Arif Ali, Harry Eldridge, John Hotchkiss, Eugene Yang
COSC-572 Empirical Methods in Natural Language Processing
November 9, 2016

Group AHJE Final Project Proposal

Github usernames
Arif Ali - arifyali
John Hotchkiss - jhotchx
Harry Eldridge - heldridge
Eugene Yang - eugene2528

Abstract

Since the dawn of the radio, and arriving fashionably late to parties, the humans of modern society have listened to partial songs and thought, “Wow, this is the jam to end all jams; I must know who and what it is.” We know a song is a lot more than just its lyrics - it’s also the beat, the instrumentation, and the artist. Musicians become famous and their producers become rich when they hit that perfect mix. We propose taking some of the mystery out of the equation by building a model that can predict the artist based on the lyrics of the song. Not only can this help the aforementioned humans finally figure out who is responsible for the song running through their heads, but it follows naturally that it could then be used by producers to assign already written songs to artists that would fit them well, and therefore be more lucrative. In addition to analyzing the lyrics directly through application of our knowledge gained in this course, we hope to explore how using other, more traditional statistical features can combine with those suggested by NLP for an even more (we hope) complete model.

Data

There are a number of potential sources of data for lyrics and other song data on the web, including Google Play Music (https://play.google.com/music/listen?u=0#/sww), AZLyrics (http://www.azlyrics.com/), Genius (http://genius.com/), and LyricWikia (http://lyrics.wikia.com/wiki/Lyrics_Wiki). Since Genius provides more than just lyrics (it is traditionally a lyrical annotation site), we believe it would be the best place to start for a healthy mix of lyrics and other song data. Although we will need to scrape the lyrics, it has an API for retrieval of other song data that could be used to create the more traditional features discussed in the abstract. Additionally, LyricWikia works as a secondary source for additional coverage, with the added benefit of having a pre-implemented python interface with the PyLyrics library.

Possible Applications

In addition to the highly motivating premise suggested in the abstract, exploring the hybrid application of classical Statistical Learning with NLP Methods could be used beyond predicting music artists to predicting the author of news articles. This could be especially useful when trying to identify otherwise anonymous authors of particular articles, such as those written and published in *The Economist*, a periodical that does not publish authors’ names. Additionally, should we expand the model to use or predict features other than authorship, such

as sentiment (which could very well be related to the author), the machine's classification of an article or song could reveal more than a normal human experience would.

Timeline

By November 15th, we should have a repository of music lyrics along with a label file that will store the artist that sang the lyrics.

By November 22th, John and Arif will have written the evaluation files that will use for each of the algorithms. At this point, hopefully basic structure of the four algorithms have been written in the form of four individual python files. These algorithms should be based on literature review.

By November 28th: We will attempt to implement a hybrid or statistical learning and NLP Empirical Methods and compare accuracy to the original four algorithms

Breakdown of Work

Eugene and Harry are the best equipped to gather the lyric and song data from Genius and other lyric sites, because in addition to getting song data from Genius's API, the website will need to be crawled in order to get the lyrics of the songs.

Everyone in the group will implement empirical models that leverage NLP knowledge and other song features in order to identify the artist. Ideally there should be a total of four models that are trained, tuned, and cross-validated before being run on a final test set.

Arif and John are better equipped to perform formal evaluation and comparison of the models and determining efficacy of results. Evaluation may involve aggregating results and creating ROC curves, confusion matrices and their standard statistics, such as accuracy, precision, and recall.