Whose Rap is it Anyways?

Determining Hip-Hop Artists from their Lyrics

CS 229: Machine Learning, Project Milestone

Alex Wang, Robin Cheong, Vince Ranganathan
{jwang98, robinc20, akranga}@stanford.edu

November 22, 2017

1 Problem Statement & Motivation

Authorship identification and verification is a task that is becoming increasingly relevant due to machine learning. The concern of this task is to select the true author from a given set of possibilities, based on characteristics of the text. Current algorithms exist to identify authors of prose using tools such as natural language processing [1], support vector machines [2], and deep learning [3]. Rap, however, is a drastically different form of writing from prose. While it shares many commonalities, it also includes aspects such as rhythm, rhyme, versatile sentence structures, and contemporary references, and is frequently focused on a different and smaller subset of themes. It is relatively modern, having been growing in popularity over the last two decades, and is arguably the most popular form of music in the United States at the time of writing.

Our task is to **identify the rapper/lyricist of a given verse from a given set of rappers**. This will involve gathering appropriate amounts of data, selecting a training and classification method, and using this method to build and train the algorithm. The scope of the project is currently limited to 12 rappers with sufficiently extensive discographies (five or more studio albums or mixtapes).

Part of our motivation for undertaking this exploration is that we are interested in using error analysis to reveal and explore fascinating and non-obvious characteristics about rap lyrics: If the algorithms is consistently struggling to choose between two lyricists, how truly unique are their verses? How does the algorithm perform on songs that are known to us to be ghost-written (i.e. worded by an individual other than the performer), and can this be used to identify the widely frowned-upon ghost-writing? What makes each rappers' lyrics special and unique (or in other words, which features does the algorithm choose to weight most heavily)? We hope that this research will shed light on these questions.

2 Data Collection and Processing

2.1 Data Collection

In order to retrieve the lyrics, we used the Genius API and PyGenius to find the list of songs of a certain artist. Unfortunately, the Genius API does not provide lyrics, so we used BeautifulSoup and scraped the lyrics off of the webpage for the song on Genius. Then, we stored all of the scraped data in csv format for later use in feature extraction.

2.2 Pre-processing

In terms of data processing, we *removed all punctuation* from the lyrics. This simplified the feature extraction process as we no longer had to worry about punctuation at the end of the words. We also decided to apply *stemming* to all of the words present using the nltk, since some words, especially verbs, have many different endings even though they be used in the same way. For example, "walk", "walks", and "walked" all refer to the same action of walking, and it would be very unnecessary for these word to be treated as different features.

We noted that neither of these actions - punctuation removal and word stemming - affected words or sounds that are included in the lyrics but not present in the English language (or are considered 'explicit'). This is important, as we would not want pre-processing to affect the validity of our results in any manner.

2.3 Feature Extraction

For our initial testing, we extracted only a single feature from each text input: the **vocabulary**. This is sparse vector representing the correspondence between each word used in the input and the number of occurrences of the word. This single feature alone (which is actually a group of sub-features) was sufficient for preliminary testing.

The program is also capable of extracting a few other features:

- **Vocabulary richness**: fraction of words that are unique
- **Explicit language**: counts of 'explicit' language in a verse
- **Length**: average word length, average line length

These features, however, have yet to be incorporated into our testing phase.

3 Methods

Thus far, we've implemented two baselines for our project using the features and data given above: Naive Bayes and Softmax Regression.

3.1 Naive Bayes

Our reasoning behind trying Naive Bayes is that NB is a fairly simple classification algorithm to implement, and for our current set of features (number of occurrences of each word), it seemed well suited to the task, especially considering that given a specific rapper, it might be reasonable to think of their selection of words as independent.

3.2 Softmax Regression

We tried Softmax Regression as just another multi-class classification algorithm, with the hopes that as we developed better features, we could outperform Naive Bayes.

