Information Processing and Retrieval Project

Beatriz Aguiar, João Marinho, Margarida Vieira*

Abstract

This report explores the processing of the creation of a search engine. It is divided into three main sections, each representing a project's milestone. Every procedure is carefully and minutely described. Data preparation is the initial and fundamental phase. Its focus is on collecting data, processing it, and conducting an exploratory analysis for further understanding of the convenient search tasks.

Keywords: datasets, data preparation, data analysis, information retrieval, search system

ACM Reference Format:
Beatriz Aguiar, João Marinho, Margarida Vieira. 2018. Information Processing and Retrieval Project. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

Introduction

Searching for information on the Web is, in the modern era, of utter importance. ICT (Information and Communications Technology) is a broad subject and the concepts are evolving. It covers any product that will store, retrieve, manipulate, transmit, or receive information electronically in a digital form (e.g., personal computers including smartphones, digital television, email, or robots). [1] Therefore, improving and optimizing the process of finding information is one of the primary focuses of those who work in the field of Data Processing & Information Retrieval.

We were proposed to develop a project to solidify the knowledge acquired throughout the semester. The end result is an information search system, which entails data collection and preparation processes, information querying and retrieval, and retrieval evaluation.

This report covers thoroughly the three milestones of the project, namely: data preparation, information retrieval, and final search system.

1 Data Preparation

The first milestone is achieved with the selection, preparation, and characterization of the datasets. The goal of the first task is then to prepare and explore these datasets, which may suffer from extraction actions such as crawling or scraping.

1.1 Data Selection

The provided guidelines shepherd the pursuit of rich and ambiguous datasets that granted both information quantity (in the thousands range) and quality. Thus, movies came across as the ideal theme of choice.

Google Dataset Search was our first approach to the search for datasets, however, after a generic analysis, we considered the available data to be quite simple, containing only one table, while lacking textual features.

Given IMDb's high demand and popularity, it seemed to be the perfect dataset source. Its API did not only provide 7 TSV files, but it also granted reliable and updated content, with each table containing different media-related data up to millions of entries. Such information covered most of the common movie attributes (duration, genre, etc.), titles, directors, actors, and ratings.

Subsets of IMDb data are available for access to customers for personal and non-commercial use. You can hold local copies of this data while being subject to their terms and conditions. Information courtesy of IMDb (https://www.imdb.com). Used with permission.

1.2 Data Processing

As demonstrated in figure 1, the original datasets underwent through a thorough processing, following an ETL (extract-transform-load) pattern. To collect the data, the makefile deliverable starts by running a series of commands that download the zip files from the IMDb API and unzip the latter to the data folder.

An initial exploratory analysis was conducted on each dataset to understand its relevance and features importance. The *title_akas.tsv* dataset, for e.g., turned out to be uninteresting as the only feature we were interested in, the language of the movie, had 6 not-null values in a 33M entries dataset. On top of that, we concluded that the datasets regarding the principal crew members and episodes of the streaming content were also irrelevant to our search engine.

In terms of data preparation, we started by filtering the dataset *title_basics.tsv*, leaving only entries corresponding to movies released after 1990. We opted to do this not only for efficiency purposes, by reducing the total number of entries, but also due to the minimal value it added to the dataset - fewer individuals will likely search for older data.

*Beatriz Aguiar, up201906230@up.pt; João Marinho: up201905952@up.pt; Margarida Vieira, up201907907@up.pt.

Conference'17, July 2017, Washington, DC, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06...$15.00
https://doi.org/XXXXXXX.XXXXXXX

However, the dataset still comprised hundreds of thousands of entries, which was far more data than necessary and would significantly increase web-scraping latency. For this reason, we decided to discard movies with average ratings below 7, always aiming to keep the most relevant data possible. The final dataset now contained around 70 thousand entries, a considerably good dataset dimension.

Posteriorly, we moved on to merge the other datasets with the main movie dataset.

At this point, our dataset's features consisted in an id - *tconst*, the original movie name - *originalTitle*, a boolean indicating whether it is an adult movie or not - *isAdult*, the release year - *startYear*, the duration of the movie - *runtimeMinutes*, its genres, the average rating, the number of votes, and, finally, the directors, a string of ids referring to the crew's dataset.

The next step was to match the directors' string of ids to the corresponding real names and convert it to a string of names.

For the most important movie's piece of information, the synopsis, we performed web-scraping on both Wikipedia and IMDb's pages, using Python's *BeautifulSoup* package. This textual feature did not, however, suffer from any processing other than cleaning up irrelevant characters. Additional processing will take place in the next milestone.

