FindMe FM

Final Project: Group 3

Dorna Abdi
Manuel Elizadi
Christian Koehlinger
Cailyn Miller
Mike Prue
Joshua Rhodes

Project Outline

Topic: Can machine learning predict what songs a user will enjoy based on the audio features of a song they like?

Reason for topic: Interest in the use of Spotify API and other available spotify datasources.

Data Source: Kaggle dataset - Spotify Dataset 1922-2021 ~600k tracks

- Contains info on the audio features of each song (danceability, acousticness, tempo, etc,)
- Dataset is created using the Spotify API

Data Structure: Tracks

Primary:

- ID

Numerical

- acousticness (ranges from 0 to 1)
- danceability (ranges from 0 to 1)
- energy (ranges from 0 to 1)
- duration_ms (ranges from 0 to 1)
- instrumentalness (ranges from 0 to 1)
- valence (ranges from 0 to 1)
- popularity (ranges from 0 to 1)
- tempo (ranges from 0 to 1)
- liveness (ranges from 0 to 1)
- loudness (ranges from 0 to 1)
- speechiness (ranges from 0 to 1)

Boolean

- mode (0 = Minor, 1 = Major)
- explicit (0 = No explicit content, 1 = explicit content)

Categorical

- key: all keys on octave encoded as values ranging from 0 to 11, starting C as 0, C# as 1 and so on...
- timesignature: the predicted timesignature, most typically 4
- artists: the artist(s) who made this song
- artists_ids: the ids for each artist
- release_date : date of when the song was released
- name: title of the song

Data Structure: Artists

Primary:

- id: ID of artist

Numerical

- number of followers: total number of followers the artist has
- popularity: popularity of artists based on all their tracks

Categorical

- name: name of artist
- genres: genres associated with the artist

Data Structure: Dictionary of Artist to Artist Relationships

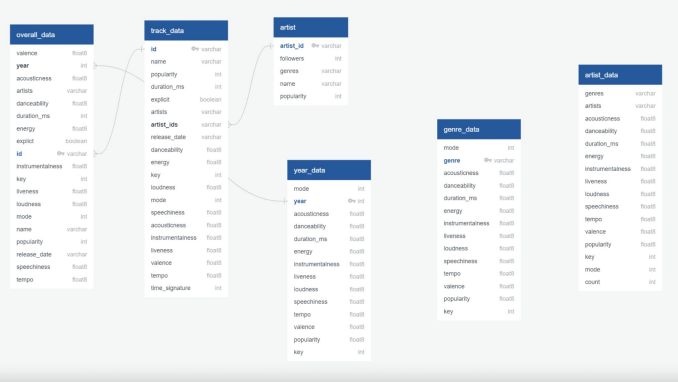
```
{
“any”: [
“first”,
“second”,
“third”,
…,
“nth”
],
“blank:[],
“first:[
“any”,
“third”,
“Second
],
…
}
```

- The lists are in descending order
- “first” - the most similar to “any”,
  “second” - the second most, and so on.
  - max 20 similar artists

Questions We Hope to Answer with Data

Can we use audio elements of a track to predict a song a user would like based on an input of another song they like?

Descriptions of the data exploration phase of the project



id id of track	name name of track	popularity popularity of track				
586672 unique values	446475 unique values	0 100	114030 unique values	19700 unique values	0 0.99	0 1
35iwgR4jXetI318WEWsa1Q	Carve	6	['Uli']	1922-02-22	0.645	0.445
021ht4sdgPcrDgSk7JTbKY	Capitulo 2.16 - Banquero Anarquista	0	['Fernando Pessoa']	1922-06-01	0.695	0.263
07A5yehtSnoedV1JAZkNnc	Vivo para Quererte - Remasterizado	0	['Ignacio Corsini']	1922-03-21	0.434	0.177
08FmqUhxtyLTn6pAh6bk45	El Prisionero - Remasterizado	0	['Ignacio Corsini']	1922-03-21	0.321	0.0946
08y9GfoqCWfOGsKdwojr5e	Lady of the Evening	0	['Dick Haymes']	1922	0.402	0.158
0BRXJHRNGQ3W4v9frnSfhu	Ave Maria	0	['Dick Haymes']	1922	0.227	0.261

Created a mapping between the different data sources

Explored the datatypes

Description of the analysis phase of the project

acousticness	danceability	energy	instrumentalness	liveness	loudness	popularity	speechiness	tempo
1.597267	-1.402608	-0.434803	0.527065	0.681875	-2.560544	0.401476	-0.384211	0.481201
-0.026886	-0.362256	0.666427	-0.519699	0.799534	-1.552361	1.193404	-0.071579	0.280969
3.212922	-1.633991	0.005412	1.118260	0.249359	-1.904048	-0.143154	1.033555	0.752712
-1.105168	-1.309363	-0.297432	0.131040	0.100094	-0.986246	-0.980184	-1.432502	0.749119
-1.037801	-0.886208	1.048483	1.097711	0.860369	-0.453503	0.875575	-1.344493	1.213830

We ran a PCA analysis to understand which audio elements had the greatest variability.

Technologies, languages, tools, and algorithms

K-Means: The Machine Learning Algorithm that is being used to cluster songs together based on elements.

Heroku: Cloud platform being used to store all of data and website.

pgAdmin: Platform being used to store the database.

Flask: Module being used to create the website.

Tableau: Platform being used to visualize the data analysis.

SQL: Language being used to query and restructure the database.

Python: Language being used to clean the data and run Machine Learning algorithms.