Lyrical Distinction

Jesse Bartola jbartola@
Phil Michalowski pmichalowski@
Nischal Tamang ntamang@
Max Berezin mberezin@

1 Introduction

Music is an important part of our everyday lives. Technological advances in modern society have made it easier than ever for new artists to distribute their songs over the Internet. Unfortunately, the massive volume of songs that are regularly released into the public can make it difficult to discern which song is made by which artist. Moreover, a more important question arises: Are the most popular artists the ones with the most distinct music?

Aside from sound, the words an artist uses in a song are perhaps the most important determining factor in how it will appeal to the public. Lyrics contain semantic relevance, which allows for the possibility of implicit analysis of a song's cultural relevance. Some of the greatest artists of all time are renown for their unique use of lyricism. Our project will aim to determine if (a) Given a set of lyrics, we can predict with a high degree of certainty who the artist is, and (b) Given our trained prediction model, we can rank artists by their relative lyrical uniqueness and use that to predict their popularity.

2 Related work

There exists many research projects on the topic of music genre classification. Alexandros (2017) attempts to classify music genres from lyrics using Hierarchical Attention Networks (HAN). HAN outperforms N-gram, SVM, KNN and Naive Bayes by utilizing the property that words combine to make lines, lines combine to make segments, and segments combine to make the entire lyrical content. It recognizes that the semantics of lyrics are embedded in the ordering of content within these levels. HAN works by extracting these layers and learning the importance of words, lines and segments. To represent the words being fed into the HAN, Tsaptsinos uses 100 dimensional GloVe embeddings. GloVe outperforms similar word2vec and SVM based models. A benefit of this approach is we do not have to hand pick features. A negative aspect of this approach is that we do not possess familiarity with the processes, and there would be a learning curve for using both HAN and GloVe. This approach performs with an accuracy of 49.50% whereas a higher value would be ideal.

In an additional paper, Michael and Caroline (2014) try to classify genre, distinguish quality, and determine the release year of a chosen lyrical passage. Unlike Tsaptsinos, Fell & Sporleder do not use a kind of Neural Network. Instead they designed 13 total features including top 100 n-grams, type-token ratio, slang words, part of speech/chunk tags, length of lines, echoisms, rhyme features, imagery, pronouns, past tense, chorus, title and repetitive structures. As baseline, they used a simple N-gram model. For their main approach they make use of the stylistic features, which are the latter 12 features, along with SVMs. Here the thoroughly developed features capture the core aspects of a songs lyrical content. The simple methods of N-grams and SVMs had on average 52.5% accuracy, so this is a process that works quite well. However, choosing features might be a tedious task, especially if we want additional features more than ones already used previously.

3 Your approach

We will begin by assembling labeled training data mapping artist names to their respective song lyrics. To start, we'll create a simple Bayes Net model to predict the most likely artist given a string of lyrics. Afterwards, we'll be moving on to more advanced models such as RNN's, where

we will aim for better accuracy and prediction of which artist is most likely to have written a given song.

When our model is developed, we will analyze the results by projecting the parameters for each artist onto a 2-dimensional space, highlighting the degree of similarity between lyrics used by all artists in our training data. This will help us better understand whether popular artists are truly the ones with unique lyricism.

What baseline algorithms will you use? A baseline algorithm is one that is very simple and trivial to implement. For example, predict the most common class, or tag all capitalized words as names, or select the first sentence in the document. Sometimes it can be difficult to get a fancy algorithm to beat a baseline. Always ask yourself, Whats the simplest experiment I could do to (in)validate my hypothesis? Talented researchers have a knack for coming up with simple baselines.

3.1 Milestones & Schedule

1. Acquire artist+lyric training data (1 week)
2. Build simple model using Bayes Net (1 weeks)
3. Build neural network model (1.5 weeks)
4. Begin creating data visualization website (0.5 weeks)
5. Write progress report (due Nov. 16)
6. Analyze model results and plot results in visualization (2 weeks)
7. Work on final report and presentation (2 weeks)

4 Data

There are many online datasets available containing song lyrics with their respective artists.

What text data do you plan to use in your project? Where will you get it from? Will you be annotating text yourselves? Convince us that it is available for you, and that you can easily get it, and that it is appropriate for the task and research questions you care about.

5 Tools

What existing libraries or toolkits are you going to use? Some questions to think about: will you be doing any preprocessing of your data such as tokenization or parsing? Will you be training logistic regression models? Will you be using deep learning libraries? Will you need to use any services for GPUs?[1] Do you need to use crowdsourcing?

References

Alexandros, T. (2017). *Lyrics Based Music Genre Classification Using a Hierarchical Attention Network*.

Michael, F. and Caroline, S. (2014). *Lyrics-based Analysis and Classification of Music*.

[1] if so, check out https://colab.research.google.com!