Automated Music Playlist Generator

Arka Sarkar
IIITD
arka18222@iiitd.ac.in

Pankil Kalra
IIITD
pankil18061@iiitd.ac.in

Daksh Thapar
IIITD
daksh18137@iiitd.ac.in

Abstract

With the growing popularity of music streaming services like Spotify, Apple Music and Wynk, the number of songs have skyrocketed globally. Creating personalized playlists for users has become tedious and challenging as it involves individual listening to various songs and categorizing them based on their audio features. The objective is to sort songs with similar musical characteristics into playlists automatically. Modern machine learning techniques and visualization tools should help us find accurate models that categorize millions of songs into user playlists based on song choices. Related works on this problem did not consider Lyrical Analysis while making playlists. We are using Topic Modelling techniques on lyrics, and will be using the extracted topics as a feature for generating playlists.

Source code : [Github]

1. Introduction

With the ever growing popularity of music streaming services and the number of songs skyrocketing , it has become increasingly challenging for a individual to construct and segregate playlists. We propose Binary Classification models that we can use for every playlist that can classify whether a new song belongs to that playlist or not. Also, we have implemented a single Multi Class Classification model which can classify a song into one of many different playlists. This classification process has diverse practical applications as it automates the assignment of playlists to songs of similar characteristics, improving the user experience and easing the process of music playback. Graph-Based approaches will be used to generate playlists containing songs with similar features from a collection. In retrospect, our work can find key applications among music enthusiasts and in mobile applications to make song collections for not only individuals, but for age groups and people with similar interests for song genres and artists.

2. Literature Review

Playlist generation is a broad problem which has been approached in many different ways.

1. Automated Playlist generation from Personal Music Libraries [1] by Diana Lin and Sampath Jayarathna provides a method to use K-means clustering, Affinity Propagation and DBSCAN on various audio features (such as ”danceability”, ”energy”, ”acousticness”) to generate an automated playlist.

2. A Comparison of Playlist Generation Strategies for Music Recommendation and a New Baseline Scheme [2] by Geoffray Bonnin and Dietmar Jannach uses multiple methods such as KNN, Popularity-based, Same artists - greatest hit, Collocated artists - greatest hit, Content-based approaches to generate playlists.

3. Music Playlist Generation based on Community Detection and Personalized PageRank [3] by Bangzheng He, Yandi Li and Bobby Nguy uses song similarity graph approaches and apply community detection and Personalized PageRank to generate playlists.

3. Dataset Features

We have picked the Million Song Dataset (MSD lyrical) which consists of 2,33,662 songs with lyrics. We collected the Top Search Results of playlists using phrases like ’Summer’, ’Love,Party’, ’Breakup’, ’Rock’, ’Country’, ’Romance’, ’Rap’, ’Metal’, ’Hip-Hop’, ’Latin’, ’Blues’, ’Soul’, ’Classic’, ’Pop’, ’Jazz’, ’Folk’, ’RB’ from Spotify API which have a significant song overlap with the MSD dataset. In total, we obtained 169 playlists amounting to 11,159 unique songs.

We extracted 14 audio features for each song using Spotify API as shown below.
We performed Topic Modelling and extracted additional 20 features using Latent Dirichlet Allocation (LDA) on the lyrics which will be described in the subsection below.

Feature	Datatype
danceability	float
energy	float
Key	int
loudness	float
mode	int
speechiness	float
instrumentalness	float
liveness	float
valence	float
duration_ms	int
release_date	int
popularity	int

Table 1. Raw Features

Example song :

{ ’song_id’: 73etijhz7pV4Wx7GTANLpq, ’danceability’: 0.332, ’energy’: 0.371, ’key’: 4, ’loudness’: -9.884, ’mode’: 1, ’speechiness’: 0.0396, ’acousticness’: 0.729, ’instrumentalness’: 0.00021, ’liveness’: 0.176, ’valence’: 0.375, ’tempo’: 173.273, ’duration_ms’ : 249093 , ’release_date’: 2002, ’popularity’ : 51 }

