

# 2020 Massive Data Analysis Term Project

Deadline: **2021/1/1(五) 23:59**

Recommendation System: **Item-item** Collaborative Filtering

From lecture **Recommendation Systems P.27 ~ 30**

Similarity: cosine similarity with subtract mean

$$\text{sim}(\mathbf{x}, \mathbf{y}) = \cos(\mathbf{r}_x, \mathbf{r}_y) = \frac{\mathbf{r}_x \cdot \mathbf{r}_y}{\|\mathbf{r}_x\| \cdot \|\mathbf{r}_y\|}$$

output format:

(item, item), similarity

**以上為基本分**

## Rating Predictions

Select top 10 similarity to calculate the movie rating for each user.

i.e.,  $N = 10$

$$r_{xi} = \frac{\sum_{j \in N(i; x)} s_{ij} \cdot r_{xj}}{\sum_{j \in N(i; x)} s_{ij}}$$

output format:

(user, item), rating

**以上為更高分**

## Input Data

From MovieLens: <https://grouplens.org/datasets/movielens/>

It contains 610 users and 9742 movies.

format:

(userID, movieID, rating)

## Other Topics

**Select one of the algorithms from Chapter 10**

You can get your dataset from any source on the Internet, ex:

Facebook, Instagram, YouTube, PTT, or any other dataset on the Internet etc.

Remember to describe the dataset you used in your report.

Filename:

Term\_Project\_Groupnumber.java or

Term\_Project\_Groupnumber.ipynb

Report:

Term\_Project\_Groupnumber.pdf

Describe what you done in your term project, includes your code and some explanation. Also, describe your dataset.

Pack the above files in Term\_Project\_Groupnumber.zip and upload to ilms.

Demo:

From 2021/1/4 ~ 2021/1/8 at 資電館 831

**Caution:**

Except web crawler, you have to complete Term Project with Hadoop or PySpark. i.e., you have to use MapReduce to complete this Term Project.