Spotify Song Feature Clustering using K-means

Nicole Aguilera
DSC 680-T302 – Summer 2021
https://github.com/ntiana55/Portfolio_NicoleAguilera/tree/main/AppliedDS/Project3

Data Domain

The domain of this data is Spotify. Spotify is a user-friendly platform that streams millions of songs to both subscribers and non-subscribers. They are the world's largest music streaming service. Spotify houses their data in an API that will give us access to the songs that are available on the platform. The reports below that reference Spotify will only be used for guidance and will not be copied, as models can vary with each construction. Below is the annotated bibliography for this project.

- Spotify Web API. Spotify for Developers. https://developer.spotify.com/documentation/web-api/
  This is the access point for the API that will be used for this project. The API houses features of the songs that the platform streams for its users.
- Spotipy. GitHub. https://github.com/plamere/spotipy
  This site on GitHub provides documention for a Python library that can be used with the API for Spotify. The library may or may not be used in this project.
- Data Science. Spotify. https://engineering.atspotify.com/category/data-science/
  This site by Spotify provides blog posts on how the company's data science department analyzes and uses data to benefit the company. The blogs range from new experiments with their data to studying audio itself.
- Spotiscience: A Tool for Data Scientists and Music Lovers. Towards Data Science. https://towardsdatascience.com/spotiscience-a-tool-for-data-scientists-and-music-lovers-a3e32bd82ed1
  This blog post provides a program created to help analyze both the Spotify API and Genius API. It has a predictor section that is able to predict the song mood to other types of songs like it.
- My Search for a Better Spotify Playlist to Use for Workouts. Towards Data Science. https://towardsdatascience.com/my-search-for-a-better-spotify-playlist-to-use-for-workouts-ed8d4a191074
  Another blog post that describes the creation of a model to help an individual find songs by searching for specific audio features.
- What Song did I miss in the 2010s? Towards Data Science. https://towardsdatascience.com/my-search-for-a-better-spotify-playlist-to-use-for-workouts-ed8d4a191074
  This blogger describes how they were able to utilize k-means clustering, like in this project, to identify songs in the Spotify API. These songs were produced in the 2010s.
- Predicting Popularity on Spotify. Towards Data Science. https://towardsdatascience.com/predicting-popularity-on-spotify-when-data-needs-culture-more-than-culture-needs-data-2ed3661f75f1
  This project constructs a predictive model to identify how popular a song on the streaming platform is. The project includes exploratory data analysis, data models, and results.
- Spotify Data Science Interview Questions. Glassdoor. https://www.glassdoor.com/Interview/Spotify-Data-Scientist-Interview-Questions-EI_IE408251.0,7_KO8,22.htm

This site is comprised of different data science interviews with Spotify and people's experiences. It is interesting to read through them because it gives a sense of what the company is looking for in their data team.

- Spotify Data. Twitter. https://twitter.com/spotify_data
  This Twitter page posts the daily statistics of the songs on Spotify. They range from top artists of all time and top songs for the day on the platform.
- Stats for Spotify. https://www.statsforspotify.com/
  This site allows you to view your personal data on the streaming platform. All that is needed is the site login information.

Data

The dataset that is being examined is found on the Spotify API. The API can be found at the link: https://developer.spotify.com/documentation/web-api/

Research Questions and Benefits

Some research questions for this clustering algorithm are, but not limited to, the following:

- Is there any correlation between the metrics of a song?
- Can a clustering algorithm break down a playlist into smaller playlists based on song metrics?

The benefits to running this clustering algorithm are to improve suggested songs for the user. If the model is able to identify clusters of metrics from a playlist, it can potentially suggest songs for the user, that they would realistically listen to, with those same metrics.

Methods

The main method being used is clustering. The clustering being used in this project will be k-means clustering.

Potential Issues

The obvious potential issue is that the model will not work because there is not enough variation in the playlist or data set that is being pulled from the API. This will not allow the model to identify successful clusters.

The opposite end of the spectrum may also cause an issue. This is that all the songs of a playlist may all be different, so it is difficult for the model to identify a cluster or identify one incorrectly.

Concluding Remarks

The goal of this model is to separate out a Spotify playlist into clusters based on metrics used by the streaming platform. The model being used will be k-means clustering. The API can be found at this link: https://developer.spotify.com/documentation/web-api/