Movie Recommendation ("MuseMovie") Proposal

Background

MuseMovie is a (fake) movie recommendation platform that utilizes advanced algorithms to provide highly personalized movie recommendations to users based on their past viewing history and preferred genres. However, the challenge arises when it comes to new users who have little or no viewing history on the platform, which is called the cold-start problem. In such cases, without sufficient data on a user's preferences and behavior, providing relevant recommendations can be difficult. Thus, MuseMovie wants to add a new feature aimed for those new users.

Project Goals and Objectives

We aim to provide relevant movie recommendations to new customers based on their interactions with the website. To achieve this, we propose using a machine learning model that calculates the similarity between different movies and incorporates the average rating of each movie into the model. Finally, we will evaluate the model's performance by testing its recommendations against a sample of customer ratings. With our platform, new customers can expect a seamless, enjoyable movie-watching experience, complete with top three movie recommendations that perfectly align with interactions with the website.

Dataset Description

This dataset, in general, contains the user's rating, tags, and tag relevance of each movie. It contains 25,000,095 ratings and 1,093,360 tag applications across 62,423 movies.

This dataset contains 4 major sub-datasets, which are:

Movies: it contains the name and dataset ID of a movie, as well as its genres.

Rating: it contains the user ID, movie ID, and the rate (from 0.5 to 5.0) this user gives to the movie.

Tags: it contains the user ID, movie ID, and the tags this user gives to the movie.

Genome_score: it contains movie ID, tag ID, and the tag's relevance of the movie, which are calculated using a machine learning algorithm mentioned in http://files.grouplens.org/papers/tag_genome.pdf

Except those four major sub-datasets, there are also 2 sub-datasets which are used to connect datasets:

Links: it contains movie ID IMDB movie ID, and TMDB ID.

Genome-tags: it contains tags ID and the name of the tags, which are written by customers with typos and punctuation characters.

With this dataset, we plan to use the Genome_score dataset to find movies with high similarity, and use Rating dataset to measure the quality of the movies, then we would provide a list of recommendations.

Metrics

We will split the dataset into a training set and a testing set. We would get the recommendation system using the training set, and use the testing set to measure how good the recommendation system is. For example, the system recommends movie B for users who have watched movie A. We will use the average rating of movie B from users who also have watched movie A as the score of this recommendation.