**CSC4008 2019 Term 2**

**Assignment #5**

**Deadline:** 2020/4/19, 23:59

(This is a strict deadline. Submissions are not acceptable after the deadline)

**Percentage:** 10%

**Purpose:** Learn to mine frequent patterns, associations, and correlations efficiently; implement Apriori and FP-growth algorithm, and compare their performance

**Requirements:**

(1) Given a transactions/items dataset in a supermarket, use a programming language that you are familiar with, such as Python, C++, and R language, to implement Apriori algorithm for mining frequent itemsets. Use it to find at least 10 most frequent 2-itemset in the supermarket dataset. Also explore at least 10 most frequent 3-itemset and 4-itemset. (60%)

(2) Compare your results of using Apriori with the results you get using Weka. Write a report to analyze the situations (e.g., different k-itemset, data size, minimal support threshold setting, etc.) (20%)

(3) Use a programming language to implement FP-growth algorithm. Compare the performance of FP-growth algorithm with Apriori algorithm. Use it to find at least 10 most frequent 2-itemset in the supermarket dataset. Also explore at least 10 most frequent 3-itemset and 4-itemset. (40%)

**Notice:**

1.Choose either (2) or (3). Selection of (1) and (2) can get a maximum point of 80, whereas selection of (1) and (3) can get a maximum of 100.

2. For supermarket.xlsx dataset, the attribute names are listed in supermarket\_attribute\_name file. “?” and “1” mean the item is not purchased and purchased, respectively. The last attribute has two values – high and low, which indicates whether the transaction is less than $100 (low) or greater than $100 (high). You need to find items that are usually bought together by customers.

**Assessment criteria:**

1. Implementation of Apriori (45%); use Apriori to find top 10 most frequent 2-itemset (5%), top 10 most frequent 3-itemset (5%) and top 10 most frequent 4-itemset (5%)
2. Compare Apriori with Weka (20%)
3. Implementation of FP-growth (30%); use FP-growth to find top 10 most frequent 2-itemset, top 10 most frequent 3-itemset and top 10 most frequent 4-itemset (10%)

**Submission:** You need to submit a **studentID.zip** (replace your ID with studentID) which contains your codes and report.