Data Management Plan

Ghazaleh Sadatfatemi – Ali Afsharian

- ## Introduction

Music plays a more important role in our life than just being a source of entertainment. As we are both into music, working on a dataset related to Music makes learning easier for us. This project analyses some datasets from Spotify and we want to use different attributes to predict the genre of the music and find the relations between these attributes to see if the song is going to be popular.

- ## Spotify Dataset

Our considering dataset can be achieved by https://developer.spotify.com/ website. Spotify APIs make it possible for us to access the dataset. What makes this dataset valuable is that there are some useful attributes which make the data analysis of our Dataset easier. We planned to get genre specific playlists for different genres including “Rock”, “Rap”, “Jazz”, and “Country”.

- ## Preparation of Dataset

We need a Spotify account and a python code for authorization to achieve Spotify APIs.
Here we use Spotipy which is a lightweight Python library for the Spotify Web API. With Spotify we get full access to all of the music data provided by the Spotify platform.
We created a function in Python to extract the information of songs (also audio features) in a particular playlist. Simultaneously, the extracted information moves to the MySQL database with the help of MySQL connector.
At below, the information which we extracted are written:

Attribute	Description	Type
artist	The name of the artist.	string
album	The name of the album. In case of an album takedown, the value may be an empty string.	string
track_name	The name of the track.	string
track_id	The Spotify ID for the track.	string
genre	The genre of the song	string
popularity	The popularity of a track is a value between 0 and 100, with 100 being the most popular. The popularity is calculated by algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are. Generally speaking, songs that are being played a lot now will have a higher popularity than songs that were played a lot in the past.	integer
acousticness	A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.	float
danceability	Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.	float

energy	Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.	float
key	The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. If no key was detected, the value is -1.	integer
loudness	The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typically range between -60 and 0 db.	float
mode	Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.	integer
spichiness	Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.	float
instrumentalness	Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.	float
liveness	Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.	float
valence	A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).	float
tempo	The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.	float
duration_ms	The duration of the track in milliseconds.	integer
time_signature	An estimated time signature. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure). The time signature ranges from 3 to 7 indicating time signatures of "3/4", to "7/4".	integer

- **Processing**

  We moved our data to the MySQL database, and genre of the music (based on playlist) to each track.

  Then we separated the features that we need for our processing. (Red attributed in the table above). Also, since the track duration attribute is expressed in milliseconds, we converted it to minutes to make much more sense.

  After preparing our data, we were going to find the relation between popularity of a track and its different audio features. Also, we were trying to classify the genre of different tracks based on their features.