CS6120 Project - Movie Recommendation

Group 24: Xuan Zhang, Yi Chen, Meishan Li, Jia Xu, Qia Lin

Description

In recent years more and more people are choosing to watch movies on the Internet. The variety and number of movies are also increasing. People often don't know which one to watch instead of facing so many movies. If there is a good recommendation application for movies, these recommendations will help people to choose a movie of the type they want to watch. We will implement a movie recommendation tool by using unsupervised learning models to group movies, and using cosine similarity to select movies the users want to watch more effectively.

Dataset

We will use the TMDB 5000 Movies dataset from Kaggle. There are two data sets. First data set is tmdb_5000_movies which contains 4083 movies information. These information include budget, genres, keywords, overview, popularity and vote_average. Second data set is tmdb_5000_credits which contains movie_id, title, cast, crew. We will combine the FEATURES of these two datasets to make a recommendation system.

Dataset link: https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata

Methods, Models and Tools

We first observe the data by various methods which include T-sne etc and select the appropriate features of the data based on the results of the observation to tokenize. Then we vectorize the tokens from the data by using the BOW method and TF-IDF method. To the BOW method and TF-IDF method, we will create them by our own code. After vectorizing the tokens, we will use clustering algorithms (like k-mean) or LSH to group data for us to find what we want more quickly and easily. And we will calculate cosine similarity between data which are movies feature vectors and input feature vectors from someone's movie viewing record, then give several movies of recommendations for the highest similarity. Finally, we will use precision, recall and accuracy to evaluate this model. We hope to get high precision, recall and accuracy.

In this model, we may use Sklearn, nltk, numpy, re, pandas, math, seaborn, matplotlib, string, contractions. Sklearn may be used to do clustering, preprocessing etc. We will use nltk, re and string to tokenize and delete stopwords etc. We will use pandas, seaborn, matplotlib to observe data and plot pictures about T-sne, RMES, precision, recall and accuracy etc . We will use math to calculate cosine similarity.

Milestones and Deliverables

Week 9	● Collect the dataset and try to clean and preprocess the dataset. ● Observe the data by various methods which include T-sne etc and select the appropriate features of the data.
Week 10	● Vectorize the tokens from the data by using the BOW method and TF-IDF method. ● Implement the BOW method and TF-IDF method.
Week 11	● Use clustering algorithms (like k-mean) or LSH to group data.
Week 12	● Calculate cosine similarity and generate recommendations. ● Evaluate the model by precision, RMES, recall and accuracy.
Week 13	● Write the final report. ● Prepare for the presentation.

Responsibilities

Xuan Zhang: Code, Proposal, Data Preprocessing, etc.
Yi Chen: Code, Presentation, Vectorize Tokens, etc.
Meishan Li: Code, Slides, Data Clustering, etc.
Jia Xu: Code, Report, Calculate Cosine Similarity, etc.
Qia Lin: Code, Model comparison, Evaluation, etc.