The class distribution of playlists is shown in figure 1 :

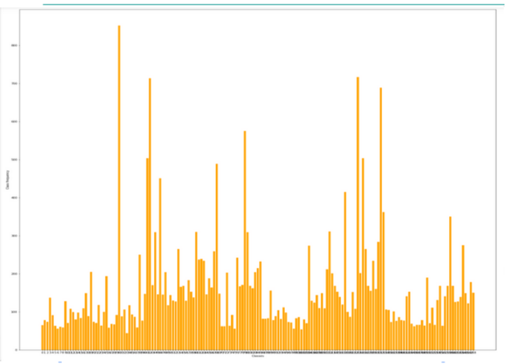

Figure 1. Class Distribution of Playlists.

3.1. Preprocessing, Feature Extraction

The lyrics obtained from Million Song Dataset were already preprocessed to some extent. The words had been stemmed and lyrics were converted to Bag of Words format. We further removed Stopwords and words with less than 3 characters for better topic creation.

3.1.1 Topic Modelling using Latent Dirichet Allocation (LDA)

We preprocessed our lyrics, applied topic modelling using Latent Dirichlet Allocation (LDA) and extracted the probabilities of each topic in every song. Latent Dirichlet allocation is a generative statistical model which provides us with a fixed number of unobserved topics which would help in analysing some similarity between playlists.

Latent Dirichlet Allocation (LDA) was applied on the lyrical corpus, we tried different values for the number of topics ranging from 3 to 30. The topics obtained were evaluated using Covariance ‘c_v’ score. *“20 number of topics”* gave the best c_v score of 0.58 and we extracted the probabilities of each topic in every song for the same. 20 new columns were added to our dataset containing the topic wise probabilities for each song.

Sample topic

Topic 18: 0.035*”dead” + 0.032*”kill” + 0.024*”head” + 0.022*”black” + 0.022*”blood” + 0.016*”fight” + 0.014*”hate” + 0.013*”lyric” + 0.013*”readi” + 0.013*”death”

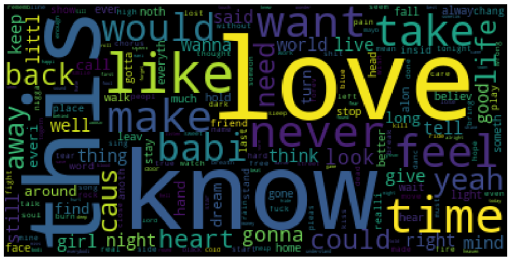

Figure 2. Word Cloud on Lyrics.

3.1.2 Dimensionality Reduction using Principal Axis Component (PCA)

Principal Component Analysis is a method of reducing dimensions of a dataset by transforming a large set of variables into a smaller set of variables that still contains most of the information in the larger data set. The various steps in PCA include standardization, computation of covariance matrix and eigenvectors to identify principle components. PCA can be thought of as fitting a “ p-dimensional ellipsoid” to the data; every axis of the ellipsoid means a single principal component. For our project, we are working with 30 principle components. Figure 3 shows the explained variance of the principle components.

3.1.3 Visualizing High Dimensional Data using t-SNE

t-Distributed Stochastic Neighbour Embedding (t-SNE) is an unsupervised, non-linear technique primarily used for data exploration and visualizing high-dimensional data. t-SNE, unlike PCA, is not a linear projection. It uses the local relationships between points to create a low-dimensional

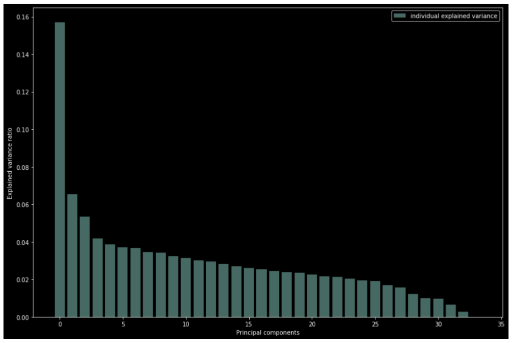

Figure 3. PCA Explained Variance

mapping. This allows it to capture non-linear structure. t-SNE creates a probability distribution using the Gaussian distribution that defines the relationships between the points in high-dimensional space.

