Panitan Satamalee

CS 235 (MSOL)

Recommendation of Movies (Software 1)

This project will be a movie recommendation system based on the Software 1 project type. The dataset that will be used will be from GroupLens Research (which collected data from MovieLens). The dataset consists of 5 different files. “ratings.csv” and “tags.csv” contains the ratings and tags of a given user ID for a given movie ID. “movies.csv” and “links.csv” contains the movie ID, its title, the movie’s genre, and the links for the movie’s IMDb and TMDb profiles. There are also the tag genome files that contain the relevance scores between the relationships of a specific tag to a movie.

By using ratings, and possibly tags and relevance, we want to try to cluster users together based on the ratings that they give to each movie and maybe find some commonality between user’s opinion on a movie or maybe towards a genre. Given a user and his ratings and tags for the movies he has seen, we want to be able to give a movie recommendation that he would enjoy. Another way of saying this would be that we want to recommend a movie where he would give a high rating, somewhere between 3-5 stars. If a list of movies a user likes is given, then we can also assume that the ratings he would have given those movies are probably 3-4, and then we can try to give a recommendation off of that.

The first step would be to combine all the files into a single table. The ratings and tags files will be combined such that there will be a row for each unique user and each column will be either a rating or tag for a movie. Then the tags can be used to find the corresponding tag ID which would give its relevance. We can then generate potentially useful features such as the average rating, or relevance, for certain movie genre for a user, or we can also generate some useful information from the movie titles such as series, location, etc. Stratified sampling, or random sampling, would be implemented to create a training and testing/validation set.

Three different methods will be implemented for recommendations given a user. The first is collaborative filtering. This finds all users who have rating similarities with the given user and will use their ratings of other movies to find the recommended movie of the given user. The second method will be matrix decomposition, and the third method will be clustering through k-means. From these methods, we do several iterations of test samples to figure out the optimal algorithm that will result in the best recommendations for a user.

** Note: If needed, I can join another group working on a similar/different topic if the number of groups currently is too high.