**CS485G:  Special Topics in Data Mining (Spring 2016)**

**Homework 1: Due Feb 18th**

**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 download the following two implementations of itemset mining algorithms from <http://www.borgelt.net/fpm.html>.

I. Apriori.

II. FPgrowth.

**Datasets:**

1)  Using [Synthetic data generator](file:///\\escalade\liuj\public_html\teaching\CS685_f10\Assn\IBM_Synthetic_Data_Generator.rar) for transaction database with embedded frequent itemsets to generate the two datasets:

      T25*.*I20*.*D50k

      T10.I4.D50k

Here

**T** corresponds to -tlen  avg\_items\_per\_transaction

**I**corresponds to -nitems number\_of\_different\_items\_in\_000s

**D** corresponds to -ntrans number\_of\_transactions\_in\_000s

2) Mushroom datasets: <http://archive.ics.uci.edu/ml/datasets/Mushroom>

**Reports of experimentation:**

0)    Please describe the configuration (CPU and MEM) of the computer where you run the experiments.

Make sure to run the whole experiment under the same configuration.

1)    For the two synthetic datasets, please report the following results for both algorithms in a plot

(1)  Running Time by varying support threshold as 0.5% ,  2%,  5%, 10%, 50%

(2)  Number of patterns by varying support threshold as 0.5% ,  2%,  5%, 10%, 50%

2)    For synthetic datasets T25*.*I20*.*D50k, please report the following results for both algorithms

Please put each set of the result into a plot.

(1) Given support threshold as 0.3%, report running time by using 20%, 40%, 60%, 80%, 100% transactions.

(2) Given support threshold as 0.3%, report number of patterns by using 20%, 40%, 60%, 80%, 100% transactions.

3)    For each figure you generated above, explain the trends shown in each figure based on your understanding of the complexity of the algorithms.

4)   Apply both algorithms on the mushroom dataset. Report the following.

4.a Report the top 10 patterns with highest support

4.b Report the top 10 maximal patterns with highest support

4.c Report the top 10 closed patterns with highest support

4.d Discuss and compare the usefulness of patterns obtained from 4.a, 4.b and 4.c.

4.e Report 5 association rules with confidence above 20% and explain whether they are important.

You may use FIMGUI or ARULEGUI from <http://www.borgelt.net/fpm.html> to explore and interpret the datasets

Note: Grading of problem 1 will be based on the correctness of the plots and their interpretation

**Problem 2 (40 points)**

Implement an apriori-like algorithm to generate the frequent pairs of the items using either python or java. The algorithm should be able to take the mushroom dataset used in problem 1 and a support threshold as input. It will output the frequent pairs together with their support count in the following form:

*Item 1, item 2, support;*

Note: The code will be graded by its correctness as well as the amount of memory used.

**Submission Instruction**

Problem 1: Please submit a pdf document to canvas containing your plots as well as interpretation of the results.

Problem 2: Please submit the code to canvas.