Genre Classification

Melody Jeng and Kyle Shankwiler

Abstract

We researched the best methods to classify songs by their genre, using bag-of-words and repetitiveness as features for our models. We found that our three selected genres could be classified well with a bag-of-words model. Repetitiveness was less successful, though we found hip hop was the least repetitive.

1 Introduction

People are quite picky with their music and everyone has their own favorite style. As a result, derogatory stereotypes of other genres are often thrown around, such as country songs always being about whiskey, dirt, and boots. To investigate the truth of such claims, it is relevant to analyze song lyrics and their ability to be classified into genres. Additionally, genre classification could help music providers make educated guesses as to how to categorize music that has no given genre.

2 Prior Work

After determining our topic, we found that there had been a few previous studies that also analyzed lyrics in ways that might be of interest to us. First, we found a study done by students at Stanford doing something very similar (Sadovsky and Chen). Their best success was using a bag-of-words model.

We found a blog asking the question of whether pop lyrics are getting more repetitive (Morris). It used the LZ77 compression algorithm to measure the compression ratio of different songs, and infer the repetitiveness from that.

3 Approach

We used a library called spotipy to make queries to the Spotify API, requesting tracks by different genres. Using the given song title and artist name for the songs that were returned, we queried lyrics.wikia.com using a library called PyLyrics. We collected as many lyrics as possible. This took several hours but we ended up with 7377 songs from each of the following genres: country, hip hop, and metal.

To process the lyrics, we first converted all the words to lowercase. Then we used regular expressions to replace all sequences of digits with the word NUM, as we did not think that the value of the numbers would be relevant in determining song genre. We did, however, think that perhaps the existence of such numbers would be informative, so we did not choose to remove them entirely. We then removed all other non-alphanumeric characters. We also removed stop words, and lastly stemmed everything using NTLK's PorterStemmer. Much of this was inspired by a Kaggle tutorial on implementing a bag of words.

We then began by looking at the bag of words model. We followed the same Kaggle tutorial to vectorize our cleaned input using sklearn's CountVectorizer and RandomForest as our classifier. We found the 5000 most frequently occurring words in the training data, and used them as features for the input to the RandomForest classifier. RandomForest creates multiple decision trees from different subsets of the training data, and averages across them.

We also wanted to measure the repetitiveness of a song. We tried many different measures to see which was most effective. For each measure we used the distribution of the occurrences of each word in the song.

1. The first measure was simply the number of words occurring more than once compared to the number of distinct words.
2. The second measure was based off an article that measured lyrics' repetitiveness using the LZ77 compression algorithm. We took each song and compressed it using an implementation of the compression algorithm found on Github. We then found the ratio be-

tween the length of the compressed song and the original and used that as an indication of the song's repetitiveness.

3. The third measure we tried was the average number of occurrences of words in a song.
4. The fourth measure was the number of times the maximally occurring word occurred, compared against the total number of distinct words.
5. For the fifth measure, we added up the total number of occurrences of words that occurred more than once compared to the total number of words.
6. We measured the kurtosis of the word distribution, which is a measure of the "tailedness" of the distribution. In this case the tail includes the words with a low number of occurrences. A song with a lot of words that occur only once and a few words that occur more often would have a larger kurtosis value than a song with words that occur the same number of times.

4 Results

The bag of words model with our cleaned data reliably gave us an error rate between 10 and 12 percent. Hip hop had the best F-measure, and while country and metal were often confused with one another, they also had F-measures over 0.8.

Our repetitiveness measures showed various results. Maximally occurring word was the most effective, but only when used with 5-grams. Interestingly, though this measure worked best with 5-grams, 5-grams were not the best for any of the other measures. Compression was by far the worst, with F-measures below 0.4 in all genres.

In all repetitiveness cases, hip hop had the best F-measure, regardless of the type of model used. This is because hip hop is less repetitive than both country and metal, which are both similarly repetitive. For example, using the kurtosis score on the training and test data, the average across hip hop was 30.72, while country was 10.34 and metal was 10.90. For this reason, any classifier we created using repetitiveness as a feature more easily distinguished hip hop from the others, but had difficulty between country and metal.

There were also relationships between size of n-gram and genre. For metal, there was a consistent decrease in F-measure with increased n-grams across the different repetition measures. Larger n-grams resulted in poorer classification for metal. Hip hop F-measure dropped specifically at 4-grams, but rebounded at 5. Country seemed to be the same regardless of size of the n-grams.

