Project II - CSC 391/691
Spring 2014

This is a two part project with two deadlines. The work you submit should be your own work.

Both parts of the project deal with the *musicXmatch* dataset (http://labrosa.ee.columbia.edu/millionsong/musixmatch#desc), a collection of lyrics associated with the Million Song Dataset. Specifically, the data set we will work with contains pre-processed information regarding the lyrics of 210,519 popular songs. The lyrics are in a "bag-of-words" format: each lyric is described as the word counts for a dictionary of the top 5,000 words across the collection. The format is:

```
# - comment, ignore
%word1,word2,... - list of top words, in popularity order
TID,MXMID,idx:cnt,idx:cnt,... - track ID from MSD, track ID from
                                musiXmatch, then word index : word
                                count (word index starts at 1!)
```

If you look at the word list you will find that the words have already been "stemmed" in a standard way and that some "words" may represent phonetic versions of several related words.

The data file we will work with can be found on Sakai under Resources → Data → mxm_dataset_train.txt.zip. You can also download from the musiXmatch web site but be sure to get the correct file. On the musicXmatch web site you will also find a SQLite version of the data if you prefer to work with a database rather than a text file. Note that the SQLite data contains both the training and testing data set.

Part 1. Warmup: Due 3/25 in class – Just e-mail me your answers.
The purpose of this part of the project is to get you started working with the data. It's not difficult. Use whatever tools you want to use to answer the following questions. The approach most likely to be successful is to write a program to do the computation but if you just want to use a calculator, good luck.

a. What is the highest TF-IDF value for these lyrics and words? In other words, for all the words, in all the lyrics, what is the highest TF-IDF value?
b. Of all 5,000 words, which word is associated with the highest TF-IDF?
c. Of all the lyrics, which song is associated with the highest TF-IDF value? (Report either the MSD ID or musiXmatch ID.)

Part 2. Similarity: Due 4/1 in class – E-mail me your answer, along with an explanation of your approach, and submit your program code through Sakai → Assignments.

Of all 210,519 lyrics, how many pairs have at least a 95% Jaccard similarity? The only constraint on your implementation is that you use some form of LSH.