CSCI 720 Big Data Analytics

Assignment 4

In this assignment, you will do frequent item set mining on the adult income dataset. Open source implementations of *Apriori* and *FPGrowth* algorithms in python are provided along with a simple test example. You many use these or any other packages you have access to or write your own.

Data

You will use the adult income dataset training data only (*adult.data*). Description of the columns is given at the bottom on the file *adult.name*. Remove the columns for *education-num* and *fnlwgt*. Convert continuous features *age, capital gain/loss, hours-per-week* to categories (using histograms or other method). Each row of features will be considered a transaction for the purpose of frequent itemset mining

Results

1. Describe your preprocessing. What are the total number of items in the transaction data?
2. Plot run times of *Apriori* and *FP Growth* versus Minimum Support. Use minSup of 2%, 5% 10%, 15% 20% , 50%. Explain the differences in the run times of the two algorithms.
3. Plot a histogram of the length of frequent itemsets obtained for minSup=10%.
4. For minSup=10%, choose 5 frequent itemsets that contains "<=50K" (choose itemsets with more than 3 items). Consider the rule X -> "<50 K", where X are the other items in your frequent itemset. Find the confidence of your rule for each of the frequent itemsets you chose.
5. Repeat (4) for 5 frequent itemsets that contains ">50K".

Include your results (include plots to support your assertions) as well as any code you developed with your submission.