5 Discussions

In attempting to classify lyrics by genre, we used a bag-of-words model as well as six different measures of repetitiveness using different sizes of n-grams in a random forest decision tree classifier.

5.1 Hip Hop

We found that hip hop differed significantly from both country and metal. Using the bag-of-words model, country and metal were more often misclassified as each other than either was mistaken for hip hop. With repetition, hip hop was the least repetitive, and easily divisible from the others, which had very similar overall repetition measures. As such, when classified using only repetitiveness as a feature, hip hop songs were almost always classified as hip hop, while country and metal were classified seemingly at random.

5.2 N-Grams

We also looked into how n-grams would affect classification and found that each genre reacted differently. Metal showed a decline at 5-grams, though it was consistent at other lengths, while hip hop dipped at 4 but came back at 5. Overall, n-grams do not seem to have a large effect on the classification, as different n-gram sizes pretty much never varied by more than 0.1.

5.3 Difficulties

One of our biggest hurdles was obtaining data. First we had trouble finding a way to get songs that had already been properly labelled with their respective genres. Our options sometimes gave us foreign language songs, which we had decided not to analyze due to the complexity of incorporating different languages into our chosen models, or songs that we had trouble finding accompanying lyrics for. After we finally settled on spotipy and PyLyrics, it took us several hours to fully acquire all the data we wanted, and even then, some re-

quests fell through, leaving us with unequal amounts for each genre.

The most egregious problem we encountered in data collection actually occurred very late in the game, when we discovered that our data contained a large amount of duplicates. We are still not entirely sure how this issue arose, but as a result, we lost about 10% of our original data.

Another challenge we faced was repetition. Since we did not know how to go about measuring such a phenomena, we resorted to trying multiple measures to determine which was most effective. The proved to be somewhat underwhelming, as most of our measures yielded results that were not as accurate as we had hoped.

6 Conclusion

Our work revealed that repetitiveness alone is not informative enough to classify songs effectively, but bag of words is extremely effective. N-grams do not affect overall classification when used with repetitiveness.

As for insights we gained in looking at the specific genres we chose, we found that hip hop is quite different from country and metal. It was the least repetitive by far, and country and metal had about the same repetitiveness. Content-wise hip hop was also quite distinctive, as country and metal were more often classified as each other rather than as hip hop.

In the future, we would like to compare words found in each genre with how common they are in the English language, incorporating different English corpora into the analysis.

References

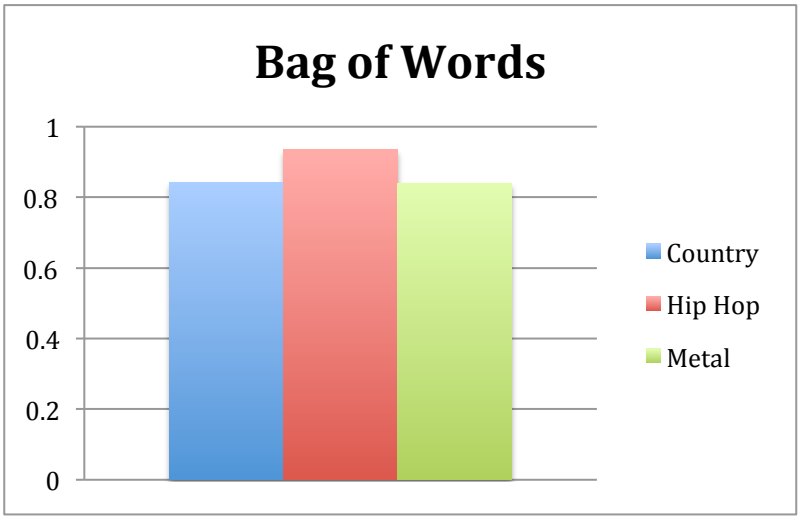
Morris, Colin. "Are Pop Lyrics Getting More Repetitive?" *The Pudding*. N.p., 2006. Web. 11 June 2017.

Sadovsky, Adam, and Xing Chen. *Song Genre and Artist Classification via Supervised Learning from Lyrics. The Stanford Natural Language Processing Group.* N.p., n.d. Web. 26 May 2017.

