MLSD: Assignment 2
Clustering
Collaborative Filtering

– Due date: May 4, 2023 –

For each of the following exercises, you should implement the solutions using Spark. Use small samples of the dataset for developing and initial testing, then run on the full data.

What to submit
For each exercise, 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.

1. Implement the BFR algorithm and apply it to the FMA dataset.

   The FMA dataset consists of 106,574 music tracks represented by 518 features, corresponding to 7 statistics (mean, standard deviation, skew, kurtosis, median, minimum, maximum) calculated from 74 time-based acoustic characteristics. See the dataset description for more details.

   The first three lines in the features.csv file identify the feature in each column: the first line indicates the feature group, the second line the statistics, and the third line the feature number within each feature group. Some examples of feature identifiers are then 'tonnetz_skew_06', 'chroma_cqt_kurtosis_09', 'mfcc_max_08', 'spectral_contrast_median_07'.

   1.1 Apply an in-memory agglomerative hierarchical clustering algorithm to the small subset (8000 songs) and calculate the radius, diameter and density (using r^2 and d^2) of all clusters for values of k between 8 and 16.

   Note: You can integrate any existing implementation of hierarchical clustering.

   Note: To select the small dataset, filter by the field subset in the tracks.csv file.

   1.2 From the results of 1.1, select the best number of clusters k and cluster the complete dataset using the BFR algorithm.

   1.3 Using the songs metadata (tracks.csv), create a (preferably visual) representation of the cluster in terms of the most common music genres found in the cluster.

2. 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 neighbours needed for predicting new rating (either LSH, clustering or matrix decomposition).

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