Automatic Playlist Recommender

Problem statement

Music recommender systems (MRS) have recently exploded in popularity thanks to music streaming services like Spotify, Pandora and Apple Music. By some accounts, almost half of all current music consumption is by the way of these services. While recommender systems have been around for quite some time and are very well researched, music recommender systems differ from their more common siblings in some characteristically important ways: the duration of the items is less (3-5 min for a song vs 90 minutes for a movie or months/years for a book or shopping item), the size of the catalog of items is larger (10s of millions of songs), the items are consumed in sequence with multiple items consumed in a session, repeated recommendations have a different significance since listening to the same song as part of different playlists may be ok, and consumption occurs passively i.e. in the background. Music Recommender Systems then require different approaches from traditional recommender systems.

One of the major problems in Music Recommender Systems is the station/ playlist generation problem. At its heart, the playlist generation is about finding the set of songs to recommend to best extend the experience of a listener in the midst of a playlist. By suggesting appropriate songs to add to a playlist, a Recommender System can increase user engagement by making playlist creation easier, as well as extending listening beyond the end of existing playlists.

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

- Automatic Playlist Generation — create a model for song discovery on the basis of the base playlist and user/context information that might be important to quality of a playlist. Some of user/context information might include intent, emotion, location, playlist "purpose" (driving/road trip, studying, etc). Use the developed model(s) for automatic playlist generation.
- Cold Start Problem — A variant of the problem of automatic playlist generation, the cold start problem involves creating a model to find good choices of songs for new playlists with relatively few prior playlist entries.

Data resources

Students will focus on the million playlist dataset, but in their exploration might use other publicly available playlist data, individual song data and/or generate data for contextual information. Spotify API is also available.

1. **Million Playlist Dataset**

    - http://recsys-challenge.spotify.com/
    - Large (5.4 GB) Playlist dataset from Spotify
    - Created in 2018

2. **Million Song Dataset**

- https://labrosa.ee.columbia.edu/millionsong/lastfm
- Generated in 2010/2011 (but you can use last.fm to regenerate)
- Freely-available collection of audio features and metadata for a million contemporary popular music tracks. Some information included per track includes:
  - Artist information
  - Audio-extracted features
  - Duration and timing information

3. **Lyrics Wiki**

- http://lyrics.wikia.com/wiki/LyricWiki
- Database of Lyrics
- May come in handy for NLP/Textual based metadata

4. **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

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).

3. Shedl, M. et al, *Current Challenges and Visions in Music Recommender Systems Research*, https://arxiv.org/pdf/1710.03208.pdf.

4. Shedl, M., Peter Knees, and Fabien Gouyon, *New Paths in Music Recommender Systems*, RecSys'17 tutorial, http://www.cp.jku.at/tutorials/mrs_recsys_2017/slides.pdf.