3.1.4 Data Standardization

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation. Here's the formula for standardization:

$$z_i = \frac{x_i - \bar{x}}{s} \tag{1}$$

$\bar{x}$ is the mean of the feature values, σ is the standard deviation of the feature values.

4. Methodologies

Our objective is to generate automated song playlists for users using Label Classification and Graph-based techniques. For the Classification problem, two different methodologies were used. We tried different ML based classification algorithms both for a single-class prediction for 169 playlist and multi-class prediction for 13 playlists.
For playlist generation, we used clustering and graph-based approaches to construct playlists based on the feature similarity of songs from certain seed songs.

4.1. Classification

We applied binary classification on a total of 169 playlists. We used a One vs All approach where the target playlist was considered positive and all other playlists were considered negative and randomly undersampled. We performed 60:20:20 train-validation-test split and stratified sampling on the dataframe constructed.
For multi-class classification, we took all the non overlapping songs among various playlists; the total songs reduced from 11159 to 6120 for 169 playlists. We filtered out playlists having more than 100 songs for our model training.

We used the following classification models: Logistic Regression, Decision Trees, Random Forests, Linear SVC, XGBoost Ensemble technique, K-Nearest Neighbours Classifier and Artificial Neural Network. To optimise various parameters in the aforementioned models, we applied GridSearchCV and 10-fold cross validation.

4.2. Playlist Generation

We performed data standardization, dimensionality reduction using Singular Value Decomposition (SVD) and used the following graph- based techniques for Playlist visualization and generation.

4.2.1 K Nearest Neighbours

Each song was represented in an N-dimensional space according to our features. The we selected the next K songs for that playlist based on the Euclidean distances measured from the average(centroid) of all of the seed songs and returned these nearest songs as a playlist collection to the user with song and artist names.

4.2.2 Clustering

We are running unsupervised clustering techniques to segregate songs into different clusters, which will act as playlists.

K-Means Clustering : We implemented K-Means Clustering algorithm, one of the most common unsupervised Machine Learning algorithm. K-Means is a centroid-based algorithm, or a distance-based algorithm, where we calculate the distances to assign a point to a cluster. The algorithm identifies k number of centroids, and then each data point is assigned to its nearest cluster, while keeping the centroids as small as possible.
Agglomerative Clustering: The agglomerative clustering is the most common type of hierarchical clustering used to group objects in clusters based on their similarity. The algorithm starts by treating each object as a singleton cluster. Next, pairs of clusters are successively merged until all clusters have been merged into one big cluster containing all objects. The result is a tree-based representation of the objects, named dendrogram.

5. Results and Analysis

5.1. Latent Dirichet Allocation (LDA)

The visualisation Figure 4 shows the importance of different topics, their separation in space in our lyrics corpus. The size of each bubble represents the importance of that topic for our LDA model. The visualisations shows 20 clearly separated topics.

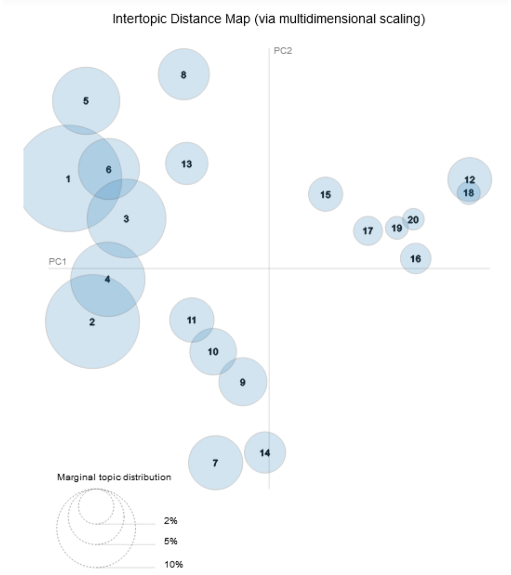

Figure 4. Visualisation of topics extracted from lyrics

5.2. Classification

5.2.1 Binary Classification

For Binary Classification we used all 169 playlists.We used a One vs All approach where the target playlist was considered positive and all other playlists were considered negative and randomly undersampled.