Finally, we chose to standardize the features' names for aesthetics and whimsy purposes, as well as to enhance the posterior visual analysis carried out - *startYear* changed to *year*, *runtimeMinutes* to *duration*, and *averageRating* to *avgRating*.

1.3 Characterization

To better understand the data in our hands, we developed a Python Notebook to plot general information about the dataset. The more information we have, the better the decisions will be made throughout the development of the project.

By taking a look at the obtained graphics, some conclusions can be easily drawn.

First, the number of movies released per year (figure 2) has increased in the last century, meaning that more information can be drawn from more recent years. This increase is probably due to the latest technological advances.

The second plot (figure 3) shows that only a small minority of films were able to get ratings above 8, with most films scoring between 7 and 8. However, it is important to keep in mind that the final dataset did not include any films with an average rating of less than 7.

When it comes to genres (figure 4), Documentary and Drama seem to be the most popular ones by a significant margin, with Comedy, Biography, and Romance completing the top 5.

Since the average rating for practically every genre falls within the same range, no meaningful conclusions can be drawn about the average rating per genre (figure 5).

The following plot, *Rating per number of votes* (figure 6), demonstrates that just a minor number of individuals give a rating above 8, leading us to believe that ratings above this threshold are simply the consequence of fewer votes when compared to ratings between 7 and 8, which usually receive the majority of votes.

We also decided to plot the number of adult vs. non-adult movies (figure 7), although no meaningful conclusions can be drawn from the said graph.

Finally, we plotted a WordCloud for 3 of the top 5 genres - Documentary, figure 8, Biography, figure 9, and Romance, figure 10.

A further step towards information extraction was conducted, Named Entity Recognition (NER), in the newest feature attained, the synopsis. This process mainly tries to locate and classify entities such as characters' names and/or specific places. Figure 11 is an example of this, applied to a biography synopsis.

In addition to NER, it was also performed a Keyword extraction analysis. For the previous NER example, the resulting keyword extraction list was:

1. 'drama follows manoel oliveira'
2. 'follows manoel oliveira life'
3. 'oliveira life times dictatorship'
4. 'life times dictatorship portugal'
5. 'manoel oliveira life'
6. 'dictatorship portugal'

1.4 Prospective Search Tasks

- Search a movie by its title, rating, year, genre, and director(s)
- Search for the most popular movies
- Search for a movie synopsis
- Search for recently released movies
- Search for +18 movies
- Search for the movies directed by a specific director

References

[1] Wikipedia. 2022. Information and communications technology — Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/w/index.php?title=Information%20and%20communications%20technology&oldid=1113616939. [Online; accessed 06-October-2022].

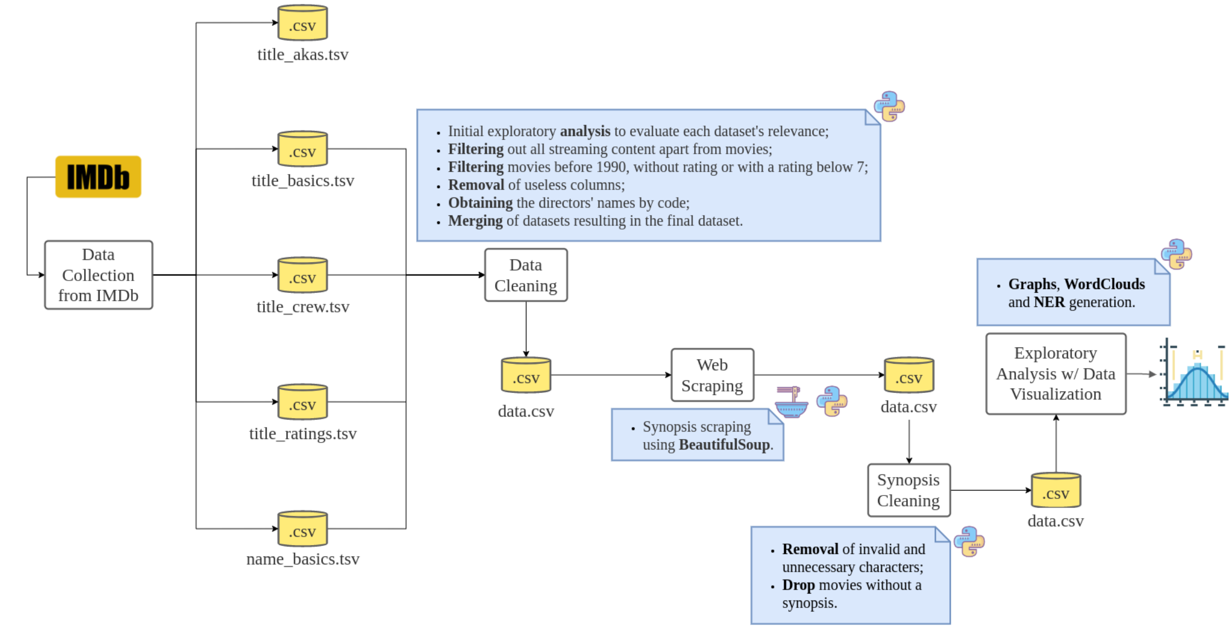
IMDb
Data Collection from IMDb
.csv
title_akas.tsv
.csv
title_basics.tsv
.csv
title_crew.tsv
.csv
title_ratings.tsv
.csv
name_basics.tsv
Initial exploratory analysis to evaluate each dataset's relevance;
Filtering out all streaming content apart from movies;
Filtering movies before 1990, without rating or with a rating below 7;
Removal of useless columns;
Obtaining the directors' names by code;
Merging of datasets resulting in the final dataset.
Data Cleaning
.csv
data.csv
Web Scraping
Synopsis scraping using BeautifulSoup.
.csv
data.csv
Synopsis Cleaning
Removal of invalid and unnecessary characters;
Drop movies without a synopsis.
.csv
data.csv
Exploratory Analysis w/ Data Visualization
Graphs, WordClouds and NER generation.

