/01

PRI 2025

Cinematic Data Retrieval: Pursuing a Better Search Mechanism

Authors:
Fernando Afonso
João Francisco Alves
João Pedro Passos
Pedro Cancela da Silva

Agenda

→ Project Theme
→ Dataset
→ Processing Pipeline
→ Conceptual Model
→ Characterization
→ Search Scenario

/03

Project Theme: **Movies/Series**

Context

Significant rise in media content → Need for efficient search systems

Theme

Movies and Series → Rich sources of structured and unstructured data

Objective

Build a **consistent dataset** → Solid Information and Search Retrieval System

/04 Dataset

Sources

IMDb, Kaggle, OpenDataBay and StreamWithVPN

Datasets

We gathered **4 datasets** from **IMDB**:

- Title Basics: 11962455 entries and 9 features
- Title Ratings: 1619879 entries and 7 features
- Title Principals: 95180101 entries and 6 features
- Name Basics: 14742901 entries and 6 features

Movie and series **descriptions** were **web scraped** from **Stream With VPN** website

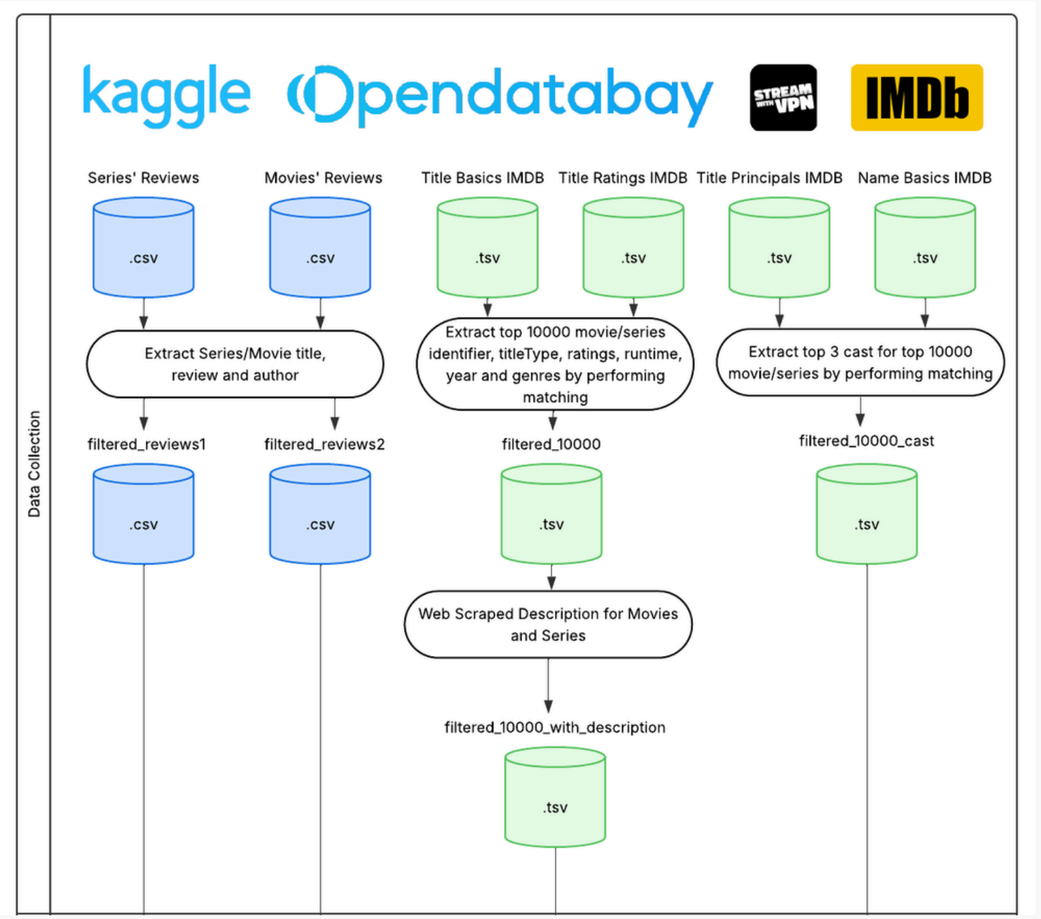
Then we also got 2 **datasets** from **Kaggle** and **OpenDataBay**:

