Data Collection | Processing

Group 14

Aditya Ratna

Umang Mistry

617-642-8352 (Aditya Ratna)

571-430-5577 (Umang Mistry)

ratna.a@northeastern.edu

mistry.u@northeastern.edu

Percentage of Effort Contributed by Student 1: 50%

Percentage of Effort Contributed by Student 2: 50%

Signature of Student 1: Aditya

Signature of Student 2: Umang

Submission Date: 10/04/2021

Data Collection Steps

1. Importing the required libraries for the data collection process.
2. Setting credentials to allow us to pull data from the API.
3. Passing list of the required artists. This is going to be from the user's music history.
4. Extracting each artist's unique (uri).
5. Using the uri's. We then used spotipy's method sp.artist_albums() to get the info about all albums of the artist in Spotify's database.
6. Defining function for adding album info like songs and track uri to a dictionary. This will be called recursively for each album.
7. Calling the function to create the required dictionary. Spotify_albums { }.
8. Defining function to extract audio features of all the songs from each album we need to train the model.
9. Running the extract audiofeatures() function on each album's every song and updating the dictionary accordingly.
10. Creating a dictionary to add all the information of the audio features and track info that we will convert to dataframe for extraction.
11. Converting the dictionary into df using Pandas.
12. Checking if any duplicities exist.
13. Exporting the dataframe as a .csv file.

To obtain a comprehensive dataset for our Recommender system, we would have to scale this for several more artists and playlists. We do not have the computing resources as making Get () requests from the API for 3 artists took us over 10 minutes.

Hence, we found a dataset with 1.2 million rows for our recommendation system. The dataset has no null values. This is consistent with the data that we pulled using the API. If additional data is required, then we will fetch selective data using the API.

Dataset link: https://drive.google.com/drive/folders/1nWFP1p93zER8FEg0E8dKGDZVJd-A4zt2?usp=sharing

Data Processing Steps

1. Checked the data types of all the columns of the dataset.
2. Outliers and IQR analysis of the numerical data type columns namely audio features.
3. Removing outliers from the “loudness” feature.
4. Plotting correlation heatmap for the numerical features.

Observations

We can high positive correlation between:

- Loudness and Energy
- Valence & Danceability

We can observe high negative correlation between:

- Acousticness & Energy
- Acousticness & Loudness