CSC 391/691 Project 4 Overview
Due 5/1/14

This project will permit you to experiment with the identification of frequent item sets in a large collection of data, the million song dataset (only 210,000+ lyrics, actually).

Download from Sakai → Data the song lyric dataset used for studying similarity (mxm_dataset_train.txt). You may still have it on your computer.

For the following experiments, do not include any word ranked 1-100 in the list of words on line 17 of the file. These are the 100 most frequent words and many of them are "stop" words. We'll omit them from our analysis.

Write your own program, in a language of your choice, to answer the following questions.

a. How many words appear in at least 1/20 (5%) of all the lyrics?
b. How many pairs of words appear in at least 1/20 (5%) of all the lyrics?
c. How many word triples (groups of 3 words) appear in at least 1/20 (5%) of all the lyrics AND what are they? [I don't plan to look at your list, just show 10-20 triples to convince me that you could have listed them all if you wanted to.]

Deliverables

Your answers from parts a-c along with a narrative description of how you arrived at those answers and any interesting observations during your exploration of the data.