- Movie Reviews: 664436 entries and 8 features
- Series Reviews: 52895 entries and 13 features

Pipeline

Data Collection

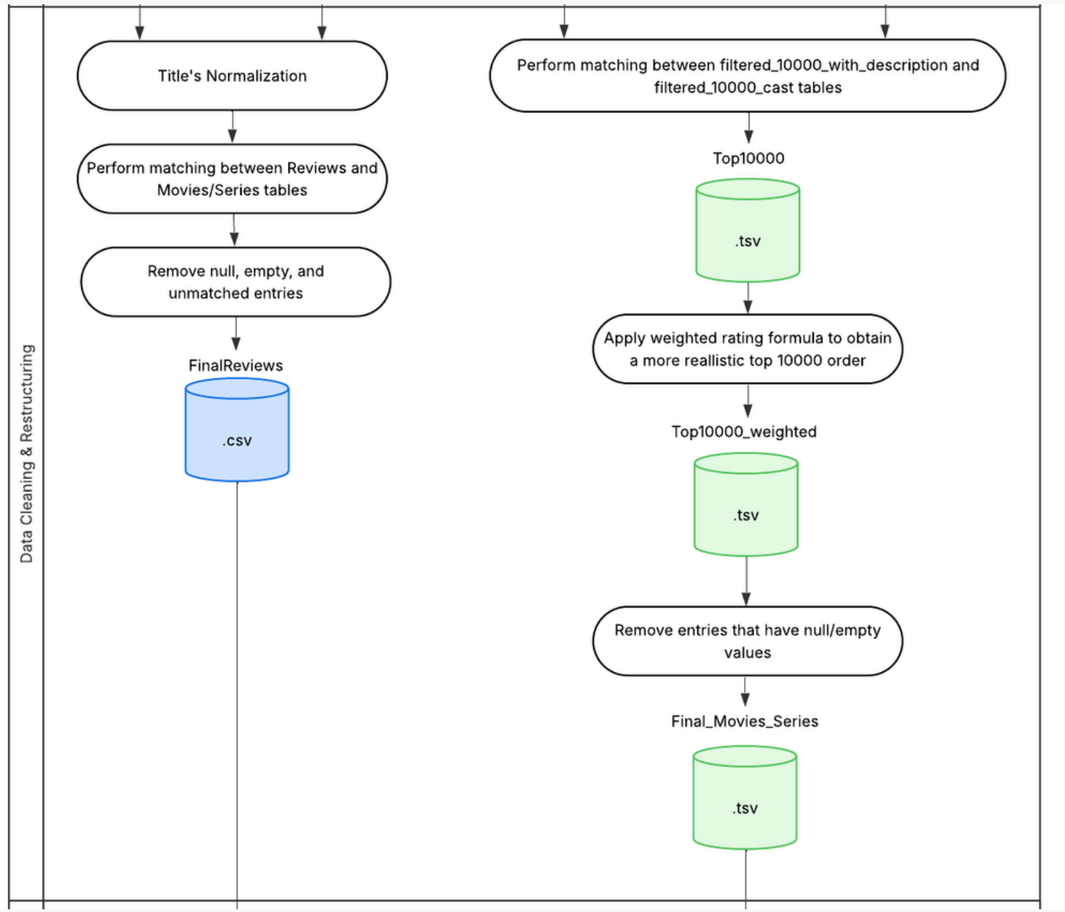
Extraction and scraping of data from different sources;



