Currently, a hot topic in data mining is figuring out how to retrieve information about music by analyzing the music itself. While the music itself has interesting properties, it seems that the lyrics themselves are somewhat ignored, as searches for studies involving analysis of the lyrics that accompany the music will not net as many relevant results, aside from a paper such as:


which was looking to see if text mining would reveal any similarities that separate top Chinese songs from less popular music. That study focused more on word meanings rather than other properties of words.

The goal would be to create a page where users can enter in a set of lyrics along with the title and artist, and then the lyrics would be processed by a text parser written in PHP. Using a machine learning/data mining algorithm (most likely k-means), the program will then attempt to guess what genre/decade the lyrics come from based solely on the lyrics (without cheating and just saving information for each artist and check to see if the artist already exists in the database). Each time a user enters in lyrics, this information will be stored as a training example in a database. As a result, if there are unique properties that separate genres and different decades of music, the program should get better at guessing the relevant information about the music.

I would like to create a text parser that will gather the following information from any set of lyrics:

- **Word Count**
  - Average word count for verse/chorus/hook
  - Number of different words in chorus/hook
- **Word Frequency**
  - Also within the verse/chorus/hook
- **Average number of syllables per word**
  - Based on some predefined rules I will set (not necessarily 100% accurate)
- Additional information will probably be gathered as data is analyzed

The results will also be displayed in a graphical format using Flash. Two programs that will be used come from the following site: http://www.maani.us/

The relevant ones will be the XML/SWF Gauge and the PHP/SWF Charts. Flash, while still a bit slow and clunky in terms of programming use, has developed into a powerful platform for displaying data. An example of its potential can be found at: http://www.gapminder.org/ Unfortunately, Google has bought this company out and has not released any sort of code or trial version of the Trendalyzer. The Gauge and Charts program are free for personal use and as a result, I will be using both to display data in an
interesting format. The charts will be used to display the various statistics found earlier, while the gauge will be a sort of representation of clustering. Both the charts and gauge have an update function which allows anything displayed in the gauge to be modified almost real-time without reloading the page, so this can be used to see how more training examples affect the graphical display of data.

Deliverables:
- Website for entering lyrics and displaying data
- PHP code
  - With an implementation of a machine learning algorithm (most likely using k-means)
  - Code to retrieve and store information in the database
  - Code to produce the relevant XML files to be used with the Flash
- mySQL Database
- Flash display of database data, in different formats, including one in the vein of k-means clustering
- A conclusion as to whether or not music can truly be classified by the aspects of lyrics noted earlier

The gauge will work well with displaying clustering because circles and polygons can be drawn easily along with lines so that it can be visualized exactly where a specific song may lie in relation to each other. The charts have the ability to serve as a more formal display of data, though much of the same functionality is available.