Predicting Genre of a Song by its Lyrics

By: Belal Taher, Rick Mortenson, Frederic Goguikian, Naman Agarwal

Description/Problem Statement

For our project, we explored whether or not its possible to predict the genre of a song using Natural Language Processing. We discussed using features such as word count, sentiment analysis, and n-grams but ultimately decided on a TF-IDF feature vector for all words present in the corpus.

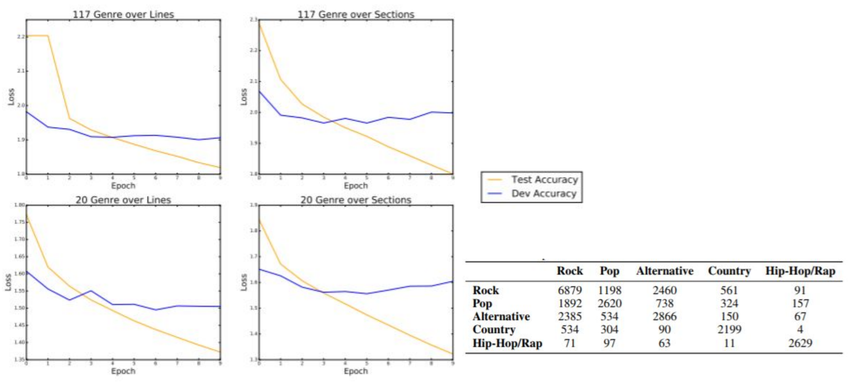
The specific hypothesis we are tackling in this project is whether or not we can use a TF-IDF feature vector of a given song's lyrics to predict that song's genre. Before we explore our methods and results, we will examine related work.

Related Work

Differences from our work

The problem we are tackling in our project is extremely similar to one addressed in a study published by Stanford titled "Music Genre Classification by Lyrics using a Hierachical Attention Network." The primary differences between our project and this one is that we are using different features and different classifiers. Below are visualizations illustrating their results.

Results



	Rock	Pop	Alternative	Country	Hip-Hop/Rap
Rock	6879	1198	2460	561	91
Pop	1892	2620	738	324	157
Alternative	2385	534	2866	150	67
Country	534	304	90	2199	4
Hip-Hop/Rap	71	97	63	11	2629

The visualizations above illustrate this studies' results. They found that, while it is possible to predict the genre of a song to some degree, classification by lyrics will always be inherently flawed because of vague genre boundaries with many genres borrowing lyrics and style from one another.

Data Collection & Cleaning

In order to have enough data to test our hypothesis, we searched Kaggle, a massive online repository of databases for a dataset of 380,000+ songs. However, this data needed to be cleaned and the lyrics needed to be preprocessed so we could build feature vectors. Below is a list of steps we took to clean our data:

- Removed all songs that were missing any required field
- Removed all songs with genre labeled as "Other"
- Removed all songs belonging to a genre of < 10,000 songs
- Removed special characters ("!","?",".", etc)
- Removed stopwords from all remaining lyrics
- Tokenized lyrics to lemmatize and stem words

Word	Stem	Lemma
Studies	Studi	Study
Hating	Hat	Hate
Fought	Fought	Fight

Methods

TF-IDF

As we previously mentioned, we used a TF-IDF feature vector for our classifiers. The reason we chose TF-IDF is because it's a measure of how important a term is to the meaning of a text document. We believe that these important terms will not just describe the song in question but also the genre due to the similarity between songs in the same genre. Below is a explanation of how to calculate the TF-IDF value of a term.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

tf_{ij} = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Dataset after cleaning

After cleaning our dataset, we were left with ~200,000 songs. The five genres that were left and that we will try to classify are Rock, Pop, Hip-Hop, Metal, & Country. The distribution of songs across these genres was as follows:

Genre	Number of Songs
Pop	40464
Hip-hop	24782
Rock	109218
Country	14386
Metal	23758

Note that the distribution is not uniform. To avoid bias in the corpus skewing our results, we chose 10,000 random samples from each genre to train and test with. The reason we chose 10,000 is because that was the largest number we could use without exceeding the number of songs present in the smallest genre.

Algorithm to create feature vector

```
tf_idf[*][*] = 0
for doc in all_docs:
  for word in doc.words:
    tf[word] = doc.findCount(word)/doc.findCount(*)
    idf[word] = all_docs.length/…
        all_docs.thatContain(word).length
    tf_idf[word] = tf[word]*idf[word]
```

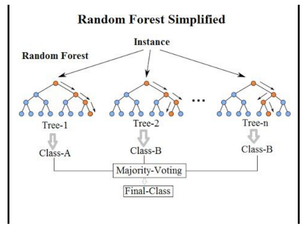
The pseudocode above provides implementation details of our feature pre-processing pipeline. After we created our feature vectors, we trained three different kinds of classifiers and juxtaposed their accuracy.

Classifiers used

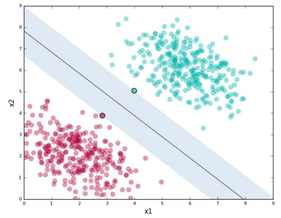
Multinomial Bayes

$$P(L \mid \text{features}) = \frac{P(\text{features} \mid L)P(L)}{P(\text{features})}$$

Random Forest



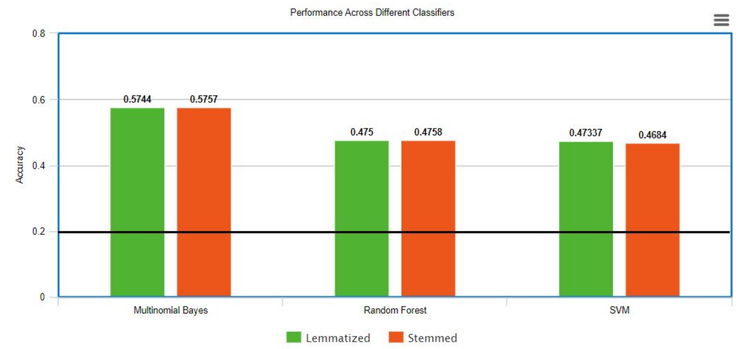
SVM



Results

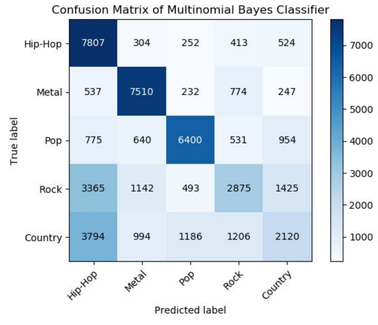
Results across different models

After building and training our models enough times to cross-validate, we took the average accuracy of 5 runs to get an average accuracy. We then juxtaposed the average accuracy of the different models to see if model choice makes a difference. The results of our juxtaposition are shown below:



The results we found very strongly supported the results of the related work we found. We were able to predict the genre to some degree (i.e. our classifiers performed better than randomly guessing), and certain classifiers are better than others, but the ambiguous definition of genres and how songs are placed in them still remain a massive problem in trying to predict the genre of a song by its lyrics. We also need to consider the fact that genres are defined by more than just the lyrics involved. It's perfectly possible for two songs to have the same exact lyrics but a different instrumental that causes them to be placed in different genres.

Confusion Matrix of Multinomial Bayes



If we look at the confusion matrix above, we can see that genres with similar lyrics are commonly mistaken for each other. For example, Hip-Hop and Metal have, arguably, the most unique lyrics since they involve darker themes/words and slangs. However, Pop, Rock, and Country are all genres that have similar lyrics which is why our classifier misclassified Country as Rock/Pop in the same proportion among various other misclassification errors.

Conclusion

Based on our experimentation, we conclude that classifiers do better than randomly guessing when trying to predict a song's genre by its lyrics, but they don't objectively predict well. As mentioned above, we believe this is due to some genres having very similar lyrics (i.e. Pop/Rock/Country), and this point is bolstered by the confusion matrix since it shows us what portion of examples in a given genre were misclassified and how they were misclassified. Ultimately, we conclude that our original hypothesis is false. We cannot reasonably predict the genre of a song based on its lyrics. This is in line with the conclusion from the related work we found.