Semester Project Proposal: Spotify Popularity and Audio Feature Analysis

Isaac Fry

October 10, 2022

1 Introduction

As one of the world's top music providers, Spotify's data infrastructure is immense and beautifully engineered while also maintaining widespread access to their backend through API calls [6]. Spotify's APIs for individual songs (i.e. tracks) contain a wealth of information, including `key`, `tempo`, `acousticness`, `speechiness`, `loudness`, `time signature`, `danceability`, and `popularity` metrics [5, 4]. Without a strong data backend, Spotify would lose some of its most prominent features that depend on critical data insights: song discoverability, artist suggestion, playlist generation, radio, marketing, and more. Thus, it seems in Spotify's best interest to continue leveraging this data and allowing developers to access it.

However, Spotify's API calls do not perform well for dataset generation (seeing as its API calls are designed for data retrieval). Any song must be called by a specific `Spotify ID`, and Spotify does not publicly list a dataset of songs that are each correlated with a `Spotify ID`. Thus, finding a dataset of songs with sufficiently interesting features is difficult.

2 Specific Dataset and Motivation

An open source developer created a large dataset (330,000 entries) of each *BillBoard Hot 100* chart from 1958-2020 and a corresponding dataset with information from a Spotify API call for each song [1] (see repository in references). This is one of the only large, publicly available datasets that is modern, includes a large number of songs, and links to a descriptive set of audio feature analysis.

Some of the features included in the base dataset are: `genre`, `Spotify ID`, `duration (ms)`, `explicitness`, `album`, `danceability`, `energy`, `key`, `loudness`, `mode`, `speechiness`, `acousticness`, `instrumentalness`, `liveness`, `valence`, `tempo`, `time signature`, and `popularity`. All of these were previously called from one of Spotify's APIs [5, 4]. However, there are other important features, such as `timbre`, `beats`, `bars`, and confidence levels that are excluded. These features could be retrieved with a call to the API and added to the dataset if needed. This ability to call APIs also allows the open source dataset's accuracy to be confirmed. (I also have a personal motivation to learn how to use Spotify's APIs).

Other datasets exist besides the one cited above, and they may have a higher quality of information. In the audio data analysis community, the Million Song Dataset (MSD) has been used regularly and remains a staple of ongoing open source research [3, 7, 2]. The proprietor of the dataset, EchoNest, published MSD with data generated from Spotify between 2010-2017. At the time, another project, known as Rosetta Stone, allowed a researcher to translate an `EchoNest ID` into a `Spotify ID`, but after Spotify acquired EchoNest in 2016, Spotify absorbed many of the translation features into its private backend [8]. After that point, MSD's insight was reduced, especially as Spotify continued to advance the data available through its API.

Thus, my motivation for using the *Billboard Hot 100* stems from a greater availability of data. With a `Spotify ID` listed for each song, the more advanced audio analysis features can be accessed through an API call. Furthermore, since the *Billboard Hot 100* is a time series data set (with a new chart each week), an extra temporal component can be added to the analysis. Finally, using songs only from the

Billboard Hot 100 gives a highly representative sample of songs. While not every chart-topping song provides accurate insight into the musical feelings of the public, the collective sum of the top songs provides a generally healthy sample of the cultural trends of the time.

3 Research Questions

With as robust of a dataset as the one listed above, multiple interesting questions can be answered. For the sake of brevity and clarity, some potential questions are listed below in order of potential:

- Can a song's analysis description (i.e. its tempo, genre, loudness, average timbre, duration, etc.) be used to predict its popularity?
- Can a song's analysis description be used to predict its period (or year) of greatest popularity?
- Can a song's valence (how emotionally positive a song is) be predicted by its key and tempo?
- Is there a correlation between a song's genre and its audio features?
- Can a song's duration be predicted by its tempo?
- Does an unclassified learning algorithm provide any interesting insights into musical trends?

References

[1] @kcmillersean. *Billboard Hot weekly charts*. URL: https://data.world/kcmillersean/billboard-hot-100-1958-2017. (accessed 10.09.2022).

[2] Thierry Bertin-Mahieux et al. "The Million Song Dataset". In: *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)* (Jan. 2011), pp. 591–596.

[3] EchoNest. *Million Song Dataset*. URL: http://millionsongdataset.com/. (accessed 10.09.2022).

[4] Spotify Engineering. *Get Track Documentation*. URL: https://developer.spotify.com/documentation/web-api/reference/#/operations/get-track. (accessed 10.09.2022).

[5] Spotify Engineering. *Get Track's Audio Features Documentation*. URL: https://developer.spotify.com/documentation/web-api/reference/#/operations/get-audio-features. (accessed 10.09.2022).

[6] Spotify Engineering. *Spotify Web API*. URL: https://developer.spotify.com/documentation/web-api/reference/#/. (accessed 10.09.2022).

[7] James Pham, Edric Kyauk, and Edwin Park. "Predicting Song Popularity". In: *Stanford Project Repository* (2015). DOI: http://cs229.stanford.edu/proj2015/140_report.pdf.

[8] Peter Skiden. *API Improvements and U*. URL: https://developer.spotify.com/community/news/2016/03/29/api-improvements-update/. (accessed 10.09.2022).