Movie Recommendation System

Jhansi Saketa B V  
 MS in Computer Science  
University of Colorado  
 Boulder, CO  
 jhbh9073@colorado.edu

FirstName Surname  
 Department Name  
 Institution/University Name  
 City State Country  
 [email@email.com](mailto:email@email.com)

FirstName Surname  
 Department Name  
 Institution/University Name  
 City State Country  
 email@email.com

MOTIVATION

All time, movies are number one source of entertainment. People enjoy watching them, reviewing them and discussing about them. For many people reviewing and rating movies is passion. A movie can have fans and critics too. By the term “Movie Recommendation System” recommend movies to the users. Based on the previous ratings given by the users on the movies, the movie recommendation system is going to recommend the movies. So, 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.

LITERATURE SURVEY

1. Netflix Recommender System, this paper discusses the various algorithms that make up the Netflix recommender system and describes its business purpose. It explains the motivation and reviews the approach that we use to improve the recommendation algorithms. [1]
2. Recommendation Systems, this course gives knowledge of recommendation systems and explain different models used in recommendation, including matrix factorization and deep neural networks. [2]

PROPOSED WORK

This Project “Movie Recommendation System” generally uses two major techniques.

1. Content Based Filtering: This algorithm will pick similar items i.e., items that have similar properties like similar content and then recommend the user with the similar movies that they liked previously.

Some example algorithms for content-based filtering are TF-IDF algorithm is “Term Frequency-Inverse Document Frequency and it is the most used algorithm to convert the text into vectors. This technique is widely used to extract features across various NLP applications.” Other algorithm is dot product / Sum of product.

1. Collaborative Based Filtering: It is User-User Collaborative Filtering i.e., it finds the similar users for a user who like and rated similar movies like them previously and recommend movies that the similar users like them liked the movies.

For Clustering, we are going to 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.

The data that we are using is almost a clean data that can be used directly. But the cleaning and pre-processing include,

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

DATA SET

The Dataset that we are using for this project is from MovieLens Dataset. There are different sizes of the data that they are providing, so we will use “Latest Full” dataset for the project which is of size 27 million data points. The “latest Small” dataset can be used for testing the model which is of size nearly 100,000 data points. The data can be accessed in the URL: <https://grouplens.org/datasets/movielens/> with the permission of the MovieLens organization. We filled a form and successfully got permission to use the dataset in our project.

EVALUATION METHODS

1. For clustering, the fit of the clusters will be evaluated by the sum of the squared error (SSE). The lower the SSE, then the more similar are the Users/Movies. A high SSE suggests that the Movies/Users in the same similar segment have a reasonable degree of differences between them and may not be a true (or usable) segment.
2. Optimum number of clusters will be chosen based on the elbow method and silhouette coefficient.
3. We will use the inbuilt accuracy methods of the KNN and SVD algorithms in sklearn.

TOOLS

The following tools are going to be used to develop the project.

1. Python

python programming language is core tool to build our project. The data extraction, data cleaning, pre-processing and all the models and algorithms are build using python programming language.

1. Pandas

Pandas is a module in python to work with the dataset. The major role of Pandas in our project is extraction of the data, creating data frames and working with the data frames.

1. Scikit Clustering modules

It is a free software library machine learning library for python programming language. It features various [classification](https://en.wikipedia.org/wiki/Statistical_classification), [regression](https://en.wikipedia.org/wiki/Regression_analysis) and [clustering](https://en.wikipedia.org/wiki/Cluster_analysis) algorithms including [support-vector machines](https://en.wikipedia.org/wiki/Support_vector_machine), [random forests](https://en.wikipedia.org/wiki/Random_forests), [gradient boosting](https://en.wikipedia.org/wiki/Gradient_boosting), [k-means](https://en.wikipedia.org/wiki/K-means_clustering) and [DBSCAN](https://en.wikipedia.org/wiki/DBSCAN), and is designed to interoperate with the Python numerical and scientific libraries [NumPy](https://en.wikipedia.org/wiki/NumPy) and [SciPy](https://en.wikipedia.org/wiki/SciPy). [3]

1. Matplotlib

Matplotlib is a plotting library for the [Python](https://en.wikipedia.org/wiki/Python_(programming_language)) programming language and its numerical mathematics extension [NumPy](https://en.wikipedia.org/wiki/NumPy). In our project, we use matplotlib to plot the graphs or visualize the results.

MILESTONES

REFERENCES

*[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-b uilding-a-movie-recommender-with-clustering-analysis-and-4bab0738e fe6*](https://towardsdatascience.com/unsupervised-classification-project-b%20uilding-a-movie-recommender-with-clustering-analysis-and-4bab0738e%20fe6)

*[3]* [*https://en.wikipedia.org/wiki/Scikit-learn*](https://en.wikipedia.org/wiki/Scikit-learn)

Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00