The XG Boost algorithm outperforms all the models with an Average F1 score of 81.0% across all playlists. Linear SVC and Logistic Regression also gave very good results posing F1 (macro) scores 78.6% and 79.2% respectively. Metrics for each model are shown in the table 5.2.1:

Model	Precision	Recall	F1
SVC	0.791	0.787	0.786
LR	0.806	0.788	0.792
DT	0.751	0.750	0.748
RF	0.839	0.758	0.774
XGB	0.834	0.800	0.810
KNN	0.774	0.728	0.726
ANN	0.769	0.781	0.771

Table 2. Binary Classification Scores

The training and Validation learning curve for SVC is shown in Figure 6 :

The ROC curve is also plotted in figure 6 and shows an excellent area under the curve.

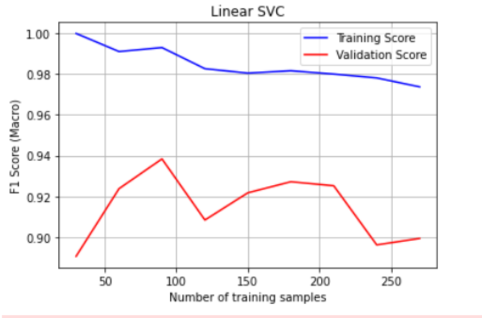

Figure 5. Learning Curve for SVC

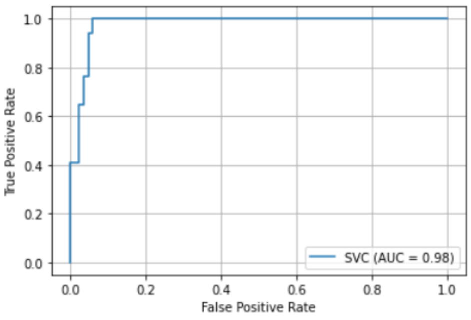

Figure 6. ROC curve for SVC

5.2.2 Multi Class Classification

For multi-class classification 5.2.2 we took 13 non-overlapping playlists each having more than 100 songs. XGB in this case also got the best performance posing a F1 score of 58.7%.

Model	Precision	Recall	F1	Acc
SVC	0.536	0.483	0.496	0.531
LR	0.510	0.505	0.503	0.526
DT	0.423	0.419	0.419	0.440
RF	0.623	0.549	0.565	0.586
XGB	0.602	0.511	0.587	0.602
KNN	0.457	0.438	0.428	0.484
ANN	0.476	0.500	0.401	0.577

Table 3. Multi Classification Scores

XBG performed well in our case as it is an ensemble method and XGBoost improves upon the basic Gradient Boosting Method framework through systems optimization and algorithmic enhancements.

We also computed the feature importance of the features which are shown below :

As we can see from figure 7 the topics obtained from the lyrics data had the highest importance (by a significant margin) for song classification. This implies that lyrics of a song are an important attribute which an user takes into ac-

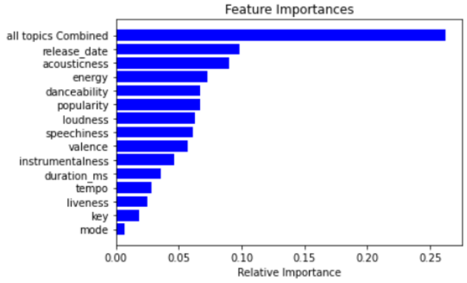

Figure 7. feature Importance for Random Forest multi classification

count while creating a playlist. Also "release_date", "acousticness" and "energy" are important attributes of a song.

