INTRODUCTION TO MACHINE LEARNING
ECE GY-6123

Music Sentiment Analysis using TF/IDF

Submitted To: Prof. Sundeep Rangan
Submitted By: Mandar Mhaske, Kunal Ninawe, Rakshita
Student ID: mm10435, kvn238, rr3653
Date of Submission: Dec 11, 2020

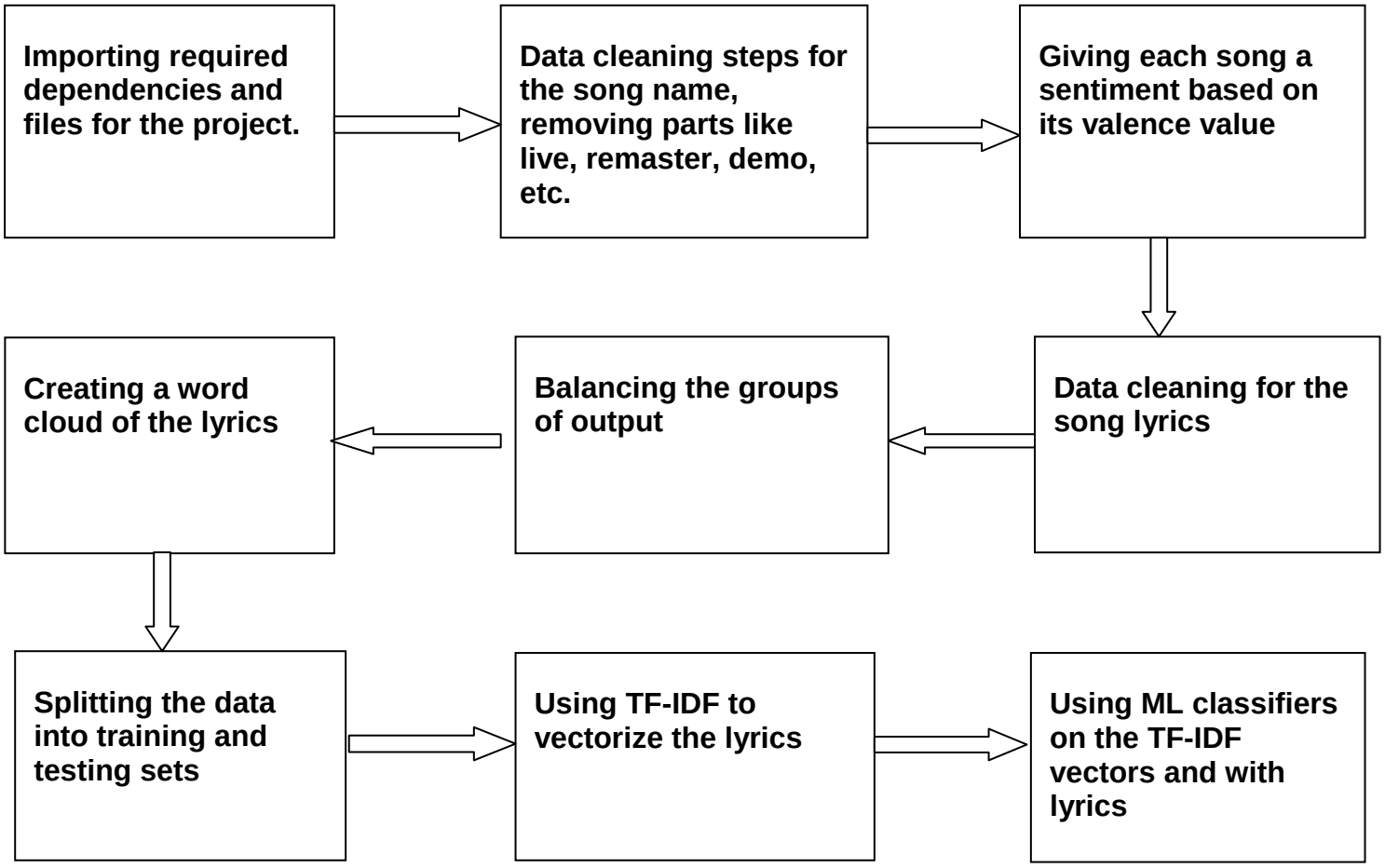
Introduction

The aim of the project is to determine the mood of the song based on the lyrics. Various components of a song contribute to its sentiment/mood in general. Some of these components include lyrics, tempo, background music etc. The approach used in this project is to simply understand the pattern of words in the song lyrics for different categories for example, Happy Songs and Sad Songs.

Background

Lots of machine learning tools are available today to perform Natural Language Processing tasks. Task like predicting the sentiment of a text are quite common in Machine Learning domain. The challenge is to perform these tasks at scale

Architecture



Machine learning models can't work with textual features and require some sort of numerical features. A common practice in training text models is by taking into account the frequency of the words in the document or corpus.
There are 2 techniques to achieve this: Count Vectorization and TFIDF Vectorization

TFIDF

TFIDF is an alternative to calculate word frequency. It assigns a word frequency score which tells how important the word is with respect to the document and the entire corpus. For example it assigns a higher value to a word that is frequent in a few documents but not across all the documents in the corpus.

Computing the sentiment of a lyrics - poetic text isn't easy as it involves considering the pattern of the words used, sequence of the words etc. Sequence of words play a vital role in determining the sentiment of the lyrics

- While it is possible to analyze the sentiment of a song through its lyrics but a song also has other properties, which are important to figure out its sentiment.

Model Building

SVM

- Support Vector Machines work by drawing a line between the different clusters of data points to group them into classes. Points on one side of the line will be one class and points on the other side belong to another class.
- The classifier will try to maximize the distance between the line it draws and the points on either side of it, to increase its confidence in which points belong to which class. When the testing points are plotted, the side of the line they fall on is the class they are put in.We tried different pipelines on SVM that fetched different accuracies

Naive Bayesian

- The classification is conducted by deriving the maximum posterior which is the maximal P(Ci|**X**) with the above assumption applying to Bayes theorem.
- This assumption greatly reduces the computational cost by only counting the class distribution.

Gradient Boosting

- A prediction model in the form of an ensemble of weak prediction models, typically decision trees
- We tried only the pipeline with " Unigrams TFIDF Vectorization and Lyrics Polarity Value" feature and fetched the accuracy

The dataset we used, could be found here: https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks?select=data.csv

Conclusion

The Music Sentiment Analysis Project helps in predicting the mood/sentiment of the song. It can be used as a recommendation system when integrated with music app to recommend songs of similar mood/sentiment to the user.

Future Work

- Merge the ML model with an existing music application to provide better recommendation of songs to users.
- To provide different tabs for recommendation of songs like by mood, genre, artist, etc
- To add more types of sentiments/ labels.