RMIT UNIVERSITY

Practical Data Science with Python
COSC 2670/2738
Assignment 3

	Assessment Type	Individual
	Due Date	23:59 on the 15th of June, 2022
	Marks	30

Please read this carefully before attempting

This is an *individual* assignment. You may not collude with any other people, or plagiarise their work. You are expected to present the results of your own thinking and writing. Never copy other student's work (even if they "explain it to you first") and never give your written work to others. Keep any conversation high-level and never show your solution to others. Never copy from the Web or any other resource. Remember you are meant to generate the solution to the questions by yourself. Suspected collusion or plagiarism will be dealt with according to RMIT policy.

In the submission (your PDF file) you will be required to certify that the submitted solution *represents your own work only* by agreeing to the following statement:

> *I certify that this is all my own original work. If I took any parts from elsewhere, then they were non-essential parts of the assignment, and they are clearly attributed in our submission. I will show I agree to this honor code by typing "Yes":*

Introduction

In this assignment, you are given a specific data science problem and one (more) potential solution(s). You are required to implement the (these) solution(s), then complete the required tasks (as detailed below) successfully.

The "Practical Data Science" Canvas contains further announcements and a discussion board for this assignment. Please be sure to check these on a regular basis – it is your responsibility to stay informed with regards to any announcements or changes. Login through https://rmit.instructure.com/.

Where to Develop Your Code

You are encouraged to develop and test your code in two environments: **Jupyter Notebook on Lab PCs** and **Anaconda 3 that was suggested in the course canvas announcement**.

Jupyter Notebook on Lab PCs

On Lab Computer, you can find Jupyter Notebook via:

Start → All Programs → Anaconda3 (64-bit) → Jupyter Notebook

Then,

- Select New → Python 3
- The new created '*.ipynd' is created at the following location:
  - C:\Users\sXXXXXXX
  - where sXXXXXXX should be replaced with a string consisting of the letter "s" followed by your student number.

Academic integrity and plagiarism (standard warning)

Academic integrity is about honest presentation of your academic work. It means acknowledging the work of others while developing your own insights, knowledge and ideas. You should take extreme care that you have:

- Acknowledged words, data, diagrams, models, frameworks and/or ideas of others you have quoted (i.e. directly copied), summarised, paraphrased, discussed or mentioned in your assessment through the appropriate referencing methods
- Provided a reference list of the publication details so your reader can locate the source if necessary. This includes material taken from Internet sites. If you do not acknowledge the sources of your material, you may be accused of plagiarism because you have passed off the work and ideas of another person without appropriate referencing, as if they were your own.

RMIT University treats plagiarism as a very serious offence constituting misconduct. Plagiarism covers a variety of inappropriate behaviours, including:

- Failure to properly document a source
- Copyright material from the internet or databases
- Collusion between students

For further information on our policies and procedures, please refer to the following: https://www.rmit.edu.au/students/student-essentials/rights-and-responsibilities/academic-integrity.

General Requirements

This section contains information about the general requirements that your assignment must meet. *Please read all requirements carefully before you start.*

- You *must* include a plain text file called "readme.txt" with your submission. This file should include your name and student ID, and instructions for how to execute your submitted script files. This is important as *automation* is part of the 6th step of data science process, and will be assessed strictly.

- Please ensure that your submission follows the file naming rules specified in the tasks below. File names are case sensitive, i.e. if it is specified that the file name is `gryphon`, then that is exactly the file name you should submit; `Gryphon`, `GRYPHON`, `griffin`, and anything else but `gryphon` will be rejected.

Overview

Although k nearest neighbour (KNN) based Collaborative Filtering method has been proved to be very successful, it still faces some challenges, including the presence of the missing values (also called the data sparsity problem) and the estimation of similarity between users or items. Luckily, data scientists and researchers have been working hard to further improve the k nearest neighbour based Collaborative Filtering method.

Specifically, an existing solution is proposed to further improve KNN-based Collaborative method by handling the data sparsity problem and utilizing the item's popularity to calculate the similarity between users. This existing solution is presented in a report named *"A New Collaborative Filtering Approach Utilizing Item's Popularity"*. Please read this report carefully, then complete the following tasks.

Tasks

Task 1: Implementation

In this task, you are required to implement the solution in the provided report to further improve user KNN-based Collaborative Filtering method.

Please note that we made some changes to the report to suit the assessment for this course. Specifically, it is on page 2 of the report (as show in Figure 1). Namely, you are required to implemented this change in your solution.

