DSA4212 Year 2022-2023
Assignment No: 2

Deadline: 23:59, 9th of April 2023

1 Task

The three following files

1. `assignment_2_anime.csv`
2. `assignment_2_ratings_train.csv`
3. `assignment_2_ratings_test.csv`

contain information on user preferences data on 12,294 anime. The purpose of this assignment is to design an anime recommendation system.

1. In a first stage of this assignment, you are asked to build models that do not rely on the ideas of *factor models* as described in class. These models can be simple, or more complex. You are encouraged to experiments with approaches not discussed in class.
2. In second stage, you are asked to study whether (some variations and extensions of) the ideas discussed in class can help improve the performance of your recommendation system.
3. As a third and optional stage, I can suggest that you try to improve further your recommendation system by any way you deem appropriate (eg. scrape data online).

Remark:

1. For building your model, you are allowed to use the list of users and films contained in the test set. But you are not allowed (indeed) to use any of the ratings contained in the test set. In other words, you are allowed to use the first two column of the test dataset in any way you deem appropriate (but not the third column).

2 CANVAS Submission

There are (at least) 2 files to submit:

1. A pdf report. This report should not include any Python code. Instead, it should give an overview of your recommendation system and its performance. It should be at the very most 5 pages, but can also be significantly shorter (ie. do **not** write a long report, just for the sake of writing a long report).

2. A Jupyter notebook describing some (not necessarily all) of the experiments that you have performed. This can be split in several notebooks if necessary. This Jupyter notebook should be reproducible: anyone should be able to run it from scratch.

For submitting your work, you will:

1. Zip all your files into a **single zip-file**

2. Use the naming convention GROUPXX.zip where XX is your 2-digit group number (i.e. 01, 02, etc...).

3. Make sure that the pdf-report includes the name and student number of all the students in the group.

4. Upload the file on CANVAS.

Do not include anything else in the zip-file except the pdf report and the jupyter notebooks.

2.1 Grading

The following components will be taken into account:

1. **[30%]** Final test Mean Squared Error, as compared to other DSA4212 groups.

2. **[30%]** Clarity and reproducibility of the Python code and pdf report

3. **[30%]** Quality and appropriateness of the numerical experiments

4. **[10%]** Proper citation and acknowledgement of resources used (eg. books, github code, articles, blog-posts, Kaggle code). For example, if your approach relies strongly on some code, or blog-post, found online (which is perfectly acceptable), please do mention it. *Failure to properly acknowledge the sources and material used will be considered plagiarism.*

3 Remarks:

Please make sure that:

1. the final accuracy that you are reporting is evaluated on the test set.
2. you do not use in any way the ratings contained in the test dataset during the training your algorithm.
3. some of the users are only present in the test set: you need to find a way to deal with this situation (that does often happen in practice)