ETL Project

Tammy Cronin
Nicole Nguyen

Topic:

To gather datasets from major media platforms such as Youtube, Spotify, MusicBrainz, Last.fm, and Billboard top charts in order to extract different popularity measures throughout the years. We want to gather data regarding top trending songs, artists, and genres.

"Scrobble" is the idea of sending information regarding the music you are listening to to suggest similar music through your service provider. This data may have an effect on music trends.

Extraction

We gathered various datasets from Kaggle and other sources. Our datasets ranged from 360MB to 5MB.

Media Platform	Data Extract
Billboard	- Artist - Song
MusicBrainz and Last.fm	- Artist - Genre
Spotify	- Top 200 - Artist - Song - Stream
Youtube	- Music videos - Views - Artists - Songs

Data Sources:
https://www.kaggle.com/datasnaek/youtube (NN)
https://www.kaggle.com/edumucelli/spotifys-worldwide-daily-song-ranking (NN)
https://www.kaggle.com/pieca111/music-artists-popularity (TC)
https://www.kaggle.com/miteshsingh/hollywood-music-dataset (TC)

Transformation

Our initial steps were to read our data frames and remove columns that were irrelevant to our overall goal.

	Position	Track Name	Artist	Streams	URL	Date	Region
0	1	Reggaetón Lento (Bailemos)	CNCO	19272	https://open.spotify.com/track/3AEZUABDXNtecAO...	2017-01-01	ec
1	2	Chantaje	Shakira	19270	https://open.spotify.com/track/6mICuAdrwEjh6Y6...	2017-01-01	ec
2	3	Otra Vez (feat. J Balvin)	Zion & Lennox	15761	https://open.spotify.com/track/3QwBODjSEzelZyV...	2017-01-01	ec
3	4	Vente Pa' Ca	Ricky Martin	14954	https://open.spotify.com/track/7DM4BPaS7uofFul...	2017-01-01	ec
4	5	Safari	J Balvin	14269	https://open.spotify.com/track/6rQSrBHf7HlZjtc...	2017-01-01	ec

We only wanted data from the US region for measures of popularity

Title	Artist	Streams	Region
Bad and Boujee (feat. Lil Uzi Vert)	Migos	1371493	us
Fake Love	Drake	1180074	us
Starboy	The Weeknd	1064351	us
Closer	The Chainsmokers	1010492	us
Black Beatles	Rae Sremmurd	874289	us

We extracted the various genre type attributed to different artists

Artist	country_mb	Genre Tags
Red Hot Chili Peppers	United States	rock; alternative rock; 80s; 90s; rap; metal; ...
Rihanna	United States	pop; dance; hip hop; reggae; contemporary r b;...
Eminem	United States	turkish; rap; american; hip-hop; hip hop; hiph...
The Killers	United States	synthpop; alternative rock; american; new wave...
Kanye West	United States	synthpop; pop; american; hip-hop; hip hop; ele...

After cleaning the datasets, we merged the csv files to create dataframes based on:
1) Measures of popularity
2) Artists and songs
3) Artists and genre tags

Figure 1. Measures of Popularity

	Artist	Artist_Pop_Rating	Listeners_lfm	Scrobbles_lfm	Spot_200_19	Spot_Streams_19	Spot_stream_17	YouTube_views
0	Nickelback	78.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
1	Usher	82.0	2650020.0	39319140.0	-1.0	-1.0	-1.0	-1.0
2	Mary J. Blige	72.0	1309603.0	15708429.0	-1.0	-1.0	-1.0	-1.0
3	P!nk	83.0	2495211.0	68613195.0	-1.0	-1.0	30836574.0	-1.0
4	Ja Rule Featuring Ashanti	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0

Load

We chose to use NoSQL utilizing MongoDB to load our tables into a database. We create 3 tables as listed above.

Different table view samples: