CS 221: Project Proposal
Mathieu Rolfo, Shalom Rottman-Yang, Nathan Tindall
21 October 2014

Rap Lyrics Analysis

Our project proposal is to utilize rap lyrics to predict the results of music reviews. Existing work has been done on the Rap Genius (lyric aggregating site) corpus, allowing us to harvest information from tens of thousands of songs and artists. The content of lyrics can be measured on several metrics, including corpus frequency, n-grams, sentiment analysis, and higher level features, making this area of investigation an interesting and fruitful one. The lyrics can be analyzed at varying levels of scope: from song, to album, to artist as a whole; allowing us to investigate the levels of structure and similarity for music. Findings that lyrical content influences perception of music would allow for prediction of audience reception to new songs and demonstrate a link between the message of music and how it is interpreted by audiences.

The input of the model will be lyrics and song, album, and artist metadata from the Rap Genius corpus, including listener reviews and comments on songs or artists. The model will be trained on the lyrics in order to predict listener perceptions, and then be tested on different artists. Listener perceptions may be recorded r (e.g. a "score" on a 1-10 scale that aggregates multiple aspects) or a list indicating sentiment reactions (e.g. "raunchy"). This data is easily accessible through online review aggregators, if not through the Rap Genius corpus.

The evaluation metric for success is the accuracy that the prediction algorithm can achieve when compared to human reviews of the song or album.

Baseline: A baseline implementation is a linear regression model, which predicts the relationship between word counts in a song and reviewer score. This problem is simple enough to be easily implemented.

Oracle: The oracle can be implemented by having people read song lyrics and then predict the way that the average reviewer would. This is a fairly difficult problem, as people have widely varying interpretations of music. However, the user has an intuition as to the quality of a song, based upon its lyrics. Although review sentiment does not map directly to lyrics, the relationship here will be explored.

A number of challenges exist as we move forward with the project. On a practical level, issues such as poor lyric transcription, unsyntactic speech structure, and transcriber preferences in punctuation may interfere with successful feature identification and lead to less effective sentiment analysis. Additionally, the feature space is quite wide. Thus, much work needs to be done in order to identify what aspects of lyrics are good indicators to the way they are perceived.

K-means clustering and classification may prove useful here, in order to both cluster artists and features into groups, and also to separate them into positive and negative attributes. This lead to interesting side projects, such as determining underlying similarities between rappers separate from group affiliation (i.e. Wu-Tang Clan).

Past work includes those of CS229 students, who have worked with attributes of movies in order to predict reviews. Additionally, the Rap Genius dataset has been worked on in the past, with data scientists showing the relative lexicons of several rappers. By adding a predictive element to this dataset, we hope to contribute to this analysis.

(https://github.com/timrogers/rapgenius http://rappers.mdaniels.com.s3-website-us-east1.amazonaws.com/)