INFO-H600 Computing Foundations of Data Sciences
Project: Million Playlist Dataset

Data

In this project, you will implement a data processing pipeline for analyzing music playlists. We will use the Million Playlist Dataset published by Spotify, discussed here: `https://www.aicrowd.com/challenges/spotify-million-playlist-dataset-challenge`, and available to download here: `https://www.kaggle.com/datasets/himanshuwagh/spotify-million/data`.

The dataset contains one million playlists, where each playlist contains information about the playlist itself (e.g., playlist title, duration) and each track it contains (e.g., track id, artist id, track name, duration). The dataset was used as part of a "playlist continuation" challenge: given a playlist name and a few tracks, predict which other tracks would fit best with that playlist.

Tasks

In this project, you are asked to address the following tasks.

1. Present your approach to handling this big dataset. Given that the dataset is big, at about 35 GB, you are not expected to be able to process all of it. Ideally, of course, you should have a solution that scales to the whole dataset. Describe the approach you took and all the alternatives you considered.
2. Present some interesting statistical/aggregate information about the dataset. Explain how you derived the presented information.
3. Propose a definition of similarity between tracks. Specifically, given two tracks, their similarity is a score from 0 to 1, where 0 indicates that the two tracks are not similar at all, and 1 indicates that the two tracks are identical. Design a method to compute those similarity scores using the dataset. Discuss the implementation and give some examples.
4. Propose a definition of similarity (score ranging from 0 to 1) between playlists. Design a method to compute this similarity score using the dataset. Discuss the implementation and give some examples.
5. Propose a method to solve the "playlist continuation" challenge. Discuss the implementation and give some examples.

You should write the code in Python, and you are free to use the tools and frameworks (e.g., Spark, Dask) that we have discussed in this course as well as others.

Submission

This project is to be made in **groups of six persons**. You are asked to form the groups via the activity "Groups for Project" on UV.

You should prepare a **report** (pdf document) describing your approach to the different tasks. You should submit a *single* zip file containing the report and well documented code. The single file has to be uploaded on UV by **December 12, 2025**. There should be one submission per group.

In addition, you should prepare a short (max. 10 minutes) **presentation** of your project solution, possibly including a live demo. You will be asked to make the presentation during the last week

of lectures (**December 15–17, 2025**). Choose a presentation date using the "Select Day for Project Presentation" link on UV.