5.3. Playlist Generation

5.3.1 Clustering

We took the song corpus consisting of 11159 songs and applied SVD to get 15 component features for each song. We ran K-Means clustering algorithm for k ranging from 2 to 10. We would report the average distances of the songs in each playlist .

As a baseline we would be using the average distance of songs in our training playlist. The results are as follows :

Model	Measure(Mean Distance)
Spotify	3.468
K-Means	3.446
Agglo(complete)	3.656
Agglo(ward)	3.974
Agglo(average)	4.137

Table 4. Mean Distances

We can observe lower mean Euclidean distance between the song nodes of our clustered playlists which implies we have achieved really good quality playlists comparable to actual top playlists on Spotify. The playlist generated from can be visualised using t-SNE as shown in figure 9 below :

To find the optimum number of clusters in both K-means and Agglomeration, we plotted the silhouette scores for clusters ranging from 1 to 30 and found out that $k = 8$ was optimum for both K-means and Agglomeration. Figure 8 shows the plot for K-Means.

The Figure 9 cluster diagram shows us a beautiful distribution of playlist clusters and shows clear distinction between the K playlists. We can clearly distinguish individual playlists with decision boundaries. The above diagram signifies the success of our clustering algorithms and feature extraction processes; we are able to distinguish songs into playlists on the basis of our extracted features from Spotify.

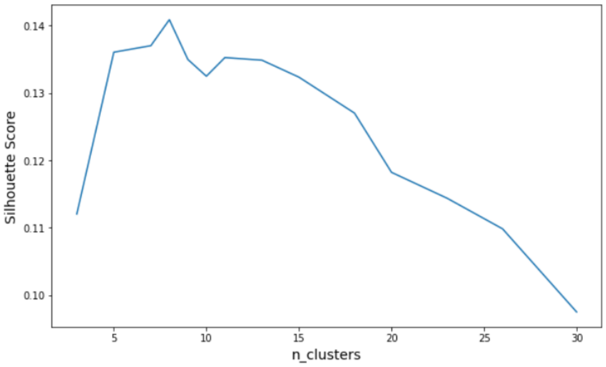

Figure 8. Silhouette Scores for K-Means

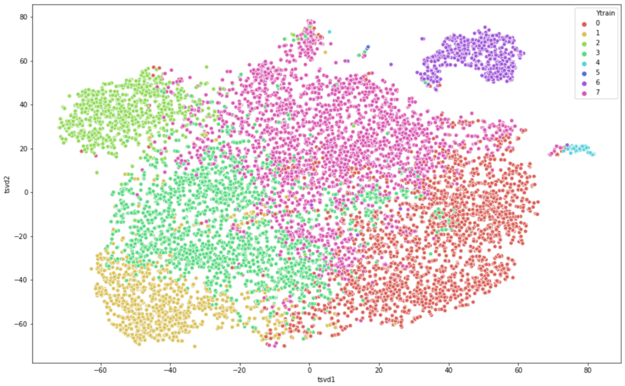

Figure 9. Cluster for k = 8 for K-Means

5.3.2 K Nearest Neighbours

For playlist generation, we sent as an input the information of a country genre song: 'On the Road Again', Bob Dylan For this corresponding input, the automatic playlist generator based on K Nearest Neighbours, where K=50 returns the following generated playlist as output (10 songs here):

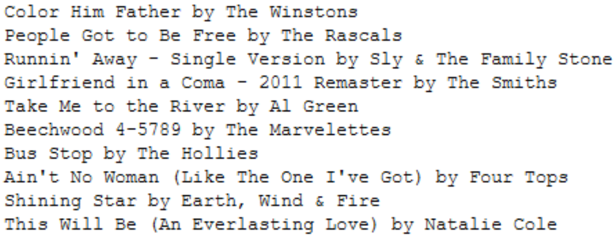
```
Color Him Father by The Winstons
People Got to Be Free by The Rascals
Runnin' Away - Single Version by Sly & The Family Stone
Girlfriend in a Coma - 2011 Remaster by The Smiths
Take Me to the River by Al Green
Beechwood 4-5789 by The Marvelettes
Bus Stop by The Hollies
Ain't No Woman (Like The One I've Got) by Four Tops
Shining Star by Earth, Wind & Fire
This Will Be (An Everlasting Love) by Natalie Cole
```

Figure 10. Songs Generated

We can clearly observe that our output generated playlist contains songs that are very similar to the input songs. On closer manual observation for this example, we see that the output playlist's songs are majorly of the 'country' genre, have slower rhythmic melodies, similar classic instrumentals and having similar lyrics and wordings.

