PROCESS 2 - ESSAY

Course: Mining Massive Data Sets

Duration: 03 weeks

I. Formation

- The project is conducted in groups with 03-05 students.
- Student groups conduct designated tasks and submit the project by the given deadline.

II. Requirements

1) Task 1 (8.0 point(s)): Collaborative Filtering

Implement the task in **Task01.ipynb.**

Data

Data sets	Description
ratings2k.csv	Product rating data set. The first line is the header. • **index**: row index • **user**: user ID • **item**: product ID • **rating**: rating (0.0-5.0) 2365 remaining lines are data samples.

Algorithms

- Implement the **Collaborative Filtering** algorithm to recommend items for individual users using *PySpark*. It should be a class for future deployment.
- The similarity between users is measured using the *Pearson correlation coefficient*.
- The constructor takes in the value of N (number of similar users) and the data set as a data frame (*PySpark*).
- The function *predict*() takes in a user (a vector of ratings) and the expected number of recommended items. It returns a data frame (*PySpark*) consisting of recommended items sorted in the descending order of scores.

Experiments

- Load the given dataset to construct the utility matrix and then infer "profile" vectors for users and items.
- Divide the dataset into a training and test sets with the ratio of 8 : 2.
- Evaluate the algorithm in the test set with *N* in the range [2, 16]. After that, draw a bar chart to illustrate the *RMSE* values for each *N*.

2) Task 2 (2.0 point(s)): Report

- Student groups compose the project report using the IEEE conference proceeding template.
- Recommended editor: Overleaf.
- Selective contents:
  - *Title*: the project title
  - *Authors*: group member's information, the lecturer is appended as the last author.
  - *Abstract*: summarize the project requirements, approaches, experimental results, and levels of completion.
  - Each following section presents a task in the project, with a meaningful and human-readable title. Briefly introduce the approach to tackle the problem and illustrate results with related figures/tables, etc.
  - "*Contributions*" section: individual tasks, individual completion levels (0%-100%).
  - "*Self-evaluation*" section: self-evaluate task completion and estimate scores.
  - "*Conclusion*" section: summarize the project requirements, approaches, experimental results, and levels of completion.
- References are in the IEEE format.
- Maximal length is 05 pages.

III. Submission Notice

- Create a folder whose name is like

process2_<Group ID>_<your student ID>

- **Source/:** consists of the project source code, each task is implemented in an individual sub-directory, preserving the outputs of all cells in ipynb files, output files as well.
- **Report/:** report source (exported from Overleaf), **report.pdf** file.

- Compress the folder as a zip file and submit by the deadline.
- Every team member must submit the project individually.

IV. Policy

- **Student groups submitting late get 0.0 points for each member.**
- **Copying source code on the internet/other students, sharing your work with other groups, etc., cause 0.0 points for all related groups.**
- **If there exist any signs of illegal copying or sharing of the assignment, then extra interviews are conducted to verify student groups' work.**
- **Evaluation scores of individual tasks are only recorded if and only if the student group give a reasonable presentation and justification to avoid cheating by AI tools, rental of doing the project, imbalance contributions, missing discussing, cooperating of group members in the project, etc.**
- **AI tools are forbidden in the project.**

-- THE END --