PRI T1G4 Project Report

FEUPFlix

Nuno Alves  
 Faculdade de Engenharia  
 Universidade do Porto  
 up202302778@fe.up.pt

FirstName Surname  
 Department Name  
 Institution/University Name  
 City State Country  
 [email@email.com](mailto:email@email.com)

FirstName Surname  
 Department Name  
 Institution/University Name  
 City State Country  
 email@email.com

FirstName Surname  
 Department Name  
 Institution/University Name  
 City State Country  
 [email@email.com](mailto:email@email.com)

ABSTRACT

In today's data-rich landscape, effective data processing and retrieval are paramount. This report documents the creation of a robust database and information processing and retrieval tool within the context of the 2023/24 Project for Information Processing and Retrieval course. Our project focuses on harnessing data related to movies to build a free search engine that assists users in selecting movies based on textual queries. We emphasize data quality, source reliability, and efficient retrieval mechanisms. This endeavor underscores the increasing importance of data management and information retrieval in modern society, particularly in the context of entertainment and movie selection.

**KEYWORDS**

*Data Processing, Data Retrieval, Information Retrieval Tool, Movie Search Engine, Project for Information Processing and Retrieval.*

1 INTRODUCTION

Our project is designed to address the challenges and opportunities presented by the vast landscape of movie-related data. We have selected the theme of movies due to the extensive historical data available in this domain, spanning decades of cinematic history. The movie industry continually accumulates data, making it a valuable area for data exploration and retrieval.

2 DATASET

2.1 Dataset Choice

To lay the foundation for our project, we embarked on a quest to find a suitable dataset that aligns with our research objectives. Initially, we explored movie datasets, which contained a wealth of information, including details on personnel, ratings, reviews, and votes. However, we encountered challenges with textual data, such as missing movie synopses or runtimes.

Ultimately, we opted for a Kaggle dataset [<https://www.kaggle.com/datasets/utsh0dey/25k-movie-dataset>], which, while smaller in scale (approximately 25000 movies), offered a much more complete data content. Although it still had some missing important fields, we decided to improve the quality of the dataset by merging it with additional data sources.

In order to enhance the dataset, we leveraged an API from [<https://developer.themoviedb.org/>] to complete missing fields or correct malformed ones, thereby adding strength to the dataset.

2.2 Dataset Content

The original dataset that we've chosen for our project is composed of 12 columns, each containing valuable information crucial for both our research and the development of our movie search engine. These columns are meticulously curated as follows:

1. **Movie Title:** The title of the movie.
2. **Total Run Time:** The duration of the movie.
3. **Movie Rating:** The assigned rating for the movie.
4. **User Rating:** Ratings contributed by users for the movie.
5. **Genres:** The genres associated with the movie.
6. **Overview:** A concise summary of the movie.
7. **Movie's Plot Keywords:** Keywords pertaining to the movie's plot or theme.
8. **Director Name:** The name of the movie's director.
9. **Top 5 Cast Members:** The names of the top five cast members.
10. **Writer Name:** The name of the movie's writer.
11. **Releasing Year:** The year in which the movie was released.
12. **IMDb Movie URL Path:** The URL path leading to the IMDb page dedicated to the movie.

Following our meticulous data processing pipeline and the integration of data from "The Movie Database" (TMDb) API, we have refined our dataset to include the following essential columns:

1. **Movie Title:** This column remains a fundamental component of our dataset, serving as a key identifier for each movie.
2. **Genres:** We have enriched our dataset with genre information, allowing users to explore movies based on their genre preferences.
3. **Overview:** A concise movie summary or overview, which provides users with a quick glimpse into the movie's content.
4. **Keywords:** Keywords related to the movie's plot or theme, facilitating more nuanced and context-aware searches.
5. **Director:** The name of the movie's director, an important factor for those who appreciate specific directors' work.
6. **Top 5 Casts:** We have retained information about the top five cast members, helping users identify movies featuring their favorite actors.
7. **Writer:** The name of the movie's writer, which can be crucial for those interested in particular screenwriters' contributions.
8. **Path:** The path to the IMDb page for the movie, ensuring easy access to additional information and user reviews.

These refined columns not only provide a more comprehensive dataset for our movie search engine but also enhance the user experience by enabling more refined and precise queries. Our dataset's enrichment align seamlessly with our project's goal of assisting users in selecting movies based on their textual preferences.

2.3 Dataset Quality and Source

Our dataset, originally obtained from IMDb.com, is now available on Kaggle. IMDb.com is a widely recognized and reputable platform for movie-related information, ensuring the reliability of our data source. The dataset's information was scraped directly from the IMDb public website, adding credibility to its content.

To enhance the dataset further, we have integrated the TMDB API [<https://developer.themoviedb.org/>], which allows us to improve data completeness and accuracy. Despite its origin as a Kaggle dataset, the core data's source remains IMDb.com, known for its trustworthiness in the realm of movie-related data.

Throughout our analysis, we found that the dataset met our criteria, containing the necessary information without any unexpected values or formats. This combination of a trusted source and quality assurance measures reaffirms the dataset's reliability and suitability for our research and movie search engine development.

3 PIPELINE

Our Data Preparation Pipeline is entirely constructed using Python scripts, where the pandas library played a crucial role in data management and manipulation. We harnessed the power of the requests library to facilitate seamless communication between our Python scripts and the API, resulting in the creation of simple yet robust scripts for data cleaning and organization.

Our primary goal was to ensure the data's cleanliness and structure, ultimately formatting it into a CSV file for easy accessibility and further analysis

3.1 Data Refinement

In our Data Refinement process, we followed a systematic approach within our pipeline:

1. **Handling Malformed Data:** The initial step involved the identification and removal of columns containing malformed or incomplete data. To address this, we leveraged the API to correct the data, ensuring its accuracy and completeness.
2. **Eliminating Table Entries with Null Values:** Subsequently, we conducted a thorough examination of the dataset, removing any rows or entries with null values. This step helped us ensure data consistency and completeness.
3. **Eliminating Duplicate Movie Titles:** We also identified and removed movies with duplicated titles, ensuring that our dataset only retained unique and distinct entries.

By implementing these refinements, we enhanced the dataset's quality, removing irrelevant or erroneous information and streamlining it for our research and movie search engine development.

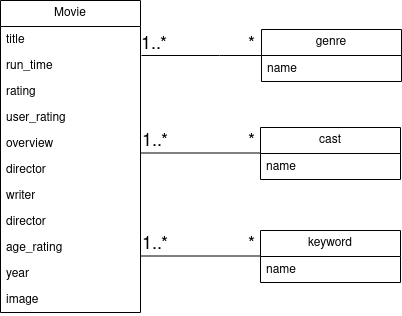
Top of Form

3.2 Data Analysis

Dummy text

Bottom of Form

3.3 Database



**Figure 3: Database Model**

In o The final step in our data refinement process involved creating a structured CSV file containing all the refined data. To accomplish this, we utilized a dedicated section of the Python script specifically designed for this task. The script was carefully crafted to include only the columns that met our research objectives and had undergone the necessary data refinement steps.

The Python script efficiently combined all these refined data elements and generated a CSV file, providing us with a clean and organized dataset in a structured format. This CSV file serves as a reliable foundation for our research and movie search engine development, enabling us to perform data analysis and deliver accurate and valuable results.

3.4 Pipeline Conclusion

Dummy text

Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00