4 Preliminary Testing & Results

Our Naive Bayes algorithm gives an overall test set accuracy of **.3137%** and a training set accuracy of **92.6%** when using a 66% / 33% train-test split, which is significantly above random chance (.083%), but still far too low to place any confidence in the algorithm's predictions.

Our Softmax algorithm performed significantly better, giving an overall test accuracy of **.4575%**, and a training set accuracy of **.987%** on the same 66% / 33% train-test split. Unfortunately, however, the accuracy is still too low to place any confidence in the algorithm's predictions.

5 Error Analysis

Interestingly, Naive Bayes did exceedingly well in classifying Eminiem (92% accuracy), but did horribly with in predicting several other artists like Kanye West (.09%), Jay Z (.08%), Kendrick Lamar (.08%), and Snoop Dogg (0%). Initially, we thought this may be because Naive Bayes guessed Eminem for the majority of the songs, however, on examining the data, we found that NB guessed Eminem for only 14 of the 153 songs, suggesting, potentially, that Eminem's word choices are just really, really, unique. Furthermore, we found that for the majority of Snoop Dogg's songs (12/16), NB mistook Snoop Dogg for 2Pac.

Given that our Softmax algorithm performed extremely well on training set, but poorly on the test set, we currently believe that the algorithm is susceptible to high variance and is overfitting the training set. Part of the reason for this may come from the fact that rappers occasionally use very strange or unique words like 'oooooooo' or a person's name in their lyrics, making it very easy to identify that particular song, but useless for identifying others. Thus, our plan is to prune the vocabulary we're using, eliminating words that occur too frequently (like 'the', 'and', 'I') as well as words that appear extremely rarely. Hopefully by eliminating these red herrings, we

can get a more representative picture of what defines a rapper's style.

We still need to perform more in-depth error analysis however, such as analyzing which words were given the highest weights, before we can confirm our conclusions.

6 Next Steps

6.1 Data Collection

- **Scrape more lyrics** for each of the rappers. We currently have 480 songs scraped (40 for each of the 12 rappers).
- **Separate/remove featured artists' lyrics**. Currently, the songs in the dataset include lyrics from vocalists other than the original artist, which is one of the primary sources of noise.

6.2 Preprocessing

- **Remove uninformative words** from the lyrics. In particular, we will create and look at the list of the most common words (with their frequencies), and fix a parameter n_{most} that determines the word frequency maximum cutoffs.

6.3 Feature Extraction

- **Implement existing features**. Our feature extractor is currently capable of extracting more than just the vocabulary, and implementing this will involve developing our infrastructure just a little bit. This is one of our first next steps.
- **Rhyme analysis**. One of the key characteristics in rap lyrics is rhyme schemes, and understanding the most rhymed words and other rhyming patterns may be highly contributive to increasing the algorithm's accuracy. We are planning to use either the `nltk` (which we are currently using for stemming) or the `pronouncing` library for accurate rhyme detection.
- **Proper nouns**. A common theme in rap music is to refer to other rappers and to hometowns, both of which can be used as identifying characteristics.
- **Parameter sharing**. If we have extra time or want to increase our accuracy even further (after first focusing on the other developments), we will explore using information (such as location references) from one set of lyrics to another.

6.4 RNN

Given the recent successes of RNNs in NLP, we're aiming to implement an RNN as well after improving our features.

7 Contributions

- Alex Wang: data collection, testing, naive bayes
- Robin Cheong: softmax regression, testing, naive bayes
- Vince Ranganathan: preprocessing, feature extraction, softmax regression

References

[1] Chen Qian, Tianchang He, and Rao Zhang. Deep Learning based Author Identification.

[2] Sean Stanko, Devin Lu, and Irving Hsu. Whose Book is it Anyway? Using Machine Learning to Identify the Author of Unknown Texts. 2013.

[3] Ahmed M. Mohsen, Nagwa M. El-Makky, and Nagia Ghanem. Author identificatino using Deep Learning.