IE7275 Fall 2023 Term Project Proposal
Saravanan Arumugam, Reenae Victor

Movie Clustering and Rating Prediction for Content Organization

1. Team Responsibility

Our team consists of Saravanan Arumugam and Reenae Victor and both of us will be sharing the project responsibilities equally.

2. Introduction

In the ever-evolving landscape of the entertainment industry, the demand for personalized content recommendations and effective content organization has surged. With the vast array of movies available, understanding and categorizing them effectively is paramount. This project aspires to address this challenge by creating a comprehensive movie clustering system and developing a movie rating prediction model.

3. Problem Definition

The project goals are two-fold:

Firstly, Movie Clustering is all about sorting movies into meaningful groups based on factors like genre, cast, crew, and user ratings. It's like finding the hidden connections between different types of movies. The goal is to craft a more personalized content experience for viewers. Whether we're suggesting movies based on your previous favorites or helping marketers fine-tune their advertising, the potential for content enhancement is immense. Secondly, Movie rating prediction involves creating a model that can accurately predict what rating a user might give to a movie. It's a game-changer for personalized recommendations. By understanding the intricate factors that influence your movie rating, we're getting closer to a tailored content experience that matches your unique taste.

4. Data Description

We will use the IMDB dataset on Kaggle (dataset link). This dataset contains 5 tab separated value files namely:

title.akas.tsv.gz: Contains information about localized titles, regions, languages, and more.

title.basics.tsv.gz: Provides details about titles, including title type, primary and original titles, genre, and more.

title.principals.tsv.gz: Contains information about the principal cast and crew for titles.

title.ratings.tsv.gz: Provides IMDb rating and vote information for titles.

name.basics.tsv.gz: Contains information about names (actors, directors, etc.), including their primary names, birth and death years, professions, and known titles.

5. Proposed Methods

We propose the following steps, In Data Preprocessing we will clean and preprocess the data, handling missing values, and converting it into a suitable format for analysis. For movie clustering we plan to use dimensionality reduction to select relevant features for clustering. Implement K-Means, Hierarchical, and Spectral clustering algorithms. Evaluate clustering results using metrics like Silhouette Score and Calinski-Harabasz Index. Create visualizations to interpret and present the clustering results. For rating prediction, we will develop a regression and a Random Forest model to predict movie ratings based on features like movie attributes, user interactions, and historical ratings. Evaluate the model's performance using metrics like Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

6. Expecting Results

A functional Movie Clustering system that categorizes movies based on content attributes. A movie rating prediction model that accurately estimates user ratings for movies. Improved content organization and marketing strategies based on the insights generated through clustering and rating predictions. This project will also demonstrate the capabilities of machine learning and data analysis in content organization and user personalization within the entertainment industry. By combining movie clustering and rating prediction, we can provide valuable insights for content marketing strategies.

References,

1. Ricci, F., Rokach, L., & Shapira, B. (2011). Recommender systems: Introduction and advances. Springer.
2. Berry, M. W. (2007). Machine learning for content organization. Springer.