# Introduction

The goal of this project is to gain experiences in data preparation.

# Spec

## The IMDB Dataset

Source: <https://www.imdb.com/interfaces/>

**title.ratings.tsv.gz** contains title id to rating mapping. However, the features are structured and not flattened in a vectorized form.

## Part 1: Data Set Generation 15%

Join **two** other sources (e.g. title.crew.tsv.gz, title.basics.tsv.gz, etc) to make some meaningful features for rating prediction.

## Part 2: Modeling 15%

Train/Dev/Test random split = 80/10/10

Model the problem as 3-class multiclass classification - split points are 3.3 and 6.6.

The grade is given based on the accuracy:

>60%: 9 points

>65%: 10 points

>70%: 11 points

>75%: 12 points

>80%: 13 points

>85%: 14 points

>90%: 15 points

# Grading

The assignment takes up 30% of the final grade.

# Submission

Please submit the following files through canvas

* Code
  + A notebook file (or link)
  + A saved notebook with everything executed. (please use this option if you cannot run your code on Google colab.)
* 100 samples of each processed data set.
* Train/Dev/Test accuracy.

Deadline: 11/1/2019

Suggested submission date: 11/1/2019