Data Processing Final Project: Project Description

The focus of my project was to analyze and utilize a database I created, which consists of 5370 Rock/Metal songs. This database includes information about the songs, sound analysis of these songs, albums these songs belong to and the artists who created them. The primary objective was to answer two main research questions: 'What are the characteristic features of the database I created?' and 'What are the most similar songs in my database to a given song?' Through this project, I aimed to gain insights into the data and develop a song recommendation system based on audio analysis data.

The first step was to compile a comprehensive playlist of 5370 Rock/Metal songs on Spotify. Using the Spotify API and the Spotipy library, I extracted detailed information about these songs, their respective albums, and the artists. This data was then structured into a SQL database, consisting of four main tables: artists, albums, tracks, and audio analysis. Ensuring the database was well-structured was crucial for efficient data retrieval and analysis.

To address the first research question 'What are the characteristic features of the database I created?' I used SQL queries to extract relevant data from the database. This data was then visualized using Python libraries such as Matplotlib, Pandas, and Seaborn. Through these visualizations, I provided a clear and comprehensive overview of the dataset. Key visualizations included: The most frequent artists in the database, the most popular artists and songs, genre analysis, top and least songs based on audio analysis data. The visualizations provided a detailed understanding of the database's characteristics. They highlighted key patterns and trends, such as the dominance of certain artists and the popularity of specific songs.

The second research question focused on identifying the most similar songs to a given song within the database. To achieve this, I first analyzed the audio features of the songs. Using feature engineering techniques, I transformed the audio analysis data to ensure a Gaussian distribution, which facilitated more effective comparisons. I also removed outliers to improve the accuracy of the recommendations. The core of the recommendation system was based on cosine similarity, a statistical method used to measure the similarity between two vectors. By comparing the audio features of songs, the algorithm was able to recommend songs. Initial results were promising but required further refinement to enhance accuracy. The recommendation algorithm, after incorporating feature engineering and data transformation techniques, showed a significant improvement in suggesting similar songs.

Through this project, which involved creating a database, analyzing this database, and developing an algorithm using the data within it, I gained new knowledge and experience in many different areas.

Here is the playlist that I created on Spotify:
https://open.spotify.com/playlist/6SefiumTj6DBQPgP5i9Wxt

Melik Kaan Şelale

15362175