Project

Context

Music streaming services have changed the way we listen to music. Platforms like Apple Music and Spotify allow us to enjoy music anytime and anywhere. You can create your own playlists, share them with your friends, and even share your subscriptions with family.

As a data analyst team at Spotify, you are responsible for managing multi-user accounts, also known as family accounts. These accounts are associate with a single subscription for multiple users, who retain the ability to create their own playlists and access personal music statistics.

On the morning of November 21, Spotify's servers were attacked by hackers. The hackers deleted users' Top-of-the-Year playlists from family accounts. These users find themselves faced with a huge playlist that mixes everything. If users on the same family account have exactly the same musical taste, that's fine. But who would listen to exactly the same thing as their parents or roommates?

Fortunately, you were able to recover some of the deleted data from Spotify's servers. The recovered data contains parts of the Top-of-the-Year playlists of different users. But some songs from the huge mixed playlist are still missing from the recovered data.

Your mission is to reconstruct each user's Top-of-the-Year playlists for each year.

Data

First, you have the single playlist that mixes everything:

- `mixed_playlist.csv`
- Some songs in this CSV file have already been labeled with the corresponding user and top year according to the recovered data.
- Some songs are missing from the recovered data, so the user and top year remain unkown.

Then, you have the recovered parts of the Top-of-the-Year playlists of different users.

- CSV files in the `recovered_data` folder.
- These playlists are not complete.
- You need to reconstruct each playlists.

For each song, your team created a number of characteristics to describe this song. These characteristics can also be found in every CSV file.

Work to be done

- Work by groups
    - **3 or 4 people per group**
- Reconstruct the Top-of-the-year playlists for different users.

- As a data analyst, your team also need to develop algorithms that recommend the next song to users. The song the user is currently listening to is an important reference. The user's music taste in previous years can also be used as a basis for recommendations. Please suggest several characteristics of song that can be used as input features for the recommendation algorithm and justify your proposal.
    - This part should be presented at the final session.

Deliverable

- The jupyter notebook
    - The program that you developed for reconstructing the playlists.
    - The proper comments of the program.
    - Please run all the cells of the notebook before submission.
- The reconstructed playlists
    - A `.zip` file composed of multiple `.csv` files.
    - Each `.csv` file is a Top-of-the-Year playlist of a specific user at a specific year.
- The slides you will present at the final session.
    - You will need to give a **10 minutes presentatio**n at the final session (06/12/2023).
    - Explain how you frame the problem, process the data, build the machine learning model and solve the problem.
    - Present your proposal for the recommendation algorithm and your reason.

Due date

All the three deliverables should be submitted on Learn **before 05/12/2023 23:59**.

Each group only needs to submit **once** by a group member.

Please indicate the names of group members in the submission.