***Whose Rap is it Anyways?*:**
Determining Hip-hop Artists from their Lyrics

CS 229 Project Proposal

Robin Cheong, Alex Wang, Arvind "Vince" Ranganathan
{robinc20, jwang98, akranga}@stanford.edu

October 27, 2017

Project Category: NLP

Goal: To determine the author of a verse of rap lyrics. The algorithm will be trained on partial discographies of a set of $n = 5$ rappers, analyzing features such as word choice, word positioning, dictionary-invalid strings, line length, rhyme scheme, and verse structure.

Motivation: Determining authorship in works of literature is an established field with a lot of potential. It can be utilized in many interesting applications, including detecting plagiarism. While doing background research on this topic, we can across many projects that involved generating lyrics for music of all genre. Our project can be used as an adversarial tool in determining the effectiveness and accuracy of these generative algorithms.

Method: We will explore using both softmax regression and a LSTM model to capture the structure of the lyrics for five different rappers. The dataset would include lyrics from several rappers scraped using the Genius API. We intend to explore the features mentioned above, perhaps using PCA to determine which features were most helpful in determining authorship.

Intended experiments: Since we're performing supervised learning, we will evaluate our machine learning algorithm on its accuracy (no. of examples correct / total no. of examples). We also plan to compare the accuracy of the softmax regression model and the LSTM model, and if time allows, testing our algorithms on generated lyrics. We would also be interested in comparing the start, middle, and end of an artist's career and seeing if our algorithm could pick up on the differences.

Relevant Research: We've found that other work has been done on identifying the authorship of documents in data sets like the C50 dataset (news articles) and the Guternberg dataset (short stories) with accuracy up to 69.1% and 89.2%, respectively. The authors of the paper also used softmax regression and LSTM, as well as Siamese networks.

(https://web.stanford.edu/class/cs224n/reports/2760185.pdf)