Movie Recommendation System

Team Members:
Danielle Aras
Soumyadeb Maity
Jhansi Saketa B V

About the Project

In a crowded entertainment market, movie streaming services and advertisers need to present customers with the most relevant recommendations possible to maintain customer interest and loyalty. This project will use a database of user-submitted movie ratings to explore ways to generate movie recommendations and predict how users may rate future movies.

Prior Work

- Netflix Recommender System

  Carlos A. Gomez-Uribe and Neil Hunt. 2015. The Netflix recommender system: Algorithms, business value, and innovation. ACM Trans. Manage. Inf. Syst. 6, 4, Article 13 (December 2015), 19 pages. DOI: http://dx.doi.org/10.1145/2843948

- https://developers.google.com/machine-learning/recommendation

DataSet

MovieLens Dataset

“Latest Full” set will be used (27 million data points)

“Latest Small” can be used for testing (100,000 data points)

URL: https://grouplens.org/datasets/movielens/

The dataset has been downloaded to Danielle’s computer

Proposed Work

Cleaning and Pre-Processing

- Minimal cleaning to check for null values (movies without genres, missing ratings, etc)
- Combine separate csv files for ratings and movies into a single table so ratings are linked to movie genres
- Create user genre rating table with average ratings for each genre of movie for each user

Proposed Work

There are two major techniques to develop any recommendation systems.

- Collaborative filtering - SVD, KNN, clustering etc
- Content based filtering - TFIDF and dot product / sum of product.

Content based filtering is good for handling the cold start problem. We will try to explore all the approaches and try to measure the advantages and disadvantages of them.

Proposed Work

Clustering

- Use k-means clustering to divide reviewers into clusters with similar taste in movies based on their average ratings of different genres
- Clusters can be used to predict a reviewer's rating for a movie they have not seen yet based on the cluster average rating
- "Top" movies for each cluster can be found by identifying the highest-rated movies for each cluster
- Different k values will be used to find optimum number of clusters

Tools

- Python
- Pandas
- Scikit-learn clustering module and others: https://scikit-learn.org/stable/modules/clustering.html
- Matplotlib

Evaluation

- For clustering, the fit of the clusters will be evaluated by the sum of the squared error (SSE)
- Optimum number of clusters will be chosen based on the elbow method and silhouette coefficient
- We will use the inbuilt accuracy methods of the KNN and SVD algorithms in sklearn.

Reference

[1] Eyrun A. Eyjolfsdottir, Gaurangi Tilak, Nan Li (2008), “MovieGEN: A Movie Recommendation System”, 2008 Conference Proceedings.

[2] Roman, Victor (2019), “Unsupervised Classification Project: Building a Movie Recommender with Clustering Analysis and K-Means”, Towards Data Science https://towardsdatascience.com/unsupervised-classification-project-building-a-movie-recommender-with-clustering-analysis-and-4bab0738efe6