Build a recommender system with Spark

In this project, you are going to experience how to implement a simple recommendation system with PySpark and do epxeriments on datasets of different scales. Work with your group members to fulfill the following requirements.

1.   Study a state-of-the-art (SOTA) collaborative filtering approach and present your findings.

-    The SOTA algorithm should be published with 5 years by reliable organizers.

-    Your findings should cite the followings from the original material: an overview of the algorithm (pseudo-code, demonstrative figures of the proposed model, et.), experimental data, and discussions on its pros and cons.

2.   Implement a recommender system and study its performance on the standard benchmark.

-    The benchmark dataset is MovieLens 1M movie ratings.

-    You need to use the ALS algorithm in Spark MLlib.

-    The evaluation metrics are Root Mean Squared Error, Hit rate and NDCG.

3.   Study the performance of the recommender systems built above on another benchmark, which is different from the standard one in term of data size (should be bigger) and/or sparsity.

You have to prepare the following materials for your group submission.

•   The link(s) to the Google Colab implementation. The datasets must be prepare in advance so that the grader can run the implementation with minor adjustments. You also need to guanratee that no edit is made after the deadline.

•   A report that addresses Question 1 and presents evaluation results in Question 2 and 3

•   A video recording that clearly demonstrate how to run the implementation and the results obtained for each part of the assignment.

Important notes:

•   This project gives you 20% course grade.

•   Strictly avoid plagiarism in any circumstance.

Nguyễn Ngọc Thảo – Lê Ngọc Thành – Phạm Trọng Nghĩa   *HCMUS*

Nguyễn Ngọc Thảo – Lê Ngọc Thành – Phạm Trọng Nghĩa   *HCMUS*