German International University of Applied Sciences Informatics and Computer Science
Dr. Nada Sharaf Eng.
Aya Abdallah
Eng. Omaima Ahmed

Big Data & NoSQL Databases, Spring 2024
Assignment 2
Due date is May 3rd, 2024 at 11:59 PM
Submitted in groups of maximum 2 (can be cross tutorial)

For this assignment, you will be using a dataset that compiles information about different songs including multiple attributes that were streamed on Spotify in 2021. You are required to:-

1. Perform any necessary data cleaning & engineering that renders your data useable (i.e. handling missing values, duplicates, classification, transformation…etc.)

2. Implement the following Queries. You can perform any necessary alterations to the data that may improve the usability of these queries or even make them possible. You should implement the task twice: once per SparkSQL and SparkDataframes.

The 'Genre' variable represents the category of each song such as Rock, Indie, Alt, Pop, Metal, HipHop, Alt_Music, Blues, Acoustic/Folk, Instrumental, Country

a) Which genre has the highest average popularity?

b) Display which artists have recorded the most number of songs with a duration of more than 5 minutes

c) How many songs are included in every Genre?

d) Which artists dominated the charts?

e) Recommend at least 5 fun/not-boring songs that can be played at a party, you can use features like energy, danceability etc.. to represent cheerfulness.

3. Split your data into training and testing sets. Perform classification on the dataset to predict which genre each song would belong to. You are required to use SparkML to apply the classification 3 times, each using a different classification method and compare the accuracies for all 3, pointing out the best classifier based on your results.

Deliverables

- Your code is to be submitted to **bigdata602.s24@gmail.com** as a zip file containing your 2 notebooks (one for SparkSQL and another for SparkDataframe & SparkML) (include the names and IDs of the team members in the body of the email with
subject: “Assignment 2 S24”).

PLAGIARISM IS NOT TOLERATED AND COPIED WORK WILL BE AWARDED 0 POINTS FOR BOTH TEAMS INVOLVED or IF YOU COPIED IT FROM THE INTERNET OR ELSEWHERE (NO. EXCEPTIONS.)!

Note: You will be asked for the reasoning of any actions, queries or decisions you have taken in the implementation of this project during the evaluations that have led to your answers/results

Spotify Audio Features

For every track on their platform, Spotify provides data for thirteen Audio Features. The Spotify Web API developer guide defines them as follows:

- **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.
- **Valence:** Describes 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).
- **Energy:** 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.
- **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.
- **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.
- **Speechiness:** This 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.
- **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".
- **Liveness:** Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live.
- **Acousticness:** A confidence measure from 0.0 to 1.0 of whether the track is acoustic.
- **Key:** The estimated overall key of the track. Integers map to pitches using standard Pitch Class notation . E.g. 0 = C, 1 = C#D, 2 = D, and so on.
- **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.
- **Duration:** The duration of the track in milliseconds.
- **Time Signature:** An estimated overall time signature of a track. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure).