BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI

HYDERABAD CAMPUS

A Project Report On

Mood Clustering of songs based on lyrics

BY

Ridam Jain

2013B5A7841

Under the supervision of

Dr. Aruna Malapati

SUBMITTED IN FULFILLMENT OF THE REQUIREMENTS OF

CS F377: DESIGN PROJECT

ACKNOWLEDGMENTS

I am highly indebted to Dr. Aruna Malapati for her help and guidance throughout the project. Her guidance has helped us understand various implementation methods and help us realize various design challenges.She supervised our project and understood and solved various problems related to the DOP.We will like to work under her guidance in future as well.

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI

HYDERABAD CAMPUS

CERTIFICATE

This is to certify that the project report entitled "**Mood clustering of songs based on lyrics**" submitted by Mr. Ridam Jain (ID No. **2013B5A7841**) in fulfillment of the requirements of the course **CS F377 - Design Project** embodies the work done by him under my supervision and guidance.

Date: 1 MAY 2017 **(Dr. Aruna Malapati)**

BITS- Pilani, Hyderabad Campus

ABSTRACT

This report seeks to explain the database collection and feature extraction for using appropriate Machine learning algorithms.Further more it also explains different clustering techniques that were used categorize songs into 5 clusters based on various machine learning models and elaborated on different techniques used to increase the accuracy.

CONTENTS

1.Introduction

Songs can be clustered and categorized based on various attributes like mood artist genre etc.But humans are very emotional beings and unconsciously our decision making and choice of things depend upon mood or emotions that person is feeling.

This project aims to implement and design various techniques to cluster songs into broad mood categories based on emotions expressed in lyrics of the songs.

In the first phase of the project lyrics were gathered from the internet along with their mood tagged.

Then in the second half we applied different models on the dataset with different features each time to check the variations in the accuracies and effect on correct prediction of cluster when lyrics or other relevant feature is provided.

2.Database Construction

For database construction we used mood tagging of allmusic.com and lyrics from metrolyrics.com and azlyrics.com to compile the database.We used beautiful soup for web crawling. The details of database constructions were already been discussed by mid-semister report as well as further elaborated in Aakash's report.

The database contains five attributes namely :

Name	Artist	Mood	Cluster	lyrics

Till the first part of the project cluster of the songs were not defined.

Further the songs were clustered based on their moods, there similarities were also obtained from allmusic using some crawling techniques.

Cluster 1	Passionate,Earnest,Dramatic,Rousing,Romantic,Freewheeling,Theatrical,Reverent,Joyous,Exuberant,Energetic,Sensual,Organic,Lush,Earthy,Brash,Raucous,Rambunctious,Boisterous,Rowdy,Confident,Carefree,Urgent,Street-Smart,Sexy,Rebellious,Playful,Confrontational,Celebratory,Ambitious,Reckless,Gleeful,Messy,Hedonistic,Manic
Cluster 2	Rollicking,Organic,Exuberant,Earthy,Amiable/Good-Natured,Fun,Freewheeling,Happy,Cheerful,Sweet,Playful,Carefree,Summery,Springlike,Sentimental,Joyous,Gleeful,Earnest,Celebratory,Irreverent,Energetic,Romantic,Gentle,Delicate,Intimate,Laid-Back/Mellow,Naive,Innocent
Cluster 3	Literate,Self-Conscious,Refined,Precious,Ironic,Elaborate,Detached,Complex,Cerebral,Acerbic,Reflective,Poignant,Melancholy,Indulgent,Earnest,Clinical,Bittersweet,Ambitious,Wistful,Sentimental,Searching,Sad,Plaintive,Intimate,Delicate,Brooding,Autumnal,Yearning,Gentle,Restrained,Springlike,Inno

	cent,Sophisticated,Elegant,Somber,Bleak,Angst-Ridden,Nocturnal,Nihilistic,Gloomy,Bitter,Weary,Paranoid,Ominous
Cluster 4	Humorous,Wry,Whimsical,Silly,Playful,Cynical/Sarcastic,Acerbic,Quirky,Outrageous,Irreverent,Ironic,Happy,Gleeful,Freewheeling,Campy,Witty,Theatrical,Carefree,Exuberant,Energetic,Precious,Cerebral,Naive,Indulgent,Refined,Self-Conscious,Trippy,Innocent,Springlike,Druggy,Sophisticated,Snide,Stylish,Detached,Sexual,Reflective
Cluster 5	Aggressive,Confrontational,Visceral,Reckless,Rebellious,Provocative,Angry,Volatile,Thuggish,Tense/Anxious,Street-Smart,Raucous,Rambunctious,Outrageous,Menacing,Malevolent,Intense,Hostile,Harsh,Fiery,Cathartic,Angst-Ridden,Energetic,Urgent,Uncompromising,Paranoid,Manic,Freewheeling,Dramatic,Complex,Cerebral,Brash,Ominous,Somber,Gloomy,Eerie,Bleak,Theatrical,Hypnotic,Elaborate,Druggy,Difficult,Ambitious,Earthy

We were unable to cluster some of the moods those rows were discarded from the database.Also some of the moods were in multiple clusters, as words counts have effect on the machine learning algorithm those rows were appended with same name,artist,moods,lyrics but different cluster.over all the final database that was feeded into the models was of size 7999 * 5 .

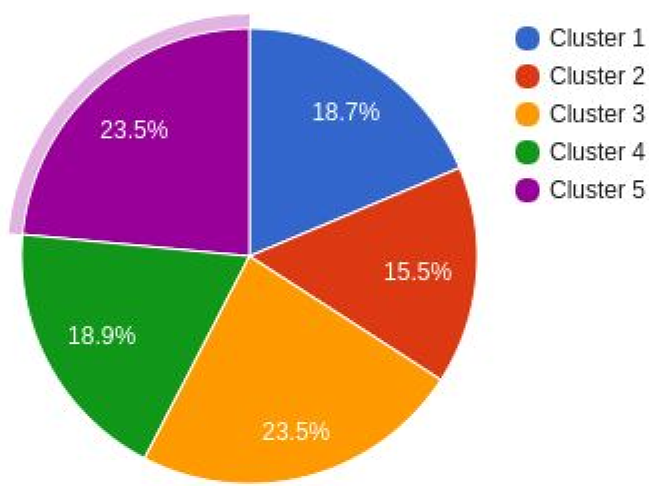
That is 7999 rows and 5 columns (for 5 different attributes of the database)

3.Categorizing songs using machine learning

For categorizing songs we used various approaches. There are some words in lyrics like “i” , “and” , “the” that doesn't mean much but can cause fluctuations in models learning capabilities set of such words are called stop words **.**Akash’s report summarizes the effect on accuracies when stop words are included into the lyrics. This report summarizes the effects when stop words are not included in the lyrics.

3.1 Preprocessing

Songs were evenly distributed among all the 5 clusters, with no cluster having more that 25 % data.

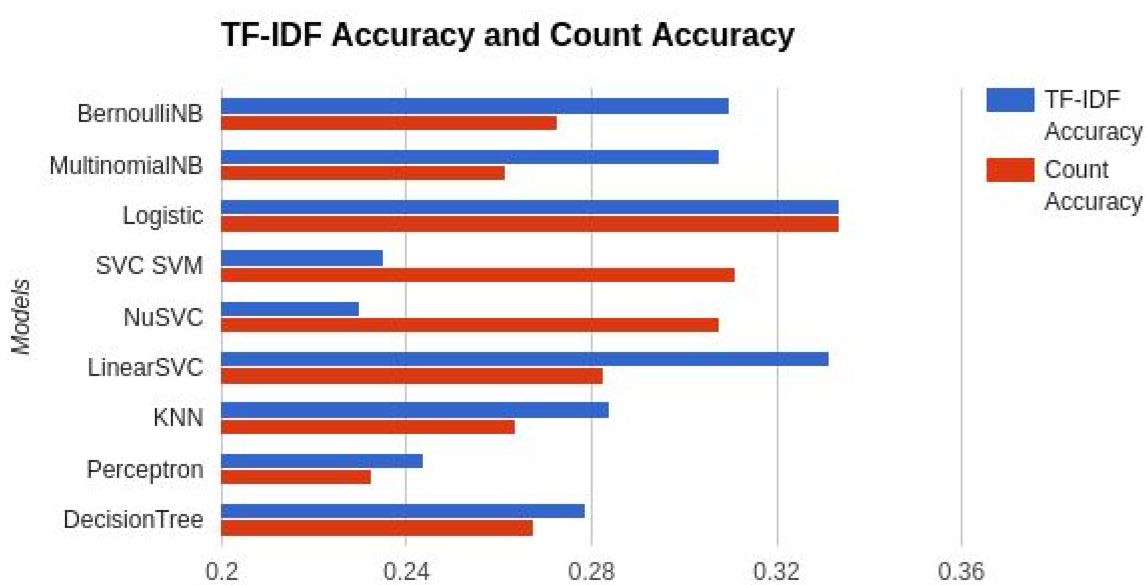


For preprocessing all the stopwords were removed from the lyrics (Akash’s report take care of the case when stop words are not removed) NLTK python library was used to lemmatized and stemmed the lyrics into its root form

3.2 PCA

Dimensionality reduction using PCA is done to a lower dimensional space for data processing and observing effect on different models . One can observe the accuracies as follows:

Models	TF-IDF Accuracy	Count Accuracy
BernoulliNB	0.31	0.2725
MultinomialNB	0.3075	0.26125
Logistic	0.33375	0.33375
SVC SVM	0.235	0.31125
NuSVC	0.23	0.3075
LinearSVC	0.33125	0.2825
KNN	0.28375	0.26375
Perceptron	0.24375	0.2325
DecisionTree	0.27875	0.2675



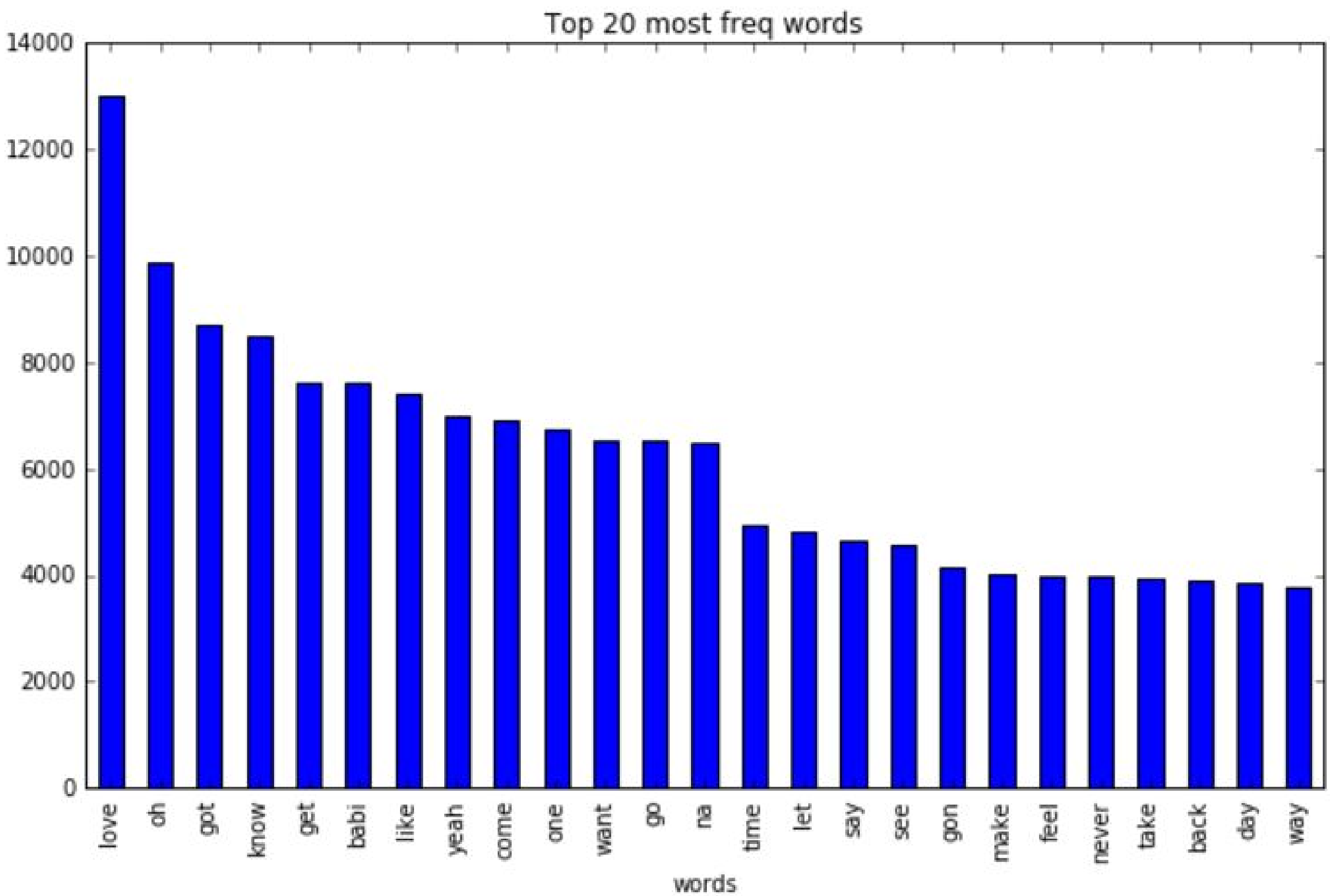
It can be observed that logistics bernoulli and svm have the highest accuracies.Term frequency Inverse document frequency is always greater that count frequency.With PCA dimensionality reduction logistic regression is provides best results.

