Music Genre Classification by Lyric Analysis

Purvank Patel (A20380792) Kartik Prakash (A20376570) Sarthak Anand (A20389087)
ppatel104@hawk.iit.edu kprakash1@hawk.iit.edu sanand13@hawk.iit.edu

Problem Overview:

Song lyrics show very specific properties such as rhyming verses and having different frequencies for certain parts of speech. Each genre possesses properties specific to its genre. This is where Lyric Analysis can be used to take advantage of those properties to efficiently classify songs by their genre

Data:

We are using data provided on Kaggle.com, which contains 380,000 lyrics with their genres. This data was initially collected from Metrolyrics.com. We are also building a new dataset from scratch by downloading lyrics from songlyrics.com and metrolyrics.com. In the new dataset we are also including year as one of the attribute which we will use to generate some interesting facts. Following is the link to the data on Kaggle :
https://www.kaggle.com/gyani95/380000-lyrics-from-metrolyrics

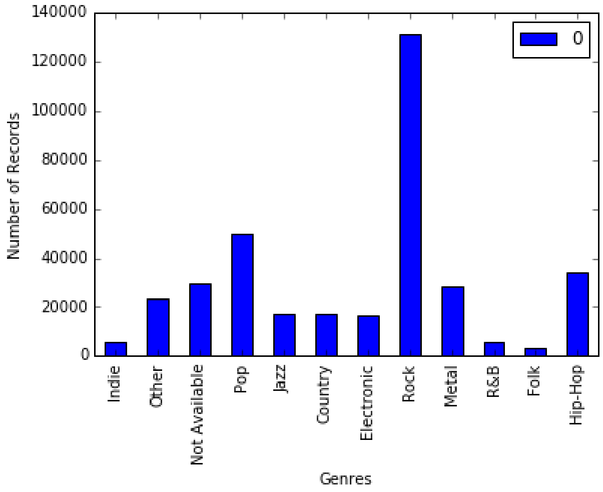
Method:

We will use classical bag-of-words, part of speech features and text statistical features like tf-idf. Genre classification will be done using naïve bayes, k-nearest neighbor with different values for k, SVMs with linear and polynomial kernels, gradient boosting with different parameters, and neural network with different parameter setting. We plan to use built in libraries in python such nltk, ftfy, enchant, scikit-learn, numpy, and scipy etc.

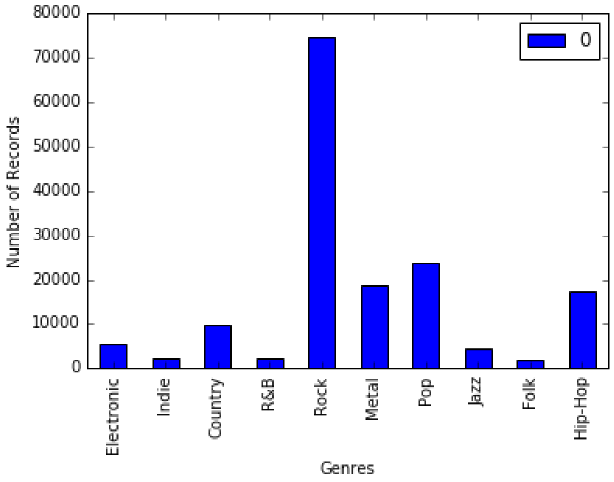
Intermediate/Preliminary Experiments & Results:

We have collected data from songlyrics.com, and grouped it with the kaggle dataset. We dropped instances which doesn't have genre. Also, we are going to focus on four genres: Rock, Hip-Hop, Pop, and Metal, because our data has a good number of songs of these genres to effectively represent them. We dropped instances which had non-english letters in their lyrics. Also, we have performed stop-word removal, bag of words, and tf-idf counts for different genres.

a) Before Preprocessing:



b) After Preprocessing:



Related Work:

1. *Integration of Text and Audio Features for Genre Classification in Music Information Retrieval* by Robert Neumayer and Andreas Rauber [1]

   The approach to solving the problem in this paper involves using lyrics as well as audio features in order to classify song genres using a corpus created from audio and song lyrics file from of a collection of music. The audio features of the corpus were computed using models such as Rhythm Patterns , Statistical Spectrum Descriptors and Rhythm Histograms. The lyric features were computed using bag of words and a tf-idf.

2. *Multimodal Music Mood Classification using Audio and Lyrics* by Cyril Laurier, Jens Grivolla, Perfecto Herrera [2]

   The approach used in this paper revolves around identifying the mood (Angry,Happy,Sad,Relaxed) of the song based on audio and lyrics. The audio classification is done using SVM, Logistic Regression and Random forest on tonal, rhythmic and temporal descriptors of songs. The lyric classification was done by applying k-NN on bag of words model.

3. *Song Genre and Artist Classification via Supervised Learning from Lyrics* by Adam Sadovsky, Xing Chen [3]

   The approach used in this paper is similar to what we are doing. We are going to use similar features such as part of speech, and bag of words. The difference between our approach and this paper is that they have used only maxent, and svm, whereas we are going to experiment with different models such as k-NN, gradient boosting, and neural networks.

4. *Semantic Analysis of Song Lyrics* by Beth Logan, Andrew Kostisky and Pedro Moreno [4]

This paper relies on the similarity of documents or in this case songs in order to classify them. The author uses PLSA to determine the most popular words for each genre then classifies new instances on songs based on the occurrences of those words. The author tries this classification using various number topics (concepts) from PLSA each with and without stemming.

5. Musical Genre Classification by ensembles of audio and lyrics features by Rudolf Mayer and Andreas Rauber. [5]

In this paper, it uses a really small dataset of around 3000 songs, whereas we are using a dataset which has a good representation of different genres. We are doing model selection and feature set selection manually, whereas in this paper they are using ensembling methods to choose the best performing feature set and classification algorithm

Who does what:

Sarthak Anand: Analysis of related work, Analysis of processed data
Purvank Patel: Data Collection, Pre-Processing of data,
Kartik Prakash: Pre-Processing of data, Analysis of data before processing

Timeline:

We are almost through with the pre processing and data collection phase of our project. We have used different libraries like ftfy and enchant to clean the data in order to apply different algorithms for further classification. Next phase of our project will be implementation of features and evaluating results by applying different algorithms and models on our clean data. After evaluation and comparison of different models with baseline we will present the results with graphs and charts. We plan to finish evaluating results in the next two to three weeks.

References:

1.*http://link.springer.com/chapter/10.1007/978-3-540-71496-5_78*
2.*http://ieeexplore.ieee.org/document/4725050/*
3.*https://nlp.stanford.edu/courses/cs224n/2006/fp/sadovsky-x1n9-1-224n_final_report.pdf*
4.*https://pdfs.semanticscholar.org/95ed/bdb583a87278c3757993a5b8078d1013bb3c.pdf*
5.*http://ieeexplore.ieee.org/document/5952341/*
6.*https://pdfs.semanticscholar.org/e658/ec86e033aae370ba680118a04431071cafe1.pdf*
7.http://cs229.stanford.edu/proj2012/BourabeeGoMohan-ClassifyingTheSubjectiveDetermining GenreOfMusicFromLyrics.pdf