Data

Bag of Words Results (for unigrams only)

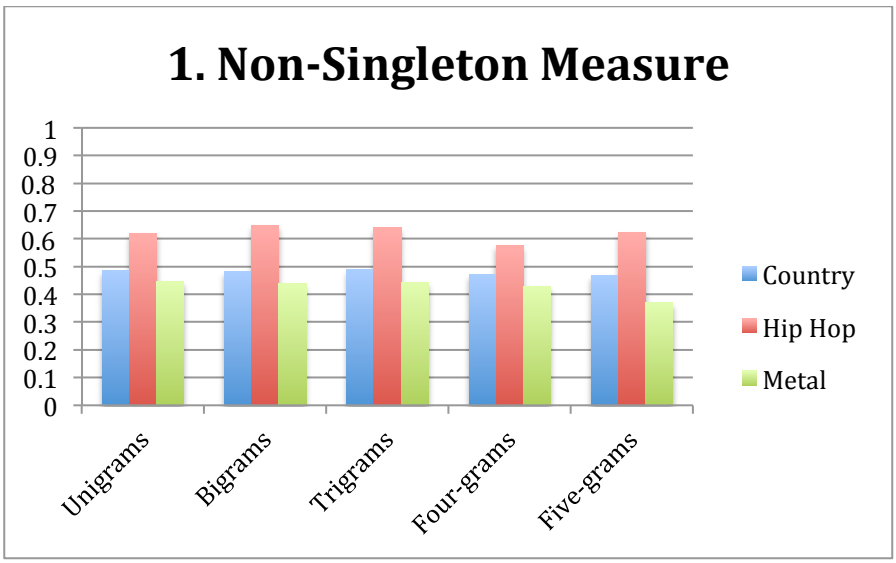
Country	Hip Hop	Metal
0.842	0.936	0.839

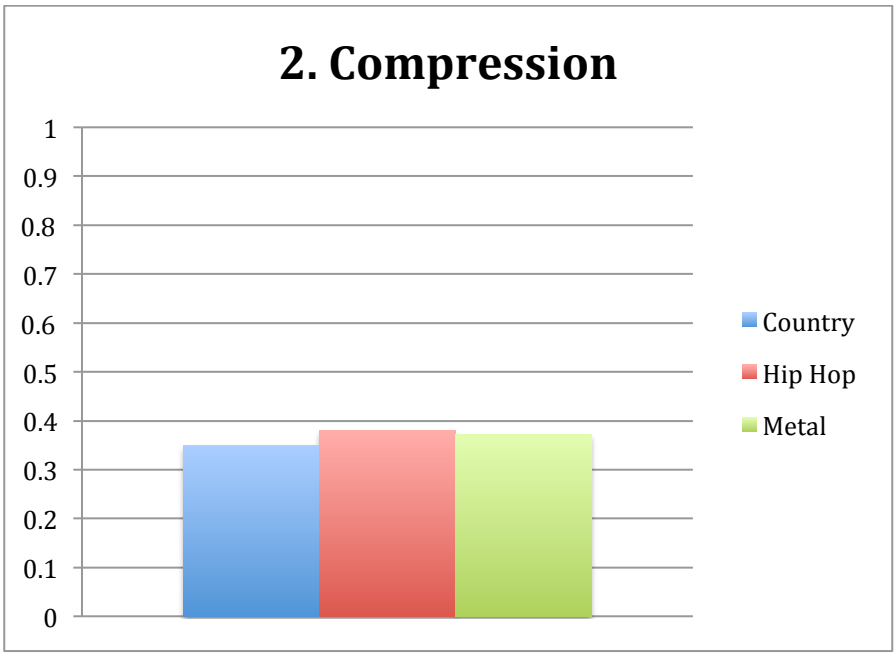


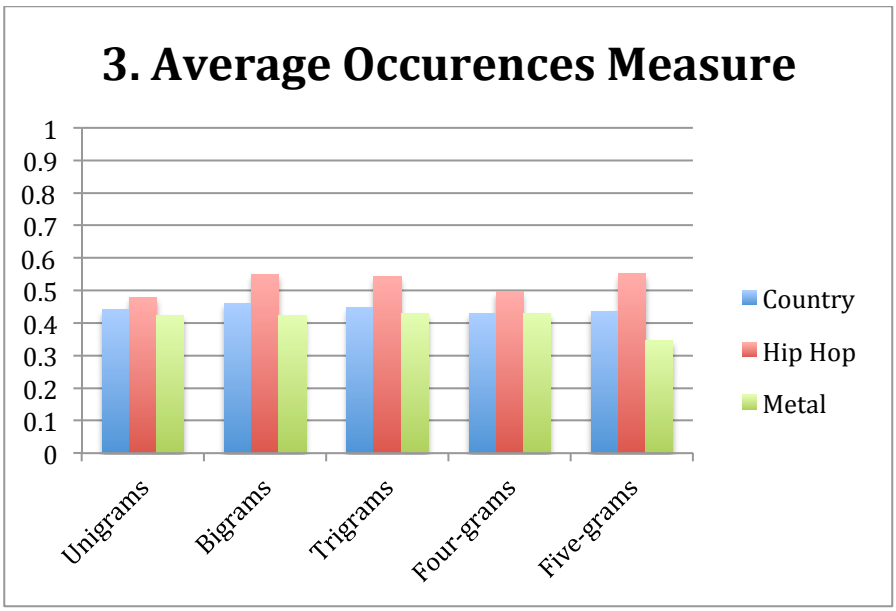
Repetitiveness Results

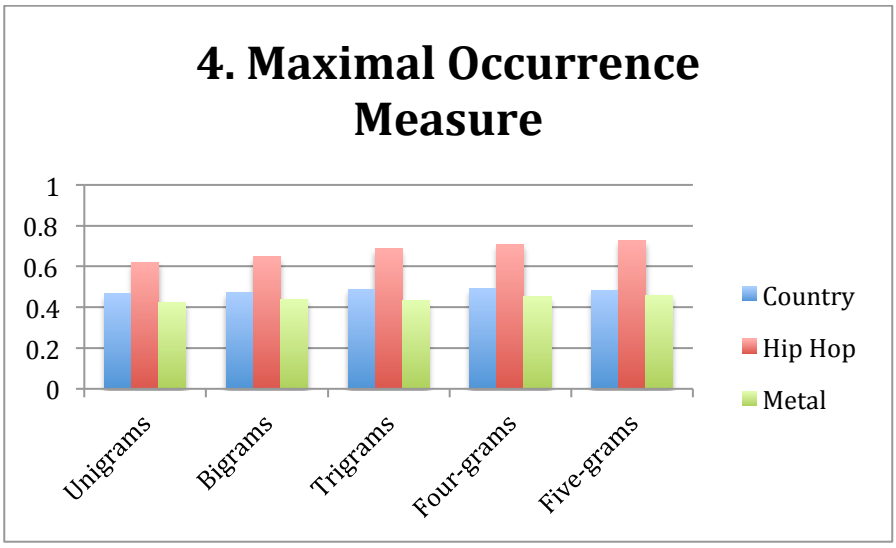
Note: We only measured unigrams with compression, as higher n-grams would produce unreliable results. Each n-gram in close proximity with one another will simply compress together. All other repetitiveness measures were tested with uni- to 5-grams.

