Spotify Playlist Challenge

Brendan Ryan
Christopher Lescinskas
Matthew Martin

Project Description

For this project we perform data mining on the dataset provided through the *Million Playlist Dataset*[1] (MPD) to provide suggested extensions to a playlist prompt. We will also seek to understand if certain types of songs (or songs) have tendencies to reside at particular points in playlists e.g. does a song tend to start off a playlist or end a playlist.

[1] AIcrowd. Spotify Million Playlist Dataset Challenge. Retrieved from
https://www.aicrowd.com/challenges/spotify-million-playlist-dataset-challenge

Prior Work

- RecSys Challenge 2018
    - Put on by Spotify, The University of Massachusetts, Amherst, and Johannes Kepler University, Linz
- Chen, Shuo & Moore, Josh & Turnbull, Douglas & Joachims, Thorsten. (2012). *Playlist prediction via metric embedding*. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 10.1145/2339530.2339643
    - Uses market basket style analysis and machine learning algorithm to generate playlists
    - Playlists represented as ‘Markov Chains in a latent space’ and songs as a point(s) in that space

Datasets

Dataset: Spotify Million Playlist Dataset

URL where found:
https://www.aicrowd.com/challenges/spotify-million-playlist-dataset-challenge/dataset_files

We have the dataset downloaded on Chris Lescinskas's machine

Dataset summary (From stats.py function, see List of Tools):

number of playlists 1,000,000

number of tracks 66,346,428

number of unique tracks 2,262,292

number of unique albums 734,684

number of unique artists 295,860

Proposed Work

- As the dataset was released for a competition, data quality is high and cleaning does not appear necessary.
- Focus will be on developing an optimal algorithm/program for generating additional recommended songs given a seed song or playlist
- Initial work will focus on developing an efficient *a priori*-like algorithm to mine frequent patterns given seed songs.
- Later work may include additional analyses and algorithms as we learn about them in class

List of Tools

1. Github -> storage and version control of documents and code
2. Jupyterhub/Python 3 -> coding environment
3. Tableau -> Data visualization
4. Tools provided by MPD
    a. Check.py → Confirm that dataset is correct/uncorrupted
    b. Print.py and show.py→ Print and show a subset of the dataset respectively
    c. Stats.py and Deeper_stats.py → Iterates through MPD and shows summary information

Evaluation

1. We will evaluate the accuracy of our playlist suggestions by testing them against existing playlists.
2. We will test the accuracy of our ability to predict a songs position in a playlist by testing against existing playlists in the dataset

Note - we plan to use ⅔ of the data in our dataset for training and ⅓ for testing