Spotify Playlist Analysis

Problem statement

Spotify is a music, podcast, and video streaming service. It provides digital rights management-protected content from record labels and media companies. Spotify is a freemium service, meaning that basic services are free with advertisements, with additional features, such as improved streaming quality, offered via paid subscriptions.

One of Spotify's primary products is Playlists, collections of tracks that individual users (or Spotify) can build for every mood or event. Spotify users can make or follow as many playlists as they like. With over 40 million songs available, the company attempts to direct the most relevant songs to users based on their preferences, and Playlists often comprise the most convenient and effective way to convey these recommended songs to a user.

Spotify participates in the creation and curation of Playlists that are *followed*, or listened to, by millions of Spotify users. These Playlists are compiled in a complex manner, involving both human-led and computer-led processes. What stands is that algorithmically-curated discovery playlists, and their effectiveness, remain an important business interest for the company. The goal is to better understand how these algorithms can be evaluated and improved with machine learning techniques learned in the class.

Project goal: Generate a machine learning model that can predict the success of a Playlist (the number of end followers), using data corresponding to the Playlist's length, genre, and component tracks. You will attempt to do this with Spotify API data. Your goal is to use this model to generate a *successful* Playlist according to a user-specified genre like Indie, Rock, Classical, and so on.

Data resources

The data will come from the Million Song Dataset and Spotify API.

1. **Million Song Dataset**
   This is a freely-available collection of audio features and metadata for a million contemporary popular music tracks. Some information included per track is below.
   - Artist information
   - Audio-extracted features
   - Duration and timing information
2. **Spotify API**
   - The component songs of a Playlist
   - The number of followers of a Playlist

- Spotify-derived audio features for each track
- An ISRC number for each track, potentially linking this API to other relevant datasets

High-level project goals

1. Derive a model that can predict Playlist success using **only** Spotify-provided predictors.
2. Combine features from other datasets such as Wikipedia, Million Song Dataset, and other external datasets, in hopes of reaching an improvement in predictive quality.
3. Use this improved model to generate new Playlists according to a user-specified genre or other search filters, and present these as potential ideas for implementation at Spotify.

References

1. Berenzweig, Adam, Beth Logan, Daniel P.W. Ellis and Brian Whitman. *A Large-Scale Evaluation of Acoustic and Subjective Music Similarity Measures.* Proceedings of the ISMIR International Conference on Music Information Retrieval (Baltimore, MD), 2003, pp. 99?105.
2. Logan, B., *A Content-Based Music Similarity Function*, (Report CRL 2001/02) Compaq Computer Corporation Cambridge Research Laboratory, Technical Report Series (Jun. 2001).