Title

Data Analysis of Movie Genres and Directors by Country

Authors

Qiansha Meng
Ni Liu

Problem Statement

The objective of this project is to analyze the movie genres and directors in different countries using the IMDb dataset. Specifically, we aim to determine the most popular movie genres and the most successful director of each country for each genre.

Introduction to the Problem

The film industry has been growing exponentially over the past few decades, and the globalization of the industry has resulted in a wide variety of movies being produced in different countries. With this growth, it has become increasingly important to understand the trends in movie genres and the success of directors in different countries. By analyzing the IMDb dataset to determine the most popular movie genres and the most successful directors for each country for each genre, it has the potential to provide some valuable insights to both the general public and film industry professionals.

It helps the public understand the cultural preferences of different countries and how these preferences may shape the movies they watch. For the film industry professionals, the result generated from the project can help them understand the market demand for different genres and the success rate of directors in different regions. This information can help filmmakers tailor their content to the preferences of specific regions and make more informed decisions when selecting directors for their projects.

Dataset Description

Dataset

Subsets of IMDb data

Data Source

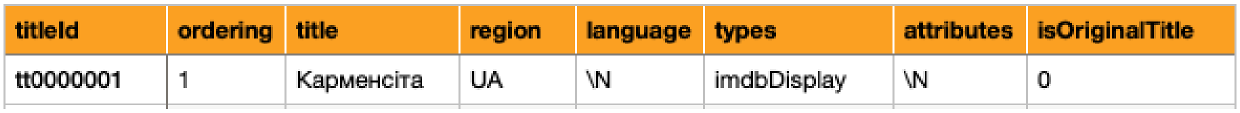
https://www.imdb.com/interfaces/
Download TSV files from https://datasets.imdbws.com/

Data Description

The dataset contains various information related to movies and TV shows. It is stored in TSV format and consists of multiple files. Each file has a different number of records, from 1 million to 35 million, and in a total of approximately 109 million records. Each file has a different number of attributes, with the lowest number of 3, the highest 9. Properties include title, region, type, year, rating, director, etc. The dataset generally spans from the early 1900s to the present day. Some titles may have information dating back to the late 1800s.

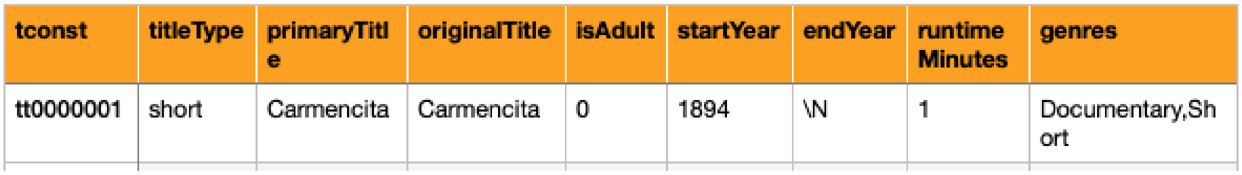
Data Snippet

title.akas.tsv



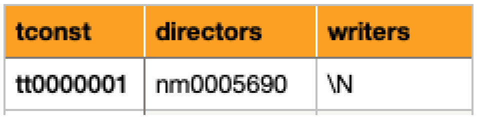
titleId	ordering	title	region	language	types	attributes	isOriginalTitle
tt0000001	1	Карменсіта	UA	\N	imdbDisplay	\N	0

title.basics.tsv



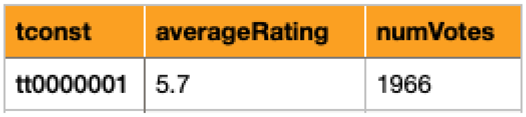
tconst	titleType	primaryTitle	originalTitle	isAdult	startYear	endYear	runtimeMinutes	genres
tt0000001	short	Carmencita	Carmencita	0	1894	\N	1	Documentary,Short

title.crew.tsv



tconst	directors	writers
tt0000001	nm0005690	\N

title.ratings.tsv



tconst	averageRating	numVotes
tt0000001	5.7	1966

Proposed Plan

Analysis Task

We want to analyze the movie genres and directors in different countries. Our goal is to get the top five highest rated movie genres for each country and the most successful director in corresponding sector and region.

Major Tasks

- Major task 1
  - Determine the most popular movie genres for each country using **Hive**.
- Major Task 2
  - Determine the most successful director for each genre for each country using **Hive**.

Helper Tasks

- Data cleaning, data transformation, and data aggregation: First remove any duplicate or irrelevant data from the dataset. And then extract the necessary data fields for analysis, such as the country, genre, and director of each movie. Finally, group the movies by country, genre, and director, and calculate a metric such as the average rating for each genre for a country, or for director in each genre in each country.
- Design a program that integrates the two outputs from the two major tasks. This could involve writing code to combine the data and perform any necessary calculations or analysis to answer the final question, such as identifying the most successful directors for the most popular movie genres in each country. This task could involve data cleaning and transformation, joining the two data sets, aggregating the data, and presenting the results in a clear and concise manner. Additionally, it may involve creating reports to help communicate the insights gained from the analysis.
- Use a query and data from the TPC-H benchmark to help evaluate the performance of the Hive queries to ensure that our queries are correct, optimized for performance, and reliable.
- Compare the performance to an implementation in Pig Latin.