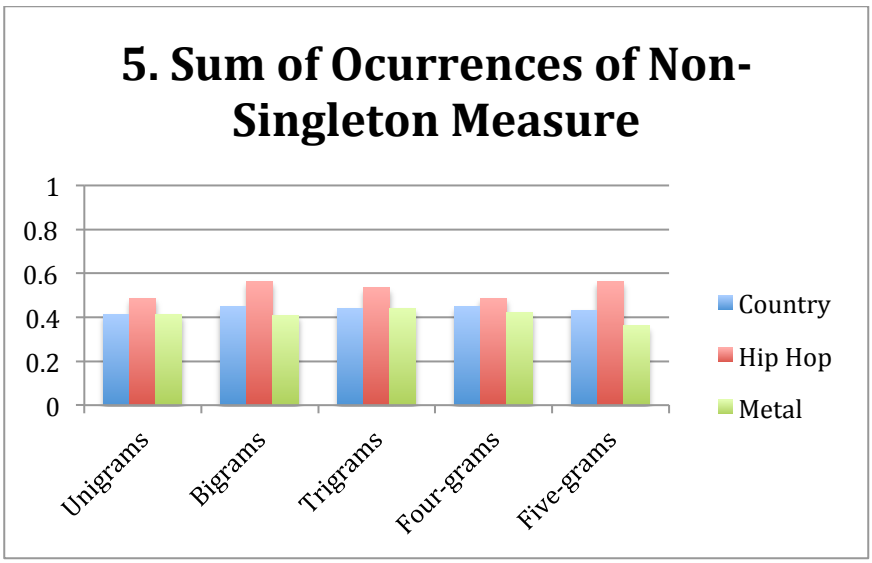
	Vectorizer	1	2	3	4	5	6
Word Types							
Unigrams		(0.486,0.619,0.448)	(0.842, 0.936, 0.839)	(0.441,0.477,0.424)	(0.466,0.62, 0.425)	(0.414,0.486,0.412)	(0.387,0.511,0.39)
Bigrams		(0.484,0.649,0.44)	-	(0.46,0.548, 0.424)	(0.47,0.649, 0.44)	(0.45,0.564, 0.407)	(0.41,0.569, 0.418)
Trigrams		(0.488,0.641,0.444)	-	(0.448,0.544,0.428)	(0.489,0.686,0.432)	(0.44,0.534, 0.438)	(0.418,0.543,0.438)
Four-grams		(0.472,0.577,0.43)	-	(0.43,0.496, 0.43)	(0.49,0.709, 0.451)	(0.45,0.486, 0.422)	(0.435,0.518,0.412)
Five-grams		(0.469,0.623,0.37)	-	(0.437,0.551,0.347)	(0.481,0.727,0.46)	(0.432,0.561,0.361)	(0.451,0.568,0.343)

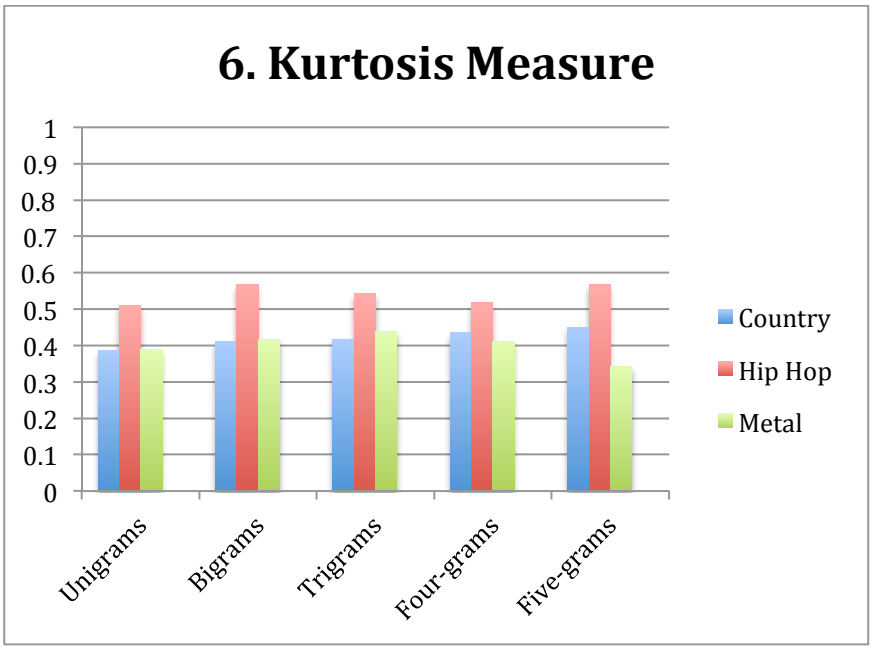

1. Non-Singleton Measure
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Unigrams
Bigrams
Trigrams
Four-grams
Five-grams
Country
Hip Hop
Metal


2. Compression
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Country
Hip Hop
Metal


3. Average Occurences Measure
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Unigrams
Bigrams
Trigrams
Four-grams
Five-grams
Country
Hip Hop
Metal


4. Maximal Occurrence Measure
1
0.8
0.6
0.4
0.2
0
Unigrams
Bigrams
Trigrams
Four-grams
Five-grams
Country
Hip Hop
Metal


5. Sum of Ocurrences of Non-Singleton Measure
1
0.8
0.6
0.4
0.2
0
Unigrams
Bigrams
Trigrams
Four-grams
Five-grams
Country
Hip Hop
Metal


6. Kurtosis Measure
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Unigrams
Bigrams
Trigrams
Four-grams
Five-grams
Country
Hip Hop
Metal