Project Proposal
Music Genre Based on Lyrics

Anna Sollazzo, Henry Hart, Jordan Jay, Juan Carlos Gallegos, And Robert Craig

January 31, 2018

***Abstract*—The primary aim for our project is to create a classifier that can accurately predict a song's genre based solely on its lyrics, with the intention of exploring the relationships between the lyrical content and different genres. In this domain, we will test our intuition about each genre's distinctive attributes and investigate more opaque or subtle patterns. However, our goal is twofold: we intend to implement both a traditional linear ML classifier that uses a derived set of NLP features, and a deep learning model, in order to compare and contrast both their accuracies and sources of error. Our models will be trained and tested on a dataset of 380,000 songs and their lyrics, acquired from Kaggle.com, and originally scraped from metrolyrics.com; the genres span: country, electronic, folk, hip-hop, indie, jazz, metal, pop, R&B, and rock. Before feeding the data to our models, will will clean it, removing songs without corresponding classifications, and ensuring lyrical notation consistency. Further processing will be done using NLP techniques and tools such as Python NLTK, to extract a set of lyrical features for analysis which may include attributes such as part-of-speech (POS) frequency, syllable count, and rhyme scheme. We hope that the results of our classifiers will illuminate the most important lyrical features with respect to genre, and provide some insight into the benefits or detriments of using deep learning for lyrical inference.**

I. Related Work

THERE exists numerous attempts at classifying songs into genres solely by their lyrical content by classification methods including neural networks and commonly used classification models such as SVM, k- nearest neighbor, etc.

One article used both recurrent neural network models and simple classifier methods to reclassify a dataset of around 500,000 song lyrics and 20 genres, similar to the dataset to be used in this study of 380,000 song lyrics and 10 genres. However, instead of just using a feature set containing syllable count, line count, word count, rhyme frequency, and rhyme scheme they also decided to use a hierarchical attention network. This relies on the model learning and using the word embedding?s, hierarchical attention, and gated recurrent units rather than hand selected features. This study resulted in a trend between the model complexity and classification accuracy, showing that as complexity increases so does the accuracy. Thus, the neural net performed much more accurately throughout their study [1].

II. Data Description and Source

THIS is very important! If you are not working with a pre-existing dataset, you must convince me that the data for your project exists and is easily accessible (e.g. can easily be scraped from a public website) Tell me what your features (attributes, columns of your data matrix X) will be.

III. Proposed project

Data, Algorithm, Evaluation... bunch of random algorithms...

IV. Estimated Timeline

Due Date	Activity
Feb. 6th	Project Proposal
Feb. 13th	Finish Data Cleaning and Mining Preparation
	Start Both Methods; SVM and Recurrent Neural Net
Feb. 23th	Progress on 'Rock or Not' (SVM) Recorded
	Progress on Recurrent Neural Net Recorded
Feb. 27th	Finish 'Rock or Not' Further the Analysis to 'Rock or Pop' etc.
	Progress on the Neural Net Method Recorded
Mar. 6th	Finish SVM and Record Results, Start Midterm Report
	All Hands On Neural Net
Mar. 11th	Prepare and Finalize Midterm Report
	All Hands On Neural Net
Mar. 13th	Midterm Report Due
	Continue Neural Net
Mar. 20th	Gather Results of Both Methods and Prepare Final Report
Mar. 25th	Meeting to Organize Presentation
	Finish Final Report
Mar. 28th	Project Presentation
Apr. 6th	Final Report Due

V. Distribution of Tasks Among Team Members

Each team member is a lead in a specific area and is required to delegate tasks in their area accordingly. This will aide in equally distributing the workload and involving team members to contribute in every aspect.

The team leaders are as follows:

Member	Task
Anna Sollazzo	TBD
Henry Hart	Data Cleaning
Jordan Jay	TBD
Juan Carlos Gallegos	Neural Net
Robert Craig	TBD

References

[1] A. Tsaptsinos, *Lyrics-Based Music Genre Classification Using a Hierarchical Attention Network*, ICME, Stanford University, USA, 2017.