Pipeline

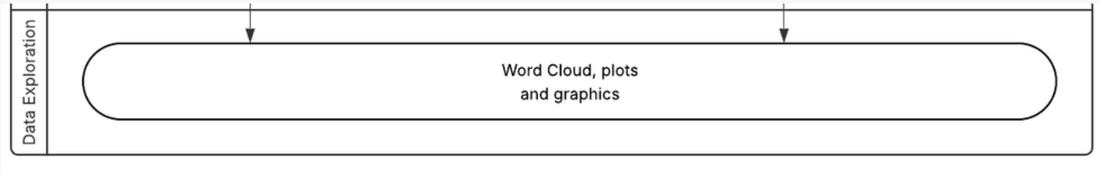
Data Cleaning & Restructuring

Normalization, adjustment and cleaning of data;



Data Exploration

Statistical analysis with graphics and plots



/06

Conceptual Model

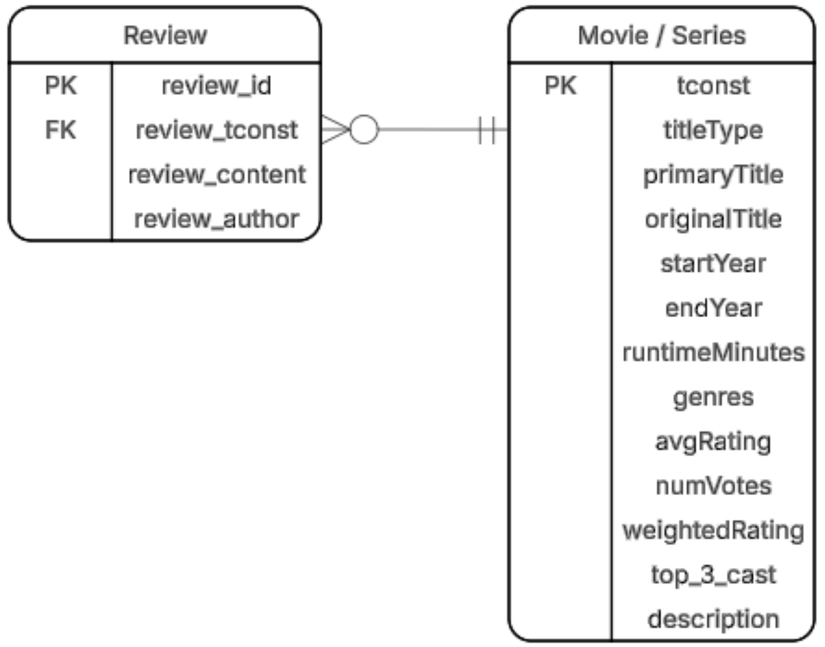
Two main entities:

Movies/Series → Each movie or series stores both structured data (titles, genre, cast, ...) and unstructured (description).

Reviews → Each review provides unstructured information (the reviewer opinion).

