Pattern Recognition and Machine Learning Project

Movie Recommendation System

March 30, 2024

1 Problem Statement

The goal of this project is to develop a movie recommendation system using machine learning techniques. This system will analyze user data and movie information to predict movies a user might be interested in watching.

These results are derived from the user's profile, search and browsing history, the viewing habits of individuals with similar characteristics or demographics, and the likelihood of the user watching those films.

2 Data Set

We will use a Movie Lens Small Latest Dataset which is available in Kaggle [link]. This dataset contains four csv files.

1. links.csv: This file acts like a translator between different movie identification systems. It contains three columns:

movieId: A unique identifier for a movie within the MovieLens dataset.

imdbId: The movie's ID on the Internet Movie Database (IMDb) website .

tmdbId: The movie's ID on The Movie Database (TMDB) website .

2. movies.csv: This file holds the core information about the movies themselves. It contains three columns:

movieId: The same unique movie identifier from the links.csv file.

title: The full title of the movie.

genres: genre of the movie .

This data allows to identify movies and potentially categorise them based on genre .

3. ratings.csv: This file captures user interactions with the movies. It contains three main columns and one additional timestamp column:

userId: A unique identifier for a user within the MovieLens dataset.

movieId: The movie identifier again, linking back to the movies.csv file.

rating: The rating a user gave to the specific movie (scale of 1 to 5).

timestamp: seconds since midnight Coordinated Universal Time (UTC) of January 1, 1970.

4. tags.csv: This file captures additional user input on movies. It contains three main columns and one timestamp column:

userId: The same user identifier from the ratings.csv file.

movieId: The movie identifier again, linking back to the movies.csv file.

tag: A keyword user assigned to describe the movie .

timestamp: seconds since midnight Coordinated Universal Time (UTC) of January 1, 1970.

2.1 Insights from the dataset

1. Total unique MovieId's in the dataset are 9742
2. Total unique userId's in the dataset are 610

3. Genres are a pipe-separated list, and are selected from the following:

   * Action * Adventure * Animation * Children's * Comedy * Crime * Documentary * Drama * Fantasy * Film-Noir * Horror * Musical * Mystery * Romance * Sci-Fi * Thriller * War * Western * (no genres listed)

4. We are dropping the timestamp column in tags.csv and movies.csv

3 Proposed Approaches

Movie Recommendation System can be done in many ways.Some of the ways are content based filtering, collaborative filtering and popularity based methods. We will be using cosine similarity, knn,decision trees for doing this.

In the frontend page end user has to enter the userId then top 5 recommended movies for that specific userId will be shown. In the other page end user has to enter the genre and rating range by which top 5 recommended movies will be shown.

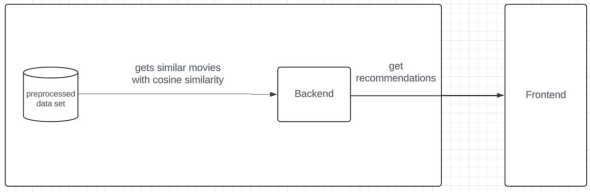

Figure 1: proposed architecture

4 Early Results

- The below is the pie chart representing the distribution of different ratings.Each slice of the pie represents a specific rating given by user's and its corresponding percentage.

  1. 26.60 % of the movies have rating 4.0.
  2. 19.88% of the movies have rating 3.0.

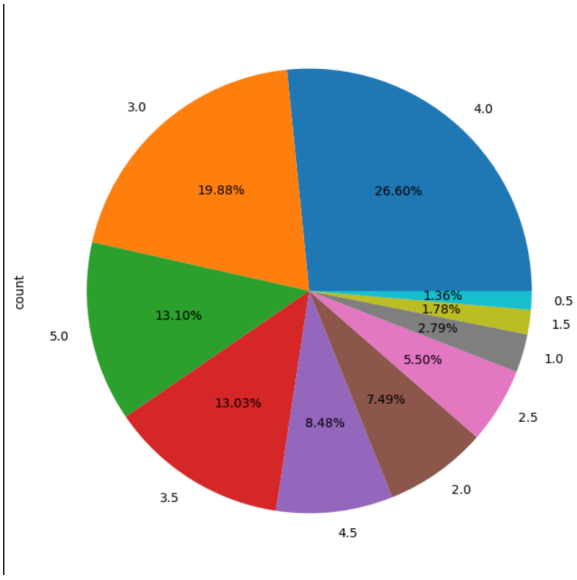

Figure 2: pie chart representing the distribution of different ratings

- The top 5 most popular movies in the dataset with a minimum of 100 user's rated are

Movie Title	**Average Rating**
Shawshank Redemption, The (1994)	4.42
Godfather, The (1972)	4.28
Fight Club (1999)	4.27
Godfather: Part II, The (1974)	4.259
Departed, The (2006)	4.252

References

[1] F. Maxwell Harper and Joseph A. Konstan. *The MovieLens Datasets: History and Context*. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4: 19:1–19:19, 2015. https://doi.org/10.1145/2827872