movies_and_reviews.csv
Analyse data
Database

Figure 1. Pipeline

We decided that we should merge all different platforms' data sets leaving only the important columns, and afterwards merge it with a clean and normalized IMBD movies data set. The normalization we applied consists in removing spaces, putting all characters on lower case and removing strange characters (we found this to be extremly efficient at corresponding movies between different data sets with different origins).

This results in a data set with all known IMDB movies, each one with columns indicating if the given movie is available in each platform.

We also chose to merge the rotten-movies data set with the rotten-reviews data set, normalizing the movie title and launch date columns and removing other irrelevant columns. This results in a new ”rt-cleaned.csv” that can be merged by the movies title and launch date with the ”platforms.csv”.

Subsequently we merged the two resulting data sets into one ”movies-and-reviews.csv” by matching the normalized title and launch date columns.

Finally the resulting ”movies-and-reviews.csv” is cleaned by deleting irrelevant columns to our project and dropping some repetitive reviews that we found after reviewing the data. After all this transformations we came up with the final data set ”final-movies-and-reviews” that can now be analysed and turned into a database for our next milestone.

3. Data Characterization

In this section we characterize our final dataset that we obtain during the preparation phase.

All the data sets we extracted from kaggle are decently sized, but the IMDB movies and reviews data set is the biggest one, containing all IMDB movies with over 100 reviews until 2020. This results on a final data set with over 270 000 lines.

Here we can understand that the average vote is a 6 on a scale form 0 to 10, showing that IMDB's rate is quite harsh being more common a movie with a rating of 2 than a movie with a rating of 9. After analysing the number of IMDB reviews we thought it would be interesting to compare the ratings given by the IMDB versus the ratings given by Rotten Tomatoes.

As we can see, the Rotten Tomatoes rating are much more permissive than the IMDB ratings, being 85 percent the most given rating. It's also very apparent that this ratings are much more distributed.

This graph demonstrates the exponential growth of movie making. The major part of the movies present on

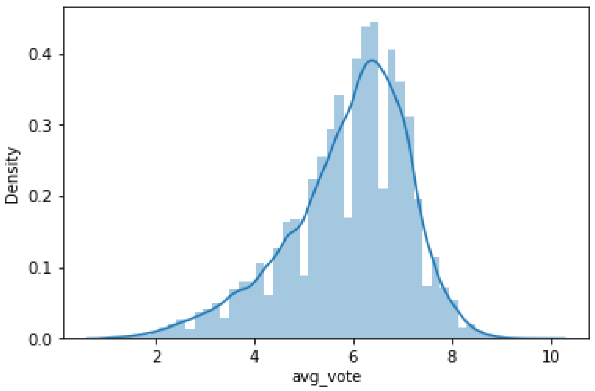

Figure 2. IMDB Rating

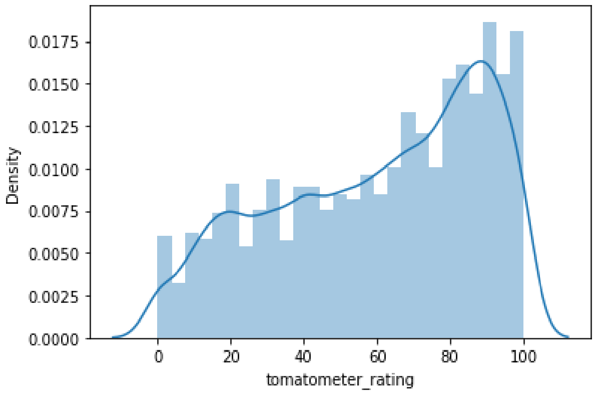

Figure 3. Rating

our database are from 2000 and afterwards so there's no need to remove old and irrelevant movies that might have been useless.

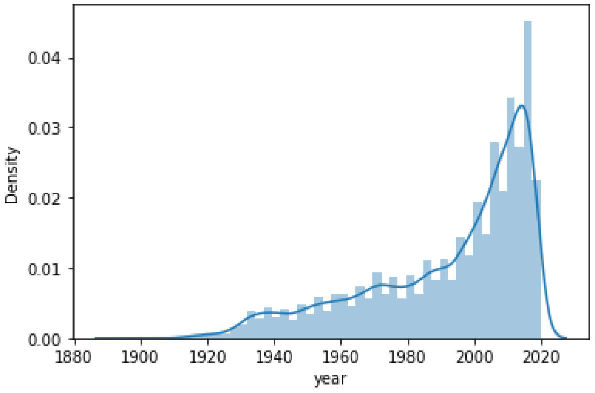

Figure 4. Density

Drama is, by a large margin, the most present movie type on our data set, followed by action and then comedy that is pretty much tied up with action.

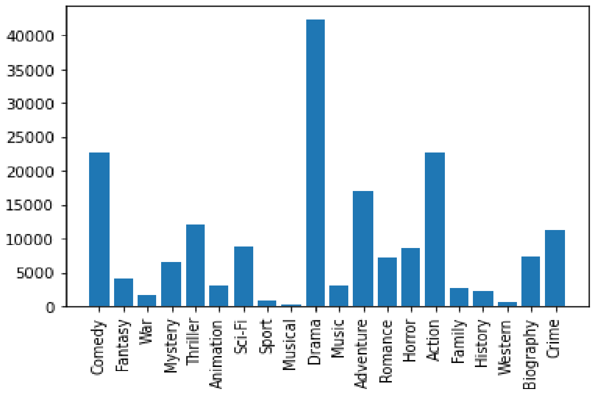

Figure 5. Geners

Netflix and Amazon Prime Video have the larger share of movies present on our data set. Nonetheless, Disney Plus and Hulu have a share big enough that shouldn't be discarded or ignored.

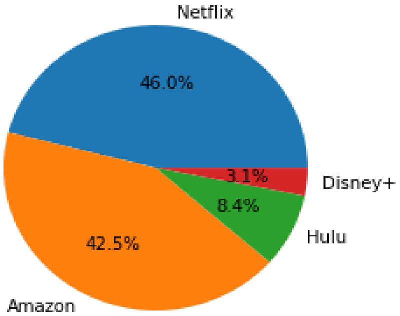

Figure 6. Platforms

Here we can find the most common word usages on the reviews from IMDB and Rotten Tomatoes, respectively, given to the movies present in our data set.

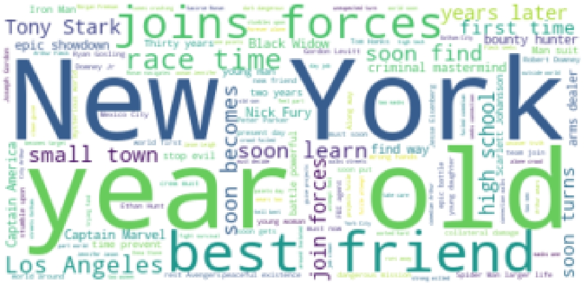

Figure 7. Reviews

4. Conceptual Data Model

After we prepare the data that we collected, we design the following conceptual data model.

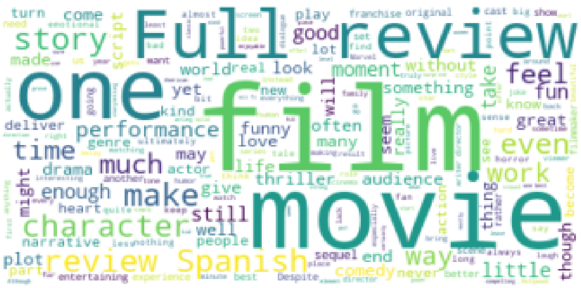

Figure 8. Words

As a main class we have the Movie class with the following attributes:

- **original_title**: This is the title of the movie.
- **year**: This is the year the movie was release in the theaters.
- **description**: This is the imdb movie's description.
- **avg_vote**: This is the average classification
- **votes**: This is the number of votes a movie had in the imdb.
- **movie_info** This is information about the movie.
- **available_hulu**: This tells if the movie is current available on hulu streaming service.
- **available_netflix**: This tells if the movie is current available on netflix streaming service.
- **available_amazon**: This tells if the movie is current available on amazon streaming service.
- **available_disney**: This tells if the movie is current available on disney streaming service.

Then each movie has a Review text that corresponds to all the reviews that people made relative to that movie. This class has the following attributes.

- **description**: This attribute corresponds to what was written about that movie.

Also, each movie has a review that corresponds to its classification in imdb and rotten tomatoes websites. This class has:

- **tomatometer_count**: This attribute is the number of classifications that was attributed to the movie.
- **tomatometer_rating**: This is the average of the classifications given by experts to the movie.
- **audience_rating**: This is the average of the classifications given by the audience to the movie.

Finally, the movie also has genres, actors and directors. This three classes have the same attribute:

- **name**: This attribute can be the name of the gener, actor or director.

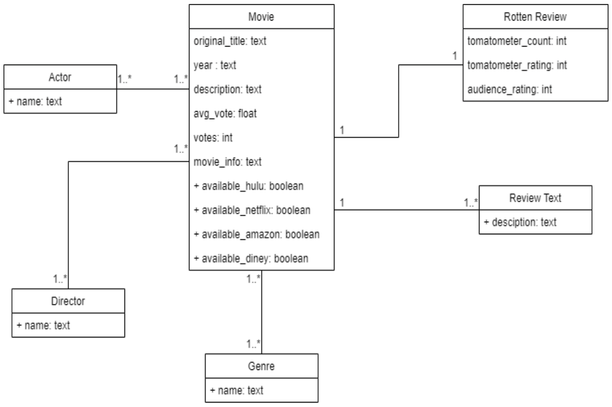

Figure 9. Conceptual Data Model