**CS685G:  Special Topics in Data Mining (Spring 2017)**

**Homework 1: Due Feb 13th**

**Goal:**  This homework would reinforce the understanding of the computational complexity of frequent itemset mining, the difference between apriori-based and depth-first frequent itemset mining algorithms and the output and interpretation of frequent patterns and association rules.  It also gives you hands on experience to apply frequent itemset mining software to solve real-world problems

**Problem 1 (60 points)**

**Software**

There are many implementations of frequent itemset mining algorithms available on the web. In this assignment, you need to choose one or a couple of appropriate frequent itemset mining softwares (or implement your own) to solve the following problem. Here are the places where existing implementations are available:

(1). <http://www.borgelt.net/fpm.html>.

(2). <http://www.philippe-fournier-viger.com/spmf/>

Problem:

You will be given a presence/absence bitmap matrix. This is a dataset containing a large matrix with 148 columns. Four small datasets are sampled from this large dataset with 10k, 100k, 1M, and 5M rows respectively. The entry of the matrix corresponds to 1/0 (presence/absence information) of the corresponding row and column. In the nutshell, the problem is to identify all unique all 1s submatrices with the number of rows larger than a given threshold (Try 1%, 0.5%, 0.1%, 0.05%, 0.01%, or more if still possible especially for larger dataset).

For the project submission, please report the following:

1. Please identify the most appropriate itemset pattern mining algorithm to address this problem.
2. For the software you choose to use, please provide its detailed information. If you decide to write your own, please include the phesudo code and detailed rationale. The performance will be compared.
3. For each of the dataset, please report the use of peak memory, running time, and the number of patterns.
4. Please summarize the patterns. The metrics of interest (You may suggest more) are:
	1. The distribution of the size of the submatrics in terms of #rows, columns and total size of the submatrices.
	2. The distribution of the overlap between submatrics.
5. Extra credit: Please suggest ways to condense the patterns
6. Extra credit: Please discuss how the algorithm will behave as more columns are added to the datasets.