Project Proposal
Music Genre Based on Lyrics

Anna Sollazzo, Henry Hart, Jordan Jay, Juan Carlos Gallegos, And Robert Craig
January 31, 2018

***Abstract*—We aim to train a classifier to predict a song's genre based on its lyrics, with the goal 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. The model will be trained and evaluated on a dataset of 380,000 song lyrics and their genres, acquired from Kaggle.com; the genres span: country, electronic, folk, hip-hop, indie, jazz, metal, pop, R/B, and rock. We will generate features describing the lyrical content with the support of existing Natural Language Processing techniques and resources, such as the Natural Language Toolkit Python library. Attributes that we are consider evaluating include: line count; syllable count; rhyme frequency; and amount of variation in rhyme scheme. Our primary goal is to produce a feature set for the song lyrics dataset along with a single classification model, on which we will iterate to increase accuracy. Our stretch goal is to develop a second model to compare to our first model, providing insight into the tradeoffs in model selection for this particular problem. The development of the second model will depend on the opportunity for insight that we perceive in our first model.**

I. Related Work

WHO has done something like what you plan to do? What did they try? How well did it work? This is an important section, so don't skimp here. Search Google Scholar to find related work.

https://arxiv.org/pdf/1707.04678.pdf

https://nlp.stanford.edu/courses/cs224n/2006/fp/sadovsky-x1n91224n_final_report.pdf

http://worldcomp-proceedings.com/proc/p2016/DMI8052.pdf

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	Data Cleaning and Mining Preparation
Feb. 23th	Algorithm Environment Setup
Feb. 27th	Run the Commonly Used Algorithms and Record Results
Mar. 6th	Deep Learning Method and Record Results
Mar. 13th	Midterm Report
Mar. 20th	Gather Results and Prepare Final Report
Mar. 28th	Project Presentation
Apr. 6th	Final Report

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	TBD
Jordan Jay	TBD
Juan Carlos Gallegos	TBD
Robert Craig	TBD

REFERENCES

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.