MDLE: Assignment 3B

– Due date: June 5th, 2025 –

Implement the solutions to the following exercise using Spark. Submit a documented Jupyter notebook, a python script to run through spark-submit, and the results of the algorithm. If the results are too large, submit a download link instead.

The comments should explain the main steps of the solution with sufficient detail.

Implement a CF algorithm, using the item-item approach, to recommend new movies to users.

Use the MovieLens dataset, available from:
https://grouplens.org/datasets/movielens/

Start with the 100,000 ratings (Small) dataset, and afterwards try to apply your methods to the larger datasets (1M, 10M, 20M, 25M).

You will need to implement an efficient approach for finding the near neighbors needed for predicting new rating (either LSH or clustering).

Validate your method by leaving out 10% of the available ratings.