Brandon Liunoras

IST 769

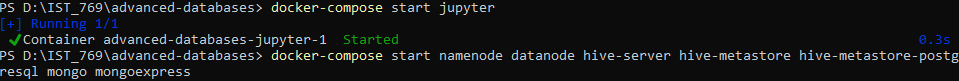
Final Project

YouTube Link: <https://youtu.be/GnYSZbvW0D0>

**Project Scope**

Utilizing Spotipy and Spotify API to extract song data and analyze its attributes. Spotipy is a Python library that grants full access to the Spotify API. The data that will be analyzed are song attributes such as valence, energy, and tempo. The goal for this project is to use the different attributes of the songs of different artists to determine what makes them popular using regression analysis. Afterwards, the data will then be added to Hadoop for future use.

**Docker Commands Used**



**Overall Agenda:**

1. API Preparation
   1. Create an application in Spotify API
   2. Download and import Spotipy
   3. Go through reference information for both applications
2. Data Acquisition
   1. Utilize Docker and Jupyter Notebook for program building
   2. Create module for API calls
   3. Build dataframes and save to csv
3. Analysis
   1. Regression Analysis for Song Attributes
4. Storing
   1. Storing data to MongoDB

**API Preparation**

In this stage, I created a developer profile to gain access to API calls as well as create an app to gain a client\_id and client\_secret. I then downloaded the Spotipy library, imported it into Jupyter Notebook and looked over the important references needed to access certain pieces of data within Spotify.

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

**Data Utilized**

The data pulled from Spotify was song data including ids, album names, song names, special identifiers known as URIs, as well as song attributes such as valence, danceability, and tempo. All the columns that do not provide a key identifier of the data, are columns that provide some type of float64 score. The data that was chosen was selected from five different song artists and all the songs they have made that were put on Spotify were pulled. The following is the module created to pull the data.

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

**Analysis**

For the exploratory analysis, the main component was to look at the different values of each value and see what can be deduced about the artists’ songs. Averages of each attribute were taken and then most were displayed via matplotlib for visual graphs.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A graph of different colored bars

Description automatically generated

As for the regression analysis, a multivariate linear regression model was created to see if and what variables could explain a song’s popularity on Spotify. This process was repeated with each artist. The following results show the r-squared values of each artists’ songs regarding song attributes versus popularity.

* Maluma r-squared = 0.09984551515258588
* Taylor Swift r-squared = 0.296912042382258
* Ice Cube r-squared = 0.10339604740270782
* Ed Sheeran r-squared = 0.08344167274465603
* Beyonce r-squared = 0.29099386768949465

A screenshot of a graph

Description automatically generatedA diagram of different colors

Description automatically generatedSince the values shown are closer to zero than one, the model shows that the attributes shown previously do little to explain what makes an artist’s song popular. After the regression model was executed, heat maps for each artist were created to show correlations between different attributes of songs. The main attribute observed was popularity and its interactions with the other attributes.

**Storage**

After all analysis is made, a dataframe was created to house all the individual dataframes into one. The large dataframe was then saved as a csv file and then transferred to a different environment to be converted from a pandas dataframe to a spark dataframe. From there, the dataframe was loaded into Mongo and then read back out to ensure that it was stored correctly.

A screenshot of a computer program

Description automatically generated

A white background with red and purple text

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

**Conclusion**

The data used was insightful when it comes to learning about song data and how users on Spotify perceive songs. This data can be very useful to derive other information such as creating song recommendations, playlist creations, and ad targeting, all based on what songs are listened to and what their attributes are. Overall, I enjoyed using Spotipy and Spotify API to look at song data and derive new insights.