Figure 1. Pipeline

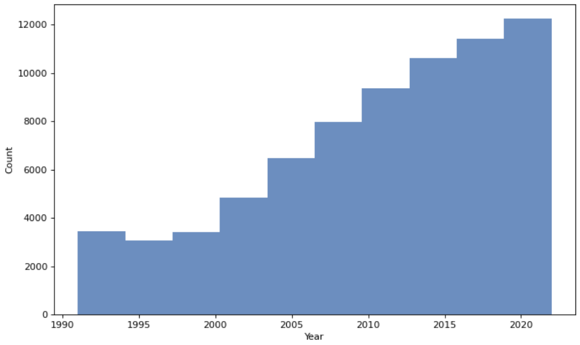

Figure 2. Number of movies per year

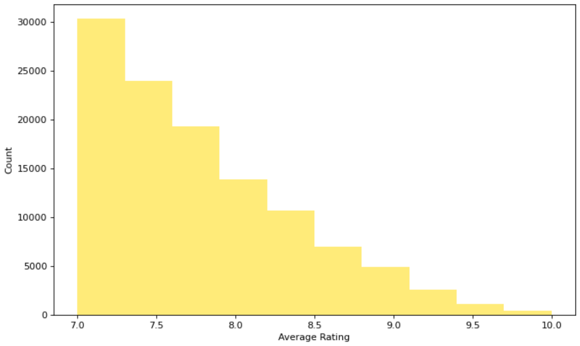

Figure 3. Number of movies per rating

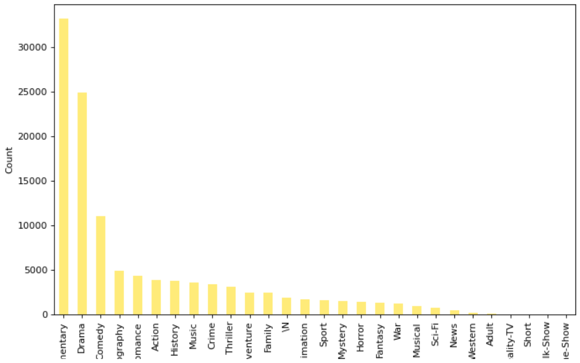

Figure 4. Number of movies per genre

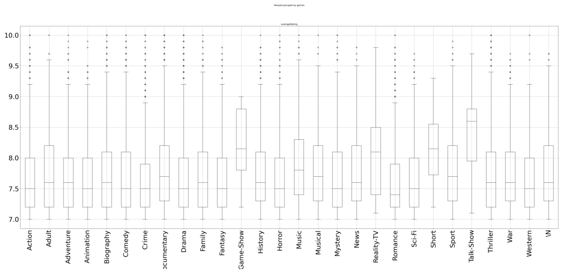

Figure 5. Average rating per genre

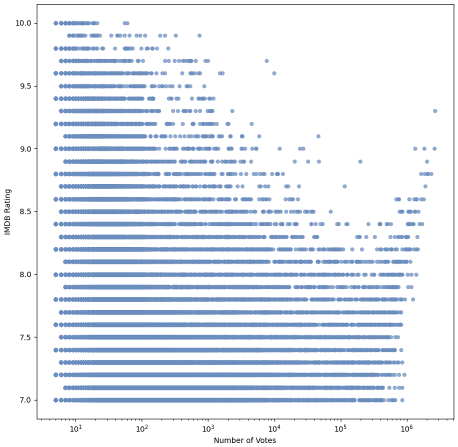

Figure 6. Rating per number of votes

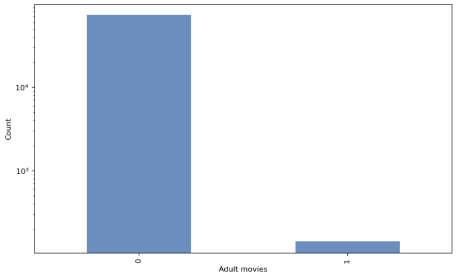

Figure 7. Number of movies per adultness

