Term Project
Recommendation System with Spark

Find a rating dataset for recommendation systems (other than MovieLens 100K rating dataset, this will be used in lecture),

- Project teams are declared as "Table of project group" in classroom page. Please obey it.
- Present/show your computer info (with code) IP address, name and configuration of your PC etc.
- Transform the data by splitting each record into fields.
- Visualize your data for your manager (important fields of your data, age group, most rated items, item categories, age group etc. – according to you); create histogram, bar chart etc.
- Build recommendation engine using collaborative filtering.
- Use %70 of dataset for training, rest of them for testing.
- Use ALS (Alternating Least Squares) for training recommendation model with last 4 digit of your student number as a "seed". Also change the parameters of ALS re-run the algorithm for parameters "rank" (10, 50, 200), "iteration" (10, 50, 200) and "lambda" (0.01, 0.1). This means 18 different model will be created using specified rank-iteration-lambda values
- Find and present MSE (Mean Squared Error), RMSE (Root Mean Squared Error) for performance evaluation of each model and explain them, indicate best model for your dataset, explain why.
- Return your code as "html" and "ipynb" file.

For datasets, you may check: https://www.kaggle.com/

YOU SHOULD SUBMIT FOLLOWING:

- Named your python file as **studentNo1_studentNo2** (Student number of each group member should be added to file name) Save the python file as a. ipynb and **.html** format. **Send both of them.**

- **Send a report (word file (.doc) or pdf)** which contains
    - Student number and student name of each group member.
    - Specify your work sharing policy (which member completed which part) in your report.
    - Your computer info (with code) IP address, name and configuration of your PC etc.
    - Report is similar to lab sheet – lab/log book/hands-on steps of your project, including with all codes, final results, outputs and screenshots related to your project.
    - Explain your dataset, show couple of records as a sample and indicate the fields.
    - Explain the result of MSE (Mean Squared Error) for performance evaluation of your model for each run, indicate best model for your dataset.

- **Send your dataset (or link).**

<u>NOTES</u>

- You will present /perform live demonstration of your project in online meeting.
- There will be Question- Answer section during demonstration of your project. You should be prepared.
- Each member of team should submit the project-report and dataset to classroom.