3.3 Feature selection

When dimensionality of the problem is reduced using feature selection .only the best features that define the dataset are taken and rest all are dropped from the learning process.

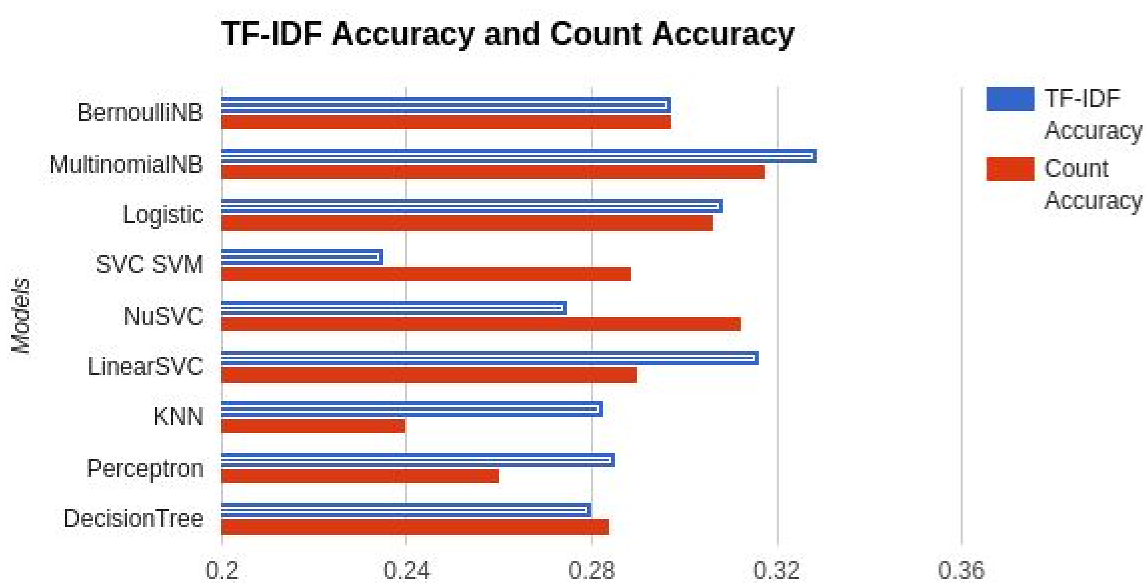
Here topwords or the most frequently occurring words are taken as desired feature .

Without stop words one can observe top 20 words with highest count.



We can observe the word love is most common in all 7999 songs followed by oh got know and so on.we can observe the changes in accuracies with feature selection in use.

Models	**TF-IDF Accuracy**	**Count Accuracy**
BernoulliNB	0.2975	0.2975
MultinomialNB	0.32875	0.3175
Logistic	0.30875	0.30625
SVC SVM	0.235	0.28875
NuSVC	0.275	0.3125
LinearSVC	0.31625	0.29
KNN	0.2825	0.24
Perceptron	0.285	0.26
Decision Tree	0.28	0.28375



Almost every time the tf idf accuracies are more than count accuracy.