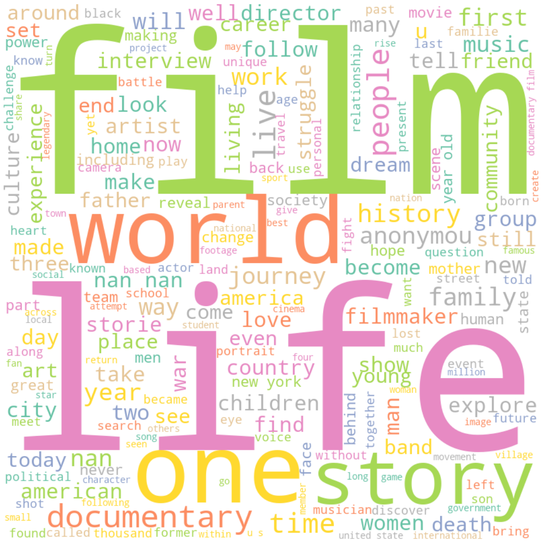

Figure 8. Documentary movies' synopsis wordcloud

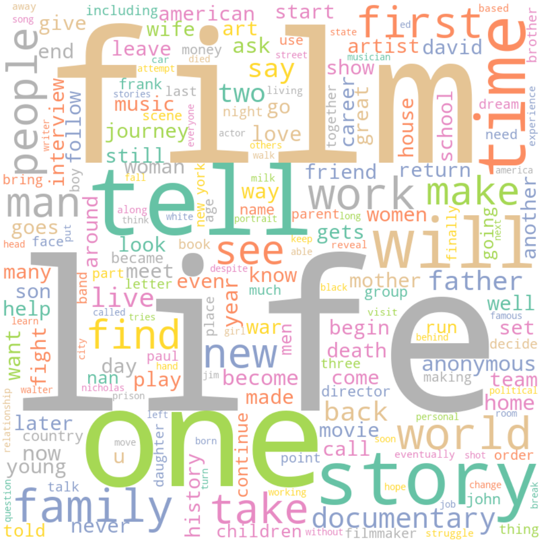

Figure 9. Biography movies' synopsis wordcloud

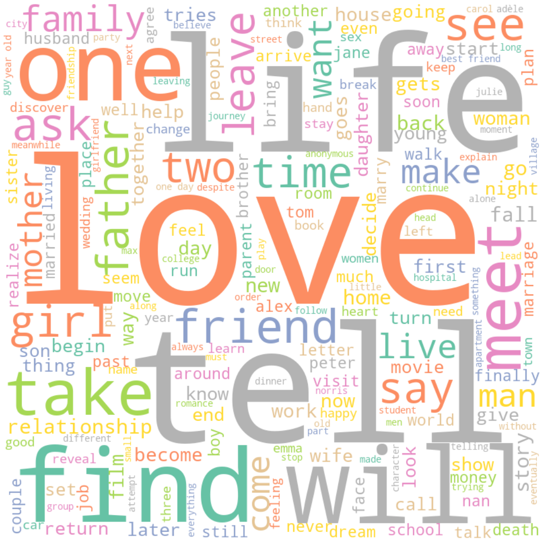

Figure 10. Romance movies' synopsis wordcloud

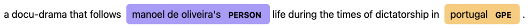
a docu-drama that follows manoel de oliveira's PERSON life during the times of dictatorship in portugal GPE .

Figure 11. Example of a NER biography synopsis.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009