INFO-511 FINAL PROJECT
CineScope: Crafting Custom Recommendations & Insights from IMDb Data

Team - Data Dazzlers

Introduction & Motivation

Introduction

- Evolution of Entertainment
- Data-Driven Decisions in OTT

Motivation

- Consumer Expectations
- Viewer Engagement Challenges

Questions we are trying to find answers

- What are the trends in the number of series vs. movies released over the last decade?
- How are the top genres distributed in ratings?

About Data

- Source: https://datasets.imdbws.com/ IMDb Dataset
- Dataset Contents:

name.basics.tsv.gz - Personalities Info

title.akas.tsv.gz - Alternative Titles

title.basics.tsv.gz - Basic Titles Data

title.crew.tsv.gz - Crew Details

title.episode.tsv.gz - Episode Information

title.principals.tsv.gz - Principal Participants

title.ratings.tsv.gz - Ratings Data

Exploring Data

title.basics.tsv.gz

- **11,290,664** entries
- 9 columns – tconst, titleType, primaryTitle, originalTitle, isAdult, startYear, endYear, runtimeMinutes, genres
- All columns are stored as objects, indicating text or mixed types.

title.ratings.tsv.gz

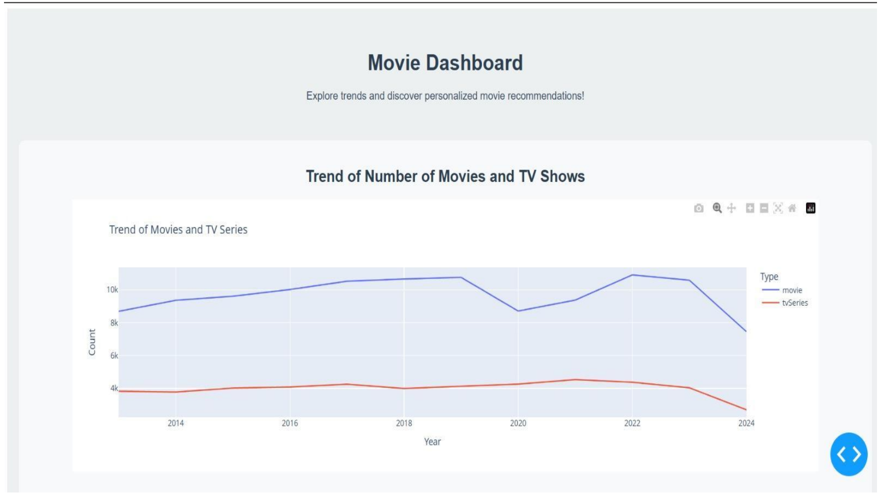
- **1,508,659** entries
- 3 columns - tconst, averageRating, numVotes
- Includes numerical data (averageRating, numVotes) for detailed statistical analysis.

Pre-Processing

- **Data Storage and Retrieval:HDF5 Files:** Utilize HDF5 format (Hierarchical Data Format version 5) for efficient storage and retrieval of movie data and feature data which supports handling large volumes of data.
- **Feature Normalization:** Using PyTorch to normalize movie features to unit vectors, a critical step for implementing cosine similarity effectively.
- **Feature Engineering:** Convert feature arrays into tensors using PyTorch, optimizing them for high-performance computations necessary in similarity calculations.
- **Error Handling and Debugging:** Incorporate error handling to manage missing titles and ensure robust system performance.
- **Feedback Loops:** Provide user feedback when titles are not found or other issues arise, enhancing user experience and system reliability.

Deployment

- **Application Framework:** Using Dash by Plotly. Allowing users to input preferences and view movie recommendations.
- **Backend Technologies:Python Libraries:** Utilizes Pandas for data handling, PyTorch for tensor operations, and h5py for efficient data storage in HDF5 format.
- **Data Management:** Utilize HDF5 files. Normalize movie features using PyTorch for cosine similarity in the recommendation algorithm.



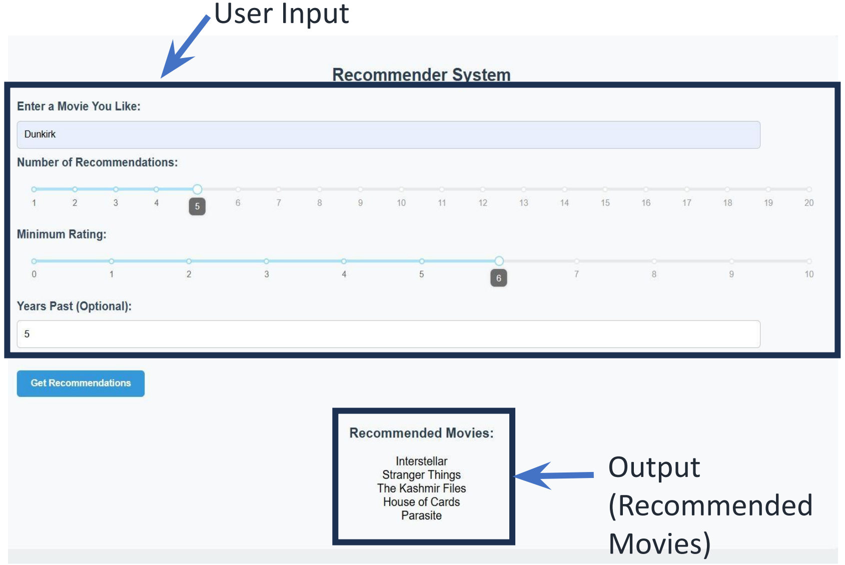
Movie Dashboard

Explore trends and discover personalized movie recommendations!

