|  |
| --- |
| AMPBA-Class of 2020 Winter  Term IV |
|  |
| ***Team: Group-1***  Unsupervised Machine Learning-II  Project Progress Report  ***Karthik Shankar S (PGID: 11920008)***  ***Pundareek Chandrashekhar (PGID: 11920049)***  ***Rajesh Rangaswamy (PGID: 11920072)***  ***Praveenkumar Agoorkaisetty (PGID: 11920096)*** |

**

***Project Goals and Objectives***

*Listening to music is obviously one’s personal choice. But with the arrival of smart gadgets and availability of popular online entertainment channels, listening to music no longer is completely one’s personal choice rather it can be rephrased like, a listener can now be coerced or recommended to listen a song based on their listening history in an entertainment ecosystem.*

*By listening history means, a song listened by an user can be tagged to ‘n’ number of attributes for suggesting any other user with same or similar taste profile.*

*The Project titled* ***“Song Recommendation System”*** *is an attempt to our Assignment Fulfilment for Unsupervised Learning-II as part of the Academic Curriculum.*

***Goal of the Project***

*To devise an Unsupervised Machine Learning model which is expected to recommend a song to an user, based on collaborative filtering techniques.*

***Objective of the Project***

*Objective of the attempt is to develop a Proof of Concept (PoC) for the Song Recommendation System by training and testing the model with the subset data first hand.*

*We aim to use machine learning to suggest or recommend a song based on both user and song characteristics with the metadata about a song (such as user listened and # of times listened) From the models, we hope to also get a deeper insight into the features that are most suggestive of a song popularity, and understand what makes certain songs more popular than others.*

*Our problem statement is thus*

1. *How to suggest or recommend a song based on user listening history?*
2. *What would be the song recommendation for a new user based on the model being developed?*
3. *What type of modelling technique would yield a better song recommendation or suggestion?*

*We propose to define success of this project by analysing the metric which would be obtained from the ROC Curves later.*

***Data Collection Steps***

*The main source of the dataset for the project in focus is referred from* [*https://www.kaggle.com/c/msdchallenge*](https://www.kaggle.com/c/msdchallenge) *and* [*www.millionsongdataset.com*](http://www.millionsongdataset.com)*.*

*The Million Song Data Set (MSDS) is a freely available collection of audio feature and one of the extensive database of popular contemporary tracks spanning decades of western music.*

*The core of the dataset is the feature analysis and metadata for one million songs, provided by* [*The Echo Nest*](http://the.echonest.com/)*. The dataset does not include any audio but only the derived features. Both the subset and the entire dataset are fairly large. The subset is 2.5GB while the full Million Songs dataset is 270GB. We initially tried building our models directly using the original dataset. However, that made it difficult to do data cleaning, imputation and statistical analysis. This has imposed the limitation on us to use the subset of the original dataset. Hardware requirements would be highly demanding for handling the bulk of dataset with even 3 to 4 key attributes for receiving a recommendation from the model. This actually would become a big data problem involving hadoop and map reduce. We have kept this issue to handle later and have decided not to make use of the entire dataset while we pursue on PoC objective at this point of time.*

*Hence, we switched to a different method based on the observation that a lot of the data in the dataset was not relevant to us at this instant (such as the actual audio tracks) so we filtered the data from original file and converted it into a much more compact CSV representation. The basic dataset with associated attributes are actually available in different files. Out of the 55 attributes associated to a particular song, only the required features were identified, grouped, extracted and then merged to get the workable subset over which the model will be run. We filtered the files by only extracting features relevant to our problem statement. For the 30,000 song subset, these included the following fields: song\_id, user id and # of times played. This filtering enabled us to convert the 270GB dataset into a significantly more manageable 370 MB CSV file.*

***Why this data suits to our Project Goals?***

*While a lot of work has been done on applying machine learning to different facets of the million song dataset, there has been little to no focus on song popularity. Research has focused on predicting the year, artist of a song, recommending songs based on genre etc. There is also a lot of existing literature on attempting to determine the characteristics of a popular song, but they have tended to focus more on evaluative or statistical approaches than a big data or machine learning approach. We present a system that ties together the two aspects we build machine learning models that are able to suggest a song based on listening history accurately and then use these models to get more insight into the features that are the strongest signals of song popularity.*

*Objective is Song Recommendation.*

*Goals:*

*SMART*

*Basic Analysis of Data (Descriptive Statistics, Visualization)*

*Preliminary Analysis*

*Description of the Way Forward.*