# Final Project

# Topic 1 – Prediction on movie rating

In final project (topic 1), we have a IMDB movie dataset, which includes thousands of movies. In the training data, all movies have been rated, ranging from 1-10. The goal of this project is to analyse data, build a machine learning model on the training data, and predict rates for movies in the testing data.

The final grade will be a combination of data processing, exploratory data analysis, machine learning model building, implementation and report writing.

**This task is for students who have even number student ID, e.g. xxx1232.**

## Data

This data set is in attributes-in-columns format, comma-separated values. It contains two files:

IMDB\_train.csv: training data, consisting of 5,000 instances with 22 attributes.

IMDB\_test: test data, consisting of 1,000 unlabelled instances with 21 attributes.

## Requirements

1. Data pre-processing
   1. Missing data – remove instances that contain missing values
   2. Duplicated data (with the same movieID) – remove duplicated instances
   3. Invalid rating – remove movies that got less than 5 votes
   4. Non-related attributes – remove attributes that are not related to class (or not help you to improve the prediction)
   5. Transform categorical variables into numeric variables (since this is a regression problem)
2. Exploratory data analysis
   1. Distribution analysis
   2. Correlation analysis
   3. Use scatterplot, histogram and boxplot, etc. to illustrate your data
   4. …
3. Prediction
   1. Build machine learning models on training data
   2. Predict rates for testing data
   3. (Optional) build multiple machine learning, find the best one and apply it to the testing data. Hints: cross-validation, train/validation set split

## Submission

* Project report (50’) in **in English**
  + **Submission – report in MS word**
  + The report has been written
  + The length of the project report must be a minimum of 1,500 words
  + The structure of the report:
    - Introduction
    - Data pre-processing
    - Exploratory data analysis
    - Prediction (introduce your algorithm, present your results, discussion on them)
    - Conclusion
  + Document each step
  + Add plots to explain
  + Discussion on your models and results
* Code (30’)
  + **Submission – source code** **in correct format (e.g. .py, .R and .m, etc.)**
  + Code can be in any programming language (e.g. Python, R and Matlab, etc.)
  + Code has been well documented
* Predicted result (20’)
  + **Submission – results in a CSV file**
  + The prediction results for the testing data
  + The file should contain two columns: movieID and predicted rates