Project Proposal

TA mentor: Pokey
Name 1: Anand Sampat (asampat)
Name 2: Megan Schoendorf (megan2)

Response to Feedback: We got feedback on an idea about trying to classify papers based on their difficulty and their subject matters. Although this would be a very col project, we realized that the scope is a bit too large and would require extensive computational power given the amount of text we would have to parse. Since we are both interested in music, we decided to, as a first pass take this text parsing and NLP thinking to music. In particular, we want to use the techniques we learned in this class to classify songs by genre.

Summary: Classify music by genre, artist, and time period using AI techniques (NLP, machine learning, logic) to analyze the lyrics and the waveforms of a variety of songs.

Input Data: We decided to start with lyrics data that we could easily scrape from the web – we found that the data on http://metrolyrics.com was the easiest to parse and so began with that. In particular, we first scraped lyrics for songs in the top 100 and then expanded to get lyrics from the artists represented in the top 100 dataset. This way we could slowly see how the error rate decreases as we add more samples.

Analysis: Using code from the spam classifier homework, we adapted the situation to reflect a text file that has words whose costs are calculated based on unigram and/or bigram cost functions. Specifically, in analyze_lyrics.py the code has been altered to reflect two labels, genre and artist. Then we learn from the OneVsAllClassifiers and instantiate a OneVsAllClassifier to be a genre classifier and one to be the artist classifier.

Output:

Just top 100 + some more (100 songs):
100% Artist Error Rate – only had a few examples for each artist.

Added songs from Justin Timberlake and Miley Cyrus (192 songs) :
TRAINING EXAMPLES:
Genre Error Rate = 0.2624113475177305
Artist Error Rate = 0.6737588652482269
DEV EXAMPLES
Genre Error Rate = 0.35766423357664234
Artist Error Rate = 0.948905109489051

Added songs from Coldplay, Lana Del Rey, Bruno Mars, Jay-Z (566 songs):
<Convergence taking a long time…>
TRAINING EXAMPLES:
Genre Error Rate = 0.384375
Artist Error Rate = 0.7375

DEV EXAMPLES
Genre Error Rate = 0.5644171779141104
Artist Error Rate = 0.9171779141104295

Interestingly, as the artist error rate goes down (as we would expect with more data from each of the artists) the genre error rate went up both on the training and dev examples which means that the new examples only added more uncertainty to the algorithm.

Plan going forward:

Continue scraping data until we have thousands of songs and then run the basic one vs. all classifier on the data – this may take a while due to computational limitations so we may have to implement this in a faster language. Otherwise, we will experiment with number of examples and the accuracy of the algorithm to classify new songs. Likewise we can rewrite the code in a quicker language for convenience.

Our next step will be to implement a more complex algorithm on the lyrics themselves. Although we're not doing sentiment analysis, we can use methods used in Kumar *et. al.* such as Naïve Bayesian (NB), k-Nearest Neighbor (KNN) and Support Vector Machine (SVM) [1]. We could even try to add a some knowledge-based such as semantic analysis as per Cambria *et. al.*

References:
[1] http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6466307
[2] http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6383145