6. Conclusion

6.1. Outcomes

We are pleased to inform that our initial hypothesis that lyrics play an significant role in curating playlists was in fact true. This can be clearly observed by looking at the feature importance graphic shown in the results section.
The metric scores obtained on our binary and multi classification were very promising and posed a higher or similar score with respect to the previous work done on this topic.
One of the limitations of the MSD dataset is that these contain information including audio features and song lyrics before 2011, and it does not contain all the trending songs from this era. These datasets mostly contains songs of the genres 'Rock' and 'Metal'.
The scores obtained by us shows the importance of audio features and lyrics in classifying playlists. Future improvements on the approach can be made by gathering more features and even creating derived features to improve the performance.

6.2. Future Work

Future general improvements could focused on extracting more song characteristics along with deriving more generated features from them to improve upon the performance of the current models. This process could be achieved by gathering more data from various providers especially some of the genres that are under represented in the MSD dataset. This entire idea could be extended to be implemented in current streaming softwares with interactive UI for musical professionals and enthusiasts. The idea of topic modelling from lyrical corpus can be extended to various text based services such as twitter, reddit and facebook data to extract topics from them and provide a similar implementation for such systems as well.

6.3. Member Contribution

1. Arka Sarkar: Dataset Preprocessing- Data Extraction from Spotify API Topic Modelling- LDA Dimensionality Reduction- PCA EDA- Word Cloud Literature Review, GridSearchCV, Logistic Regression and Analysis, Support Vector Machine and Analysis, K-Means Clustering, Metrics comparison for Graph-based Techniques

2. Daksh Thapar: Dataset Preprocessing- Feature Extraction, removing stop words Topic Modelling-LDA Data Visualization for Playlists- t-SNE EDA-Data Correlation Heatmap Literature Review, GridSearchCV, Decision Tree and Analysis, XGB and Analysis, K- Nearest Neighbours, Metrics comparison for ML classification models

3. Pankil Kalra: Dataset Preprocessing- Extracting song lyrics from Million Song Dataset (MSD) Topic Modelling- LDA Dimensionality Reduction- PCA EDA- Feature Histogram Literature Review, GridSearchCV, KNN and Analysis, ANN and Analysis, K-Nearest Neighbours, ROC Plots

References

[1] Lin, D. and Jayarathna, S. (2018). Conferences 2018 IEEE International Confe... Automated Playlist Generation from Personal Music Libraries. IEEE International Conference on Information Reuse and Integration (IRI). [online] Available at: https://ieeexplore.ieee.org/abstract/document/8424710.

[2] Bonnin, G. and Jannach, D. (2013). AAAI 2013 Workshop. [online] Available at:https://www.aaai.org/ocs/index.php/WS/AAAIW13

[3] Music Playlist Generation based on Community Detection and Personalized PageRank. Stanford University Social and Information Network Analysis Autumn 2015. [online] Available at: http://snap.stanford.edu/class/cs224w-2015/

[4] Pichl, M., Zangerle, E. and Specht, G. (2017). Understanding Playlist Creation on Music Streaming Platforms. IEEE International Symposium on Multimedia (ISM). [online] Available at: https://ieeexplore.ieee.org/document/7823674.

[5] Pampalk, E. and Gasser, M. (2006). An Implementation of a Simple Playlist Generator Based on Audio Similarity Measures and User Feedback. ISMIR 2006, 7th International Conference on Music Information Retrieval. [online] Available at: https://www.researchgate.net/publication/