Automatic Playlist Generation

Gökçe Uludoğan , Beyza Gül Gürbüz , Hilal Dönmez

Boğaziçi University , Computer Engineering

Introduction

Musical playlists have become very popular means of listening to music. User can create their own musical lists according to the preference at some applications such as Spotify, Pandora, Youtube. Such applications recommend songs to the user with regard to the user preference. Since there are more songs than ever at music listeners' fingertips, generating or recommending musical playlists has been more challenging. For this reason, our project explores generating the musical playlist based on seed songs and their features extracted from various methods.

Dataset

Our data comes from the following sources:

- Million Song Genre dataset (MSD)[1]
- Musicxmatch dataset (mxm)[2]
- Spotify[3]

The final data is created by combining these data sources.

Genre	Number of Tracks
classic pop and rock	5772
folk	2370
soul and reggae	796
metal	780
punk	660
pop	523
dance and electronica	417
jazz and blues	258
hip-hop	64
classical	27
total	11667

Table: Distribution of tracks in dataset over genres

Features

Features gathered from MSD and Spotify are taken as raw. On the other hand, using the lyrics from mxm dataset, lyrics categories are generated with Latent Dirichlet Allocation and ten features regarding these categories are derived.

- **MSD features:**
  - Tempo, Timbre, Genre, Danceability, Energy
- **Spotify features:**
  - Tempo, Danceability, Energy, Loudness, Speechiness, Acousticness, Instrumentalness, Liveness, Valence
- **mxm features:**
  - Lyrics Categories

Methods

For playlist generation, both supervised and unsupervised methodologies are investigated. We have used a few classifiers on both binary and multiclass genre classification problem. For graph models, two clustering models and one probabilistic model are generated for predicting playlists based on similarity with a seed song or weighted feature.

1. **Classification**
   - Logistic Regression
   - Support Vector Machine
   - Random Forest Classifier
2. **Graph Models**
   - KMeans Clustering
   - Affinity Propagation
   - Probabilistic Song Generation with distance

Results

Results for multiclass classification is below:

Classifier	Training Acc	CV Acc	Test Acc
Random Forest	0.99	0.61	0.64
Logistic Regression	0.66	0.64	0.67
SVM	1.0	0.49	0.50

Table: Multiclass Classification

For each genre, binary classification is applied and average accuracies for classifiers can be seen below:

Classifier	Training Acc	CV Acc	Test Acc
Random Forest	0.99	0.91	0.93
Logistic Regression	0.93	0.92	0.93
SVM	1.0	0.9	0.9

Table: Binary Classification Overall

For each model, ten playlists are generated with choosing random seed songs and several metrics are calculated. For each playlist, distance of seed song to playlist songs and distance within playlist songs are computed. These distances are normalized by dividing maximum distance in dataset. The average of these metrics over playlists are below:

Model	AvgDisToSeed	AvgDistBtwPL
KMeans	0.128	0.121
Affinity Propagation	0.148	0.147
Probabilistic Model	0.169	0.210

Table: Graph Models Distances

For comparison, average distance in dataset is 0.164.

Conclusion

From the results we obtained, the best performing classifier in the dataset is Logistic Regression in both binary and multiclass classification. From graph models' results, we see that clustering models generate playlists whose intraclass similarities is high whereas probabilistic model creates more diversified playlists by nature. This model is suitable for generating discovery playlists.

References

[1] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere.
The million song dataset.
In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

[2] `http://labrosa.ee.columbia.edu/millionsong/musixmatch`.

[3] `https://www.spotify.com/`.

[4] Kade Keith and Demetrios Fassois.
Automated playlist generation.

[5] `https://github.com/plamere/spotipy`.

[6] `http://scikit-learn.org/stable/index.html`.

Contact Information

- Gökçe Uludoğan: gokceuludogan@gmail.com
- Beyza Gül Gürbüz: beygrbz@gmail.com
- Hilal Dönmez: hilaldonmez14@gmail.com