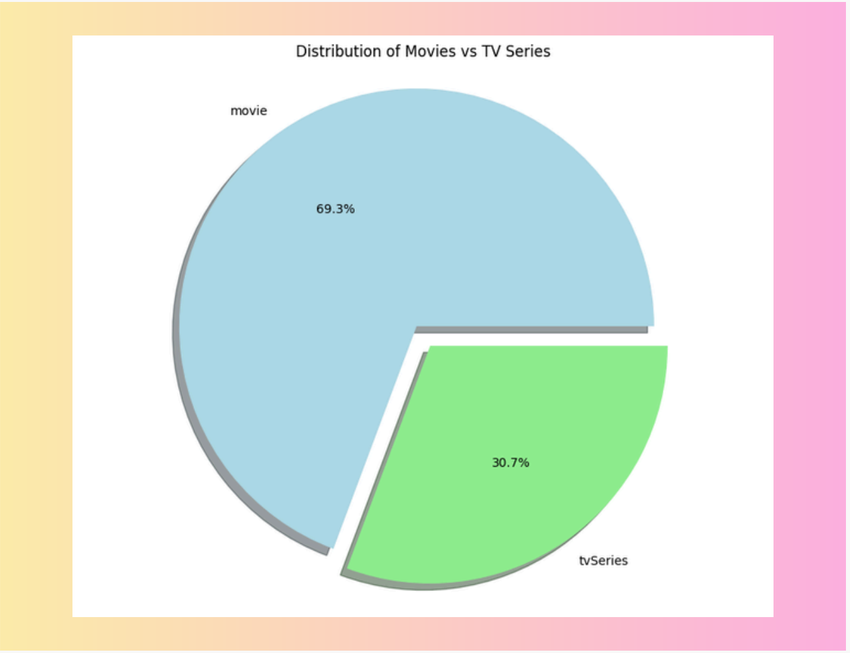
0-to-many Relationship

Each movie/series can have none or several reviews associated with it.

