

## Task 5 – Recommender Systems

The aim of the work is to implement and evaluate an object-object collaborative filtering recommendation system for movies. The implementation can be done in any programming language you prefer (python is recommended). You will use MovieLens data (<https://grouplens.org/datasets/movielens/latest/>) and specifically the small set with 100000 ratings (only the ratings file is required).

1. As a pretreatment
  - filter movies to keep only those that have at least 5 ratings.
  - filter users to keep only those who have given at least 5 ratings.
2. To implement the system in terms of similarity you will use adjusted cosine (or pearson). To evaluate the system, the data will be divided into two parts: training set and control set. The control set will consist of **10% of the data** (after preprocessing) and will be retained **Stable**.
3. To implement the system in terms of the prediction function, the following options will be implemented:
  - Average
  - weighted average
  - weighted average in which, however, the weighting will be based on the number of of common users who have rated the two objects and not on the similarity of the objects as calculated by the adjusted cosine. Define weighting function of your choice.
4. Your program should accept as a parameter the number of nearest neighbors K based on which the prediction should be made as well as the percentage of the training set.
5. The evaluation measures will be the mean absolute error (MAE) and the precision (precision) and recall (recall). For the calculation of the binary valuation measures, a movie is considered relevant if its score is  $\geq 3$ .

Perform the following experiments:

- - For a fixed 90% training set and for 5 K values compare the 3 prediction functions.
- - Using the best K for each method as obtained from the previous experiment, compare the 3 methods with a training set: 50%, 70%, and 90% of the original data set. For control we always use the same 10%.

You will deliver:

- a) your code
- b) a report in which you will present and comment on your results.

In the report against your code, comment out any optimizations or assumptions you made about anything not specified by the utterance.

**Note:** If you computationally experience a problem with the size of the data, reduce it (state the new file sizes) E.g. preprocessing with a threshold greater than 5.