Also in this case multinomial naive bayes have better accuracy that PCA and over all the best accuracy amongst all the models

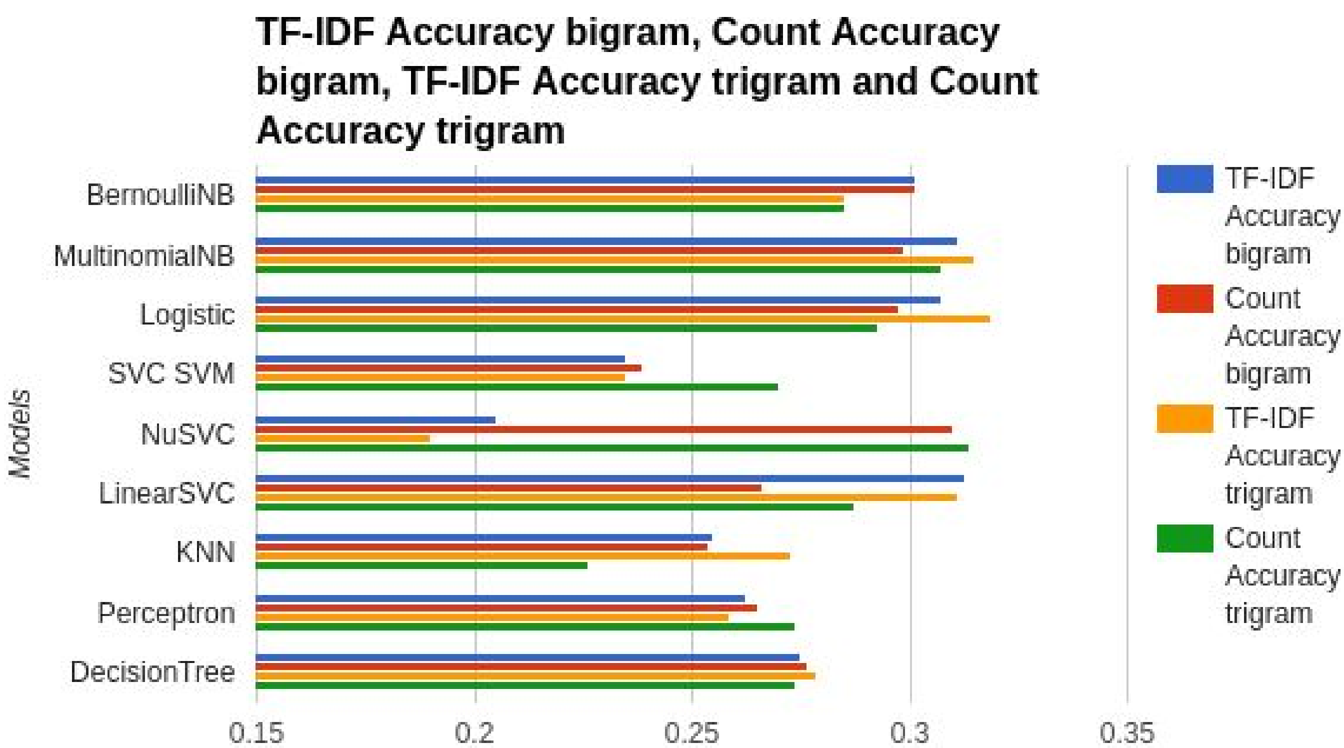
3.4 Bigram and Trigram

Instead of bag of words model one can use bigram and trigram models as these models provide more input as it take cares of sentence formation and positions of words in a sentence unlike bag of words model.making them ideal for emotion recognition , but due to removal of stop words the meaning is again lost, so even though these accuracies may be higher that previously obtained .it is incorrect to use these models without stop word (Akash's report take care of this problem)

Models	TF-IDF Accuracy bigram	Count Accuracy bigram	TF-IDF Accuracy trigram	Count Accuracy trigram
BernoulliNB	0.30125	0.30125	0.285	0.285
MultinomialNB	0.31125	0.29875	0.315	0.3075
Logistic	0.3075	0.2975	0.31875	0.2925
SVC SVM	0.235	0.23875	0.235	0.27
NuSVC	0.205	0.31	0.19	0.31375
LinearSVC	0.3125	0.26625	0.31125	0.2875
KNN	0.255	0.25375	0.2725	0.22625
Perceptron	0.2625	0.265	0.25875	0.27375
DecisionTree	0.275	0.27625	0.27875	0.27375

This time again naive bayes with multinomial distribution assumption outperformed all other models

From the graph below one can compare and observe that no generalization can be made whether trigram or bigram model is better.