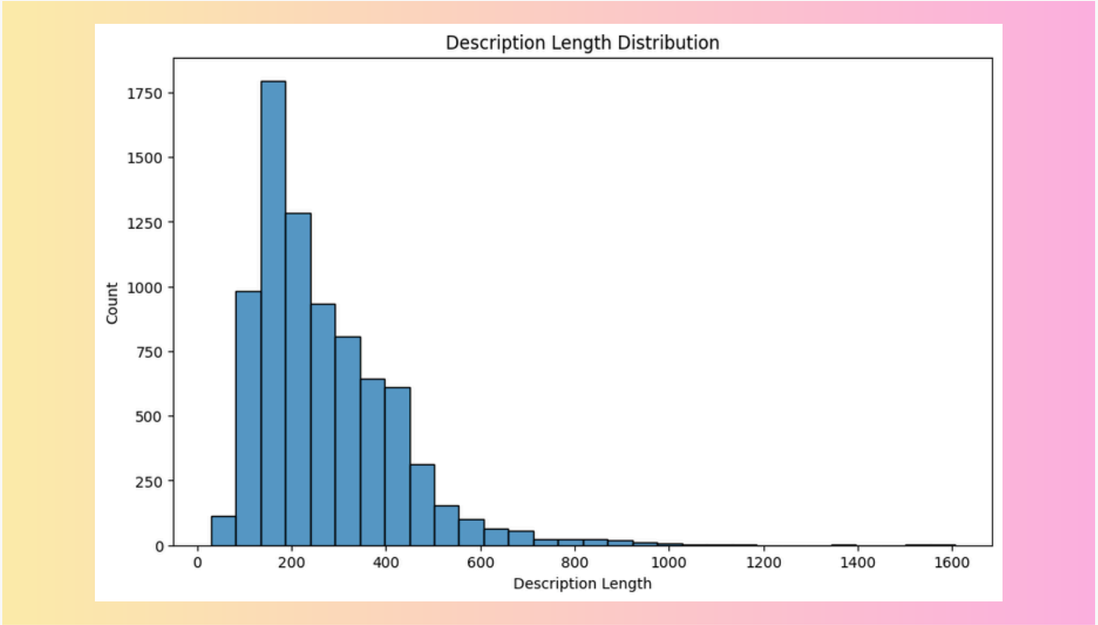


/07

Distribution by Title Type



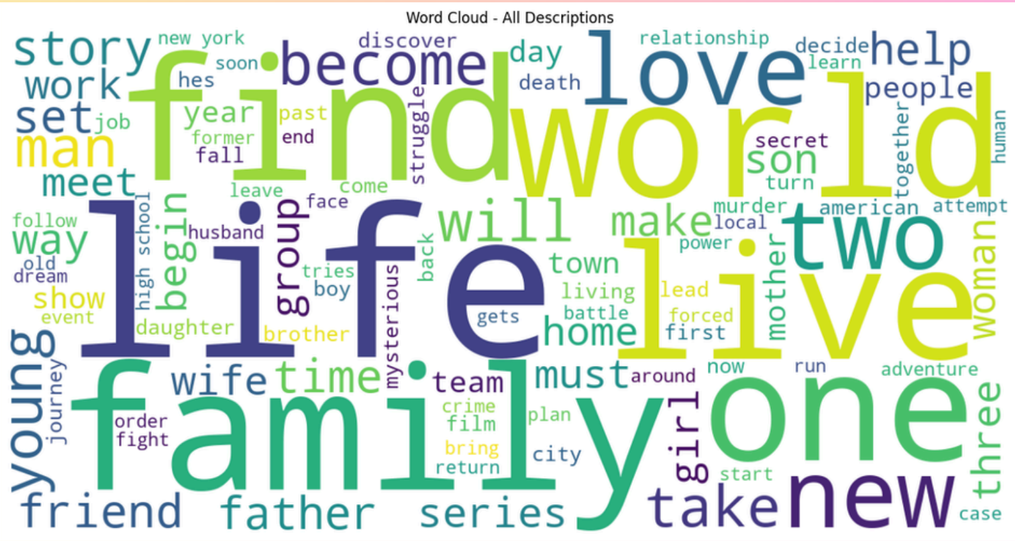
Description Length Distribution



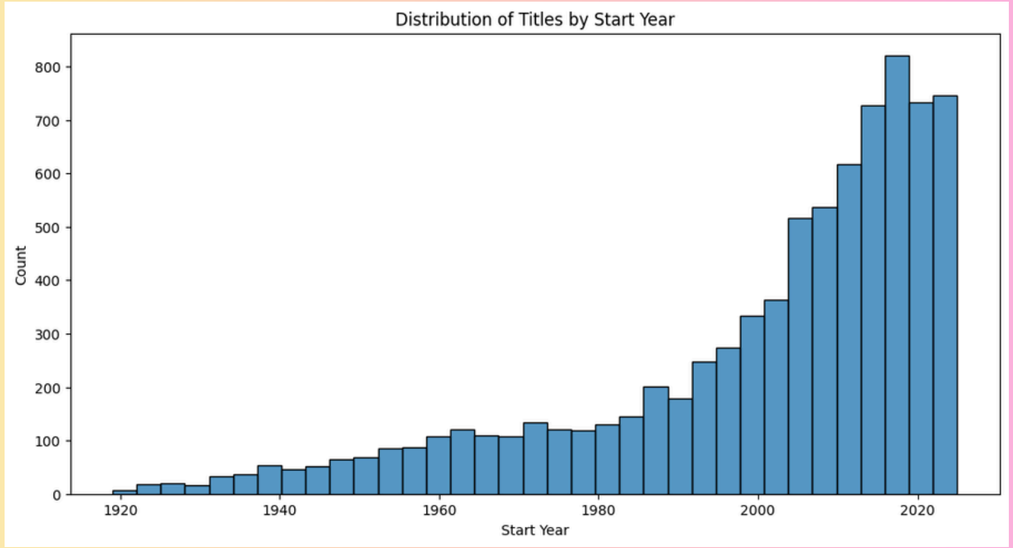
Data Characterization

/07-1

Descriptions Word Cloud



Distribution of Titles by Year

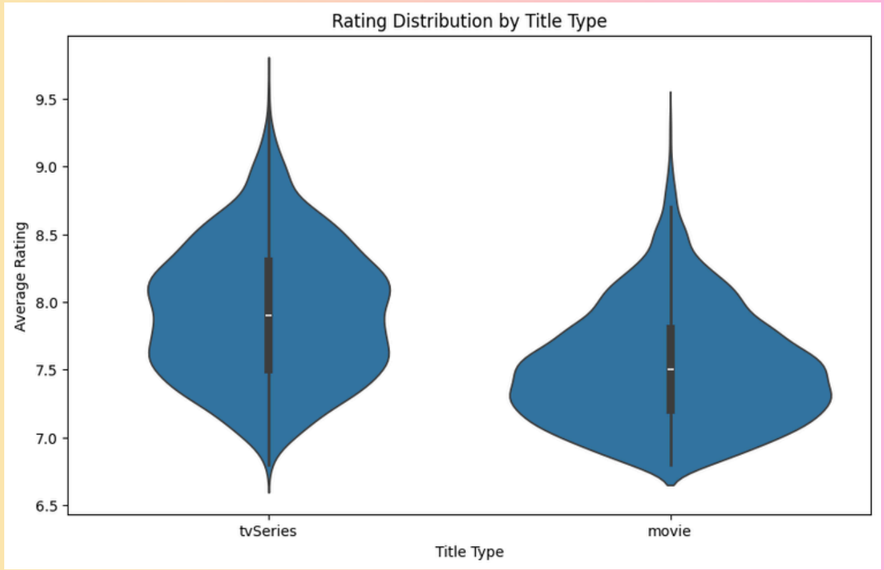


Data Characterization

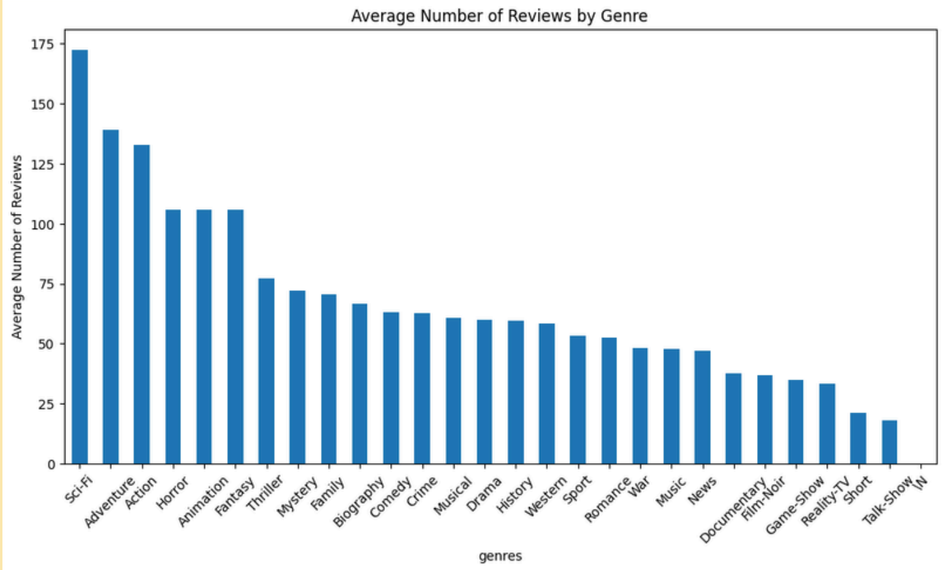
PRI | 2025

/07-2

Rating Distribution by Title Type



Average Number of Reviews by Genre



PRI | 2025

Data Characterization

Search Scenario

Similarity Search

“Find movies similar to *Inception*”

Theme Based Discovery

“Retrieve movies about artificial intelligence”

Era and Genre Filtering

“List comedies from the 90s”

Runtime Finder

“Get movies with a runtimes below 90 minutes”

Cross-Media Recommendations

“Films similar to *Game of Thrones* but standalone”

Search Scenario

Actor Based Search

“Series where Bryan Cranston is featured”

Rating-Weighted Discovery

“Best reviewed content during pandemic era (2020-2022)”

Reviewer Trust Metrics

“Films recommended by Roger Ebert”

Length-Aware Review Filtering

“Movies with detailed analytical reviews (average review above 500 words)”

/9

Conclusions

We might not have included every movie or series ever made, and some had to be removed because their descriptions were missing or incomplete. Still, we believe we have put together a solid and diverse collection. The detailed information we kept will help us discover interesting trends and improve the search and recommendation features of our movie and series search engine.

Thank you!