INFO411/INFO911
Data Mining and Knowledge Discovery

Project 2

Instructions:
You are required to prepare a set of presentation slides which must include (1) the full name and student number of each student in the group, the contribution (in percent) of each group member, (2) a description of the task, (3) your proposed data mining approach and methodology; (4) the results and an analysis of the results; (5) the strengths and weaknesses of your proposed approach; (6) the results a brief discussion and a conclusion.

Below is the recommended structure of your slides:

- Introduction (define the problem and the goal)
- Methods (propose approaches, and discuss their strengths and weaknesses)
- Results (illustrative Figures and Tables and explanations)
- Discussion (discovered knowledge?)

Task: Recommender System for Movies
Background:
Recommendation systems are designed to recommend items to customers. These recommendations are created on the basis of statistics of a customer's history and the histories of customers with similar interests. Recommendation systems are used at Amazon, Youtube, Netflix, eBay, hotels.com, just to name a few.
MovieLens helps movie enthusiasts to find movies they will like. Customers rate movies to build a custom taste profile, then MovieLens recommends movies that match a given profile.

Definition of the task:
You are to investigate whether it is possible to predict the rating of a movie on the basis of a given user profile. To answer this question: Develop a prediction model which is capable of predicting the rating of a movie by any given user. Your model must be a non-trivial one, where a trivial model is one that simply predicts averages. The model that you implement should have an RMSE score that is lower than the RMSE's produced by trivial models.
You first have to create a training and a test set through random selection of data (ensure that the entries in the training and test set are unique), then create a model based on data available in a training set then apply the model to a test set in order to predict the movie ratings.
Datasets and corresponding description can be found via the following link:
https://grouplens.org/datasets/movielens/
You are to use the ML20M dataset for this project.
Your model should incorporate descriptive features (where possible) of the movies. A movie database containing many different attributes about movies can be accessed via the following link:

http://www.imdb.com/interfaces

Requirements:

1. Present a general description of the dataset and present the general properties of the dataset.
2. Describe the problem and the methods that you used for this task. Discuss the strengths and limitation of these approaches.
3. Present, explain, and analyse your results.

4. Offer a discussion on (the suitability of) alternative approaches that may have been considered for this task.
5. Summarize: What new and interesting things did you discover while working on this project?