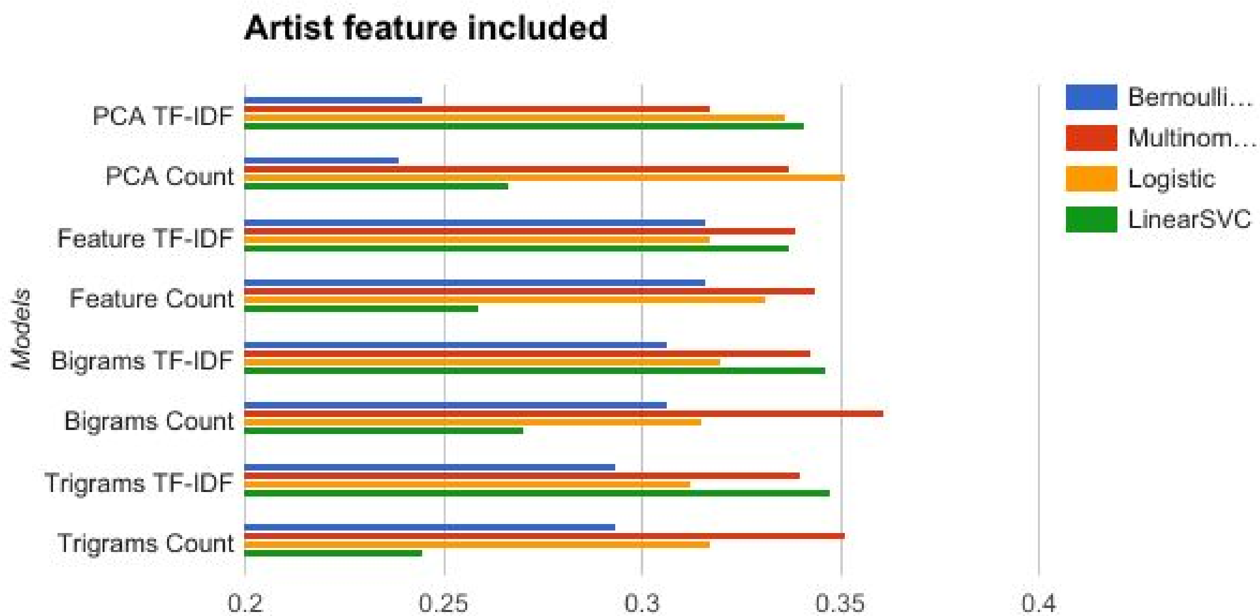


3.5 Artist and lyrics based clustering

Sometimes artists are also biased towards composing songs of same genre and mood , so artist can be another useful feature to categorize songs.As for most of the algorithms only one vector or attribute like word count vector or feature vector is given as input , so to include artist and imcrease its weightage we appended name of the artist repeatedly into the lyrics of the song .

And the results were as follows:

	PCA		Feature Selection		Bi-grmas		Tri-grams	
	TF-IDF Accuracy	Count Accuracy	TF-IDF Accuracy	Count Accuracy	TF-IDF Accuracy	Count Accuracy	TF-IDF Accuracy	Count Accuracy
BernoulliNB	0.245	0.23875	0.31625	0.31625	0.30625	0.30625	0.29375	0.29375
Multinomial NB	0.3175	0.3375	0.33875	0.34375	0.3425	0.36125	0.34	0.35125
Logistic	0.33625	0.35125	0.3175	0.33125	0.32	0.315	0.3125	0.3175
LinearSVC	0.34125	0.26625	0.3375	0.25875	0.34625	0.27	0.3475	0.245



One can observe that multinomial and logistic regression performing very well after addition of artist as attribute also Bigrams count frequency is 36.125% that is the highest accuracy we got.

4.Conclusion

For natural language processing it is becoming very important to understand emotions in text and voice for making more life like AI and assistants.We concluded that clustering of songs is possible into moods based on lyrics and artist as feature.After analysis it can be easily seen Multinomial naive bayes and logistic regression to outperform all other algorithms in the list.
With over all best count accuracy with artist as feature to be around 36% . that is our model can cluster songs into various mood categories with 36% accuracy .
For future improvement of the project better dataset and cluster definitions can be used and a transition algorithm can be implemented as well to navigate from one mood to another by generating a queue of songs using different shortest path algorithms on weighted graph.

5. References

[1] Ujlambkar, Amey, et al. "Mood based music categorization system for bollywood music." International Journal of Advanced Computer Research 4.1 (2014)
[2] Dang, Trung-Thanh, and Kiyoaki Shirai. "Machine learning approaches for mood classification of songs toward music search engine." Knowledge and Systems Engineering, 2009. KSE'09. International Conference on. IEEE, 2009.
[3] http://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html
[4] http://scikit-learn.org/stable/tutorial/basic/tutorial.html
[5] http://www.nltk.org/book/ch01.html