CPE695 Team Project Proposal
Group Member: Mu,Yu; Pang,Liangfan, Das, Himel

Preparation: We choose to work on the MovieLens dataset since we are all huge movie fans. After carefully examine the description posted on MovieLens, we decide to choose the dataset with the size of 1,000,000 so that we can better analyze the data and reach our goal.

Goal: According to the readme.txt posted on the MovieLens, there are a total of 3,600 different tags applied to 9,000 movies. Before we enter the movie theatre or watch movies online, we want to know about the detail information of movies such as "tags" and "ratings" in order to have a satisfying experience. Therefore, our goal in this project is to link a connection between the tags and rating. Moreover, we would predict the rating of a movie with specific tags.

Content:

- Read in csv files and clean the data.
- Display the top 100 appearances from the 3,600 tags, use these 100 top tags as our [feature]. Considering that some movie may have multiple tags, we would have to first analyze the number of tags for each movie. Suppose the median for tags is X, then we create 10 [multi_features], each consisting X tags. Also, some movie may not have multiple tags. Then we select the top 10 tags from the 100 tags as the [single_feature].
- Sort the "rating.csv" file with different stars, then use the [multi_feature] and [single_feature] to filter the rating again. Make plots to find some connection between specific features and ratings. Use ratings as the [target]
- Create DecisionTree and CrossValidation model, randomly select 70% from the whole set as the train set, and the rest 30% as the test set.
- Find the best parameter for a the DecisionTree model by using CV. Fit the train set and generate a proper model for prediction.
- Make predictions and calculate the accuracy for each feature.

(P.S. If there exist such model that the accuracy is lower than 0.75, then we use other models to do the prediction again. If the accuracy is still not acceptable, we would change the current feature)