Brendan Herger
Layla Martin
Chandrashekar Konda
Adv. Machine Learning
February 17, 2015

Project Proposal
Clustering Lyrics by Genre

Project Title

Identifying Genre through Lyrics

Background and Motivation

This project would seek to continue the work that Layla and I did during Module 1 to create a database of lyrics (https://github.com/bjherger/Lyric-Analysis). We think it would be interesting to identify the features which separate the songs on Billboard's Hot Country list (http://www.billboard.com/archive/charts/2014/country-songs) from those on Billboard's Hot R&B/Hip-Hop list (http://www.billboard.com/archive/charts/2014/r-b-hip-hop-songs).

This will allow us to familiarize ourselves with the Natural Language Processing, while pursuing a project that is topically interesting and rather widely accessible.

Project Objectives

Our objective is to train a learning algorithm on a set of data labeled either Country or R&B/Hip-Hop, which can then accurately predict whether an song's lyrics are Country or R&B/Hip-Hop.

What Data?

We will modify the code from Module 1 (available at https://github.com/bjherger/Lyric-Analysis) to scrape the Billboard's Hot Country list (http://www.billboard.com/archive/charts/

2014/country-songs) and Billboard's Hot R&B/Hip-Hop list (http://www.billboard.com/archive/charts/2014/r-b-hip-hop-songs) and compile the lyrics to these songs. We have yet to determine how far back we will go, but will likely limit ourselves to lyrics within the charts over the last 2 years (to reduce the effects of trends within genre).

For each song, we hope to have the lyrics as well as a body of metadata including whether the track is explicit, whether it is instrumental and producers. We will likely limit our modeling to lyrical data, but may expand it to include this metadata if needed.

Techniques Overview

We aim to take a multi-pronged approach, including Naive Bayes, TFIDF, and feature extraction / modeling. Whereas Naive Bayes and TFIDF will require little tuning, we plan on spending a significant portion of our time creating and extracting features, and then modeling those features in various machine learning algorithm (e.g. SVM, logistic regression).

At this time, we have yet to determine whether we will essentially treat our output as binary (e.g. Country or Not Country), or will have multiple classes output by each model.

Additionally, we would like to explore combining the aforementioned models into an ensemble approach.

Must-Have Project Outcomes

At the end of modeling, we expect to have a model which can separate Country songs from R&B/Hip-Hop songs.

Optional Outcomes

We would like to expand our model to be able to identify multiple genre (e.g. Pop, Gospel, Holiday). However, this would require considerable generalization of our model.

Evaluation

We will randomly split our dataset into 90%/10% train/test subsets. Within the train test set we will use 10 fold cross validation to check model accuracy. We will only use the test set once we have a fully developed model.

Schedule, Timeline, and Team Responsibilities

Rather than provide distinct weekly timelines of each partner's work, we have elected to provide 'spheres of influence', over which each member will lead our efforts. We will all assist in all areas as needed, but each team member will take the lead for their respect sphere of influence.

Brendan: Data gathering from MusicMatch Lyrics API, data quality checks, Naive Bayes modeling, final code assembly, presentation

Layla: Data gathering from Billboard charts, feature extraction (assistance), final project write up

Chandra: feature extraction, EDA, feature extraction, NLP

Shared: Modeling