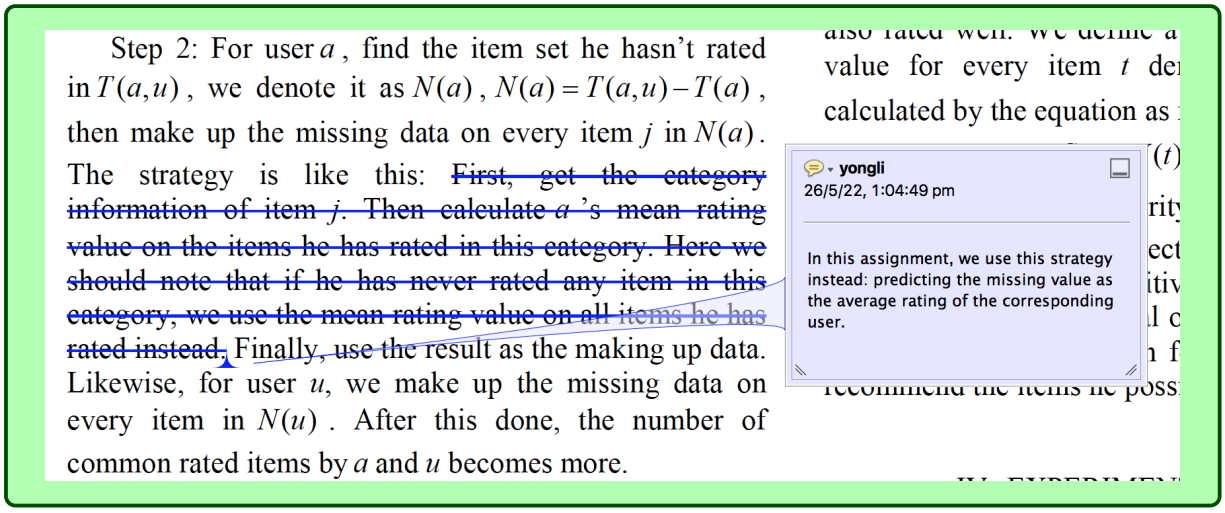

Figure 1: The Change in the Provided report

Note, you are required to implement your own implementation, and please do not use any other libraries that are related to Recommender Systems or Collaborative Filtering. If you use any of these libraries, your implementation part will be invalid.

We provide Python framework code (named *assignment3_framework.ipynb*) to help you get started, and this will also automate the correctness marking. The framework also includes the training data and the test data.

Please read the comments in the provided assignment framework carefully, and only put your own code in the provided cell as indicated (also as shown in Figure 2). Please DO NOT CHANGE anything else in the rest cells of the framework, otherwise they might cause errors during the automatic marking.

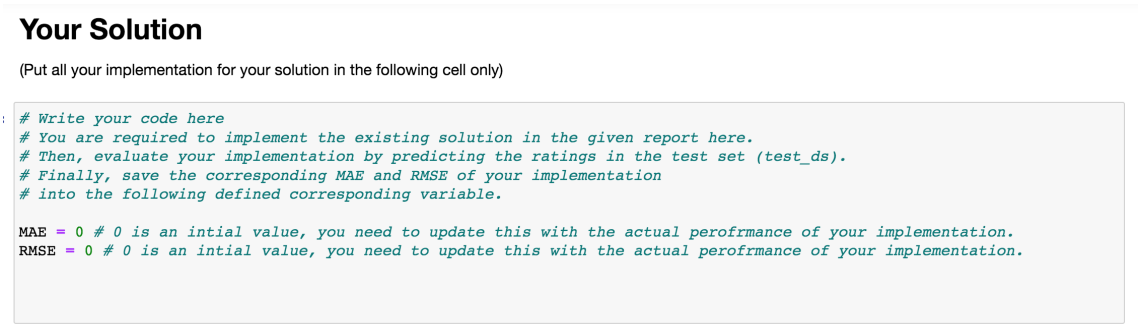

Figure 2: Where to put your implementation in the provided framework (*assignment3_framework.ipynb*)

Please provide detailed comments to explain your implementation. To what level of details should you provide in your solution? Please take the comments in the *ipynb* files in Week 10 (*knn_based_cf_updated.zip*) as examples for the level of detailed comments you are expected to put for your solution.

You might find the following information uesful:

- https://www.w3schools.com/python/python_comments.asp
- https://numpy.org/doc/stable/reference/generated/numpy.log.html
- https://numpy.org/doc/stable/reference/generated/numpy.union1d.html

Task 2: Presentation

- The presentation should
  - Explain how the solution in the provided report can improve the performance of the user KNN-based Collaborative Filtering method by using your own language clearly and completely.
  - Explain why the solution in the provided report can improve the performance of the user KNN-based Collaborative Filtering method by using your own language clearly and completely.
  - Explain how you implement the solution clearly and completely.
- The presentation should be no more than 10 minutes.

- Your presentation slides should be:
  - Microsoft PowerPoint slides (with audio inserted for each slide by using: `Insert` $->$ `Audio` $->$ `Record Audio`).
  - or you can create your own presentation slides (e.g. PDF version) and please submit your own recording (in the format of mp4 or avi) of your presentation as well.

What to Submit, When, and How

The assignment is due at

23:59 on the 15th of June, 2022.

Assignments submitted after this time will be subject to standard late submission penalties.

The following files should be submitted:

- Notebook file containing your python implementation, 'Assignment3_framework.ipynb'.

For the notebook file, follow these steps before submission:

1. Main menu → Kernel → Restart & Run All
2. Wait till you see the output displayed properly. You should see all the data printed and graphs displayed.

- One of the following:
  - Your `Slides.pdf` file and your presentation recording in the required format. Or,
  - Your Microsoft PowerPoint slides (with audio inserted for each slide).
- The "readme.txt": includes your name and student ID, and instructions for how to execute your submitted script files.
- Please note: there is no need to submit the data sets, as you are not allowed to change them.

They must be submitted as ONE single zip file, named as your student number (for example, 1234567.zip if your student ID is s1234567). The zip file must be submitted in Canvas:

Assignments/Assignment 3.

Please do NOT submit other unnecessary files.