Trend of Number of Movies and TV Shows

Deployment

- **Recommendation Engines:** Users can enter movie titles to get personalized recommendations based on cosine similarity of feature vectors.
- **Functionality:** Allows users to specify the number of recommendations, minimum rating, and recency of movies through interactive sliders and inputs.
- **Visualization and analytics:** Using Plotly Express to generate interactive plots to show current trends of different platforms.



User Input

Recommender System

Enter a Movie You Like: Dunkirk

Number of Recommendations: 5

Minimum Rating: 6

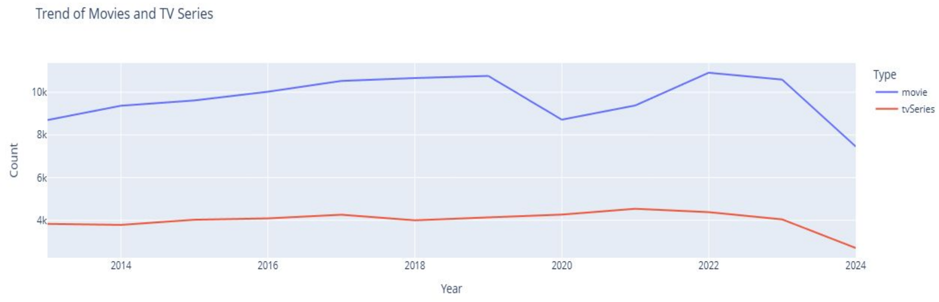
Years Past (Optional): 5

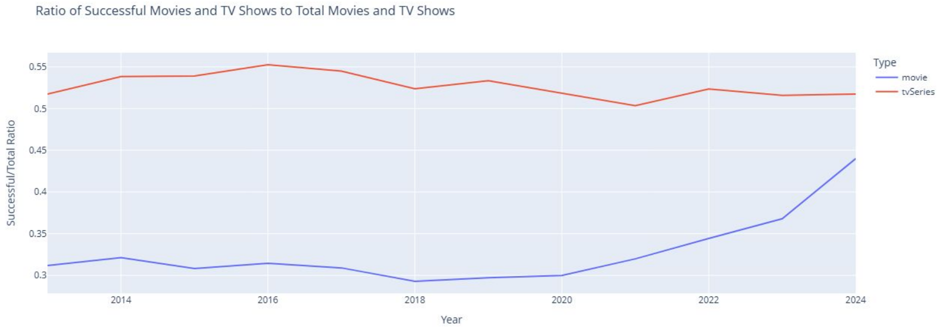
Recommended Movies:

Interstellar
Stranger Things
The Kashmir Files
House of Cards
Parasite

Output (Recommended Movies)

Analysis

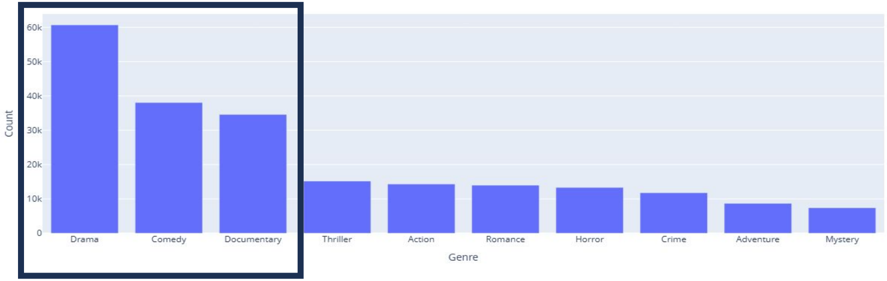




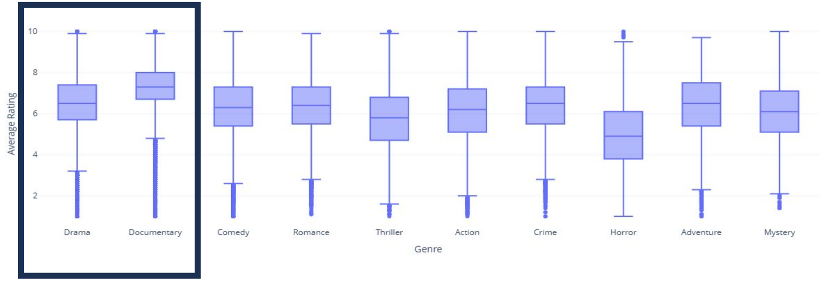
The graph shows stability in movie productions from 2014 to 2022 with some decline from 2023, while TV series production remains relatively stable.

The graph shows a stable success ratio for TV series, while the success ratio for movies significantly increases from 2020, indicating improved performance of movies in recent years.

Analysis



Top 10 Genres by Count



Distribution of Ratings for Top 10 Genres

Drama is the most prevalent, followed by Comedy and Documentary, indicating a higher production or availability of these genres compared to others like Mystery and Adventure.

Drama and Documentary genres generally receive higher median ratings compared to others like Horror and Action, with a wide range of ratings observed in most genres.

Conclusion & Future Work

- **Conclusion:**

1. The visualizations show us that while tv shows are overall more successful but in the recent years (particularly post covid) success ratios of movies has increased significantly.
2. Genres like Drama and Action show a wide range of ratings, indicating diverse content, while Documentary and Mystery genres exhibit more consistent ratings, suggesting uniform audience reception.

- **Enhanced Personalization through User Feedback:** Develop a feature allowing users to rate the accuracy and relevance of recommendations. Utilize these feedbacks to adjust the weights of the recommendations and enhancing the system's ability to improve the learning process.
- **User Experience:** Increase user engagement by making the recommendation process interactive, fostering a more personalized and satisfying experience.

Thank You