CSCI 5523 **Homework 4** **Due: March 25th, 2019**

Please clearly mention your ***name*****,** ***email*****,** ***Student ID*** and ***Section***(1 for morning, 2 for evening or UNITE) on your submission.

Question 1

Consider the following frequent 3-itemsets:

{a, b, c}, {b, c, d}, {a, b, d}, {a, c, d}, {a, d, e}, {a, c, e}

The book presents two algorithms for generating candidate 4-itemsets, the Fk−1 x F1 method (Page 369) and Fk−1 x Fk−1 method (Page 371).

a. List all the 4-itemsets that will be generated by Fk−1 x F1 candidate generation method and the 4-itemsets that will be selected after the pruning step of the Apriori algorithm.

b. List all the 4-itemsets that will be generated by Fk−1 x Fk−1 candidate generation method and the 4-itemsets that will be selected after the pruning step of the Apriori algorithm.

c. Consider a new candidate generation method in which a pair of frequent (k-1)-itemsets are merged only if the last (k-2) items of the first one are identical to the first (k-2) items of the second one. For example, {x,y,z} and {y,z,w} can be merged to generated {x,y,z,w}.

List all the 4-itemsets that will be generated by this method and the 4-itemsets that will be selected after the pruning step of the Apriori algorithm.

d. How do the two sets of results (i.e., the candidates generated from candidate generation step and the itemsets after the pruning step) you obtained from part (b) related to the results you obtained from part (c).

Question 2

For each of the following statements, state if it is True or False, and provide a brief explanation.

a. Given that {a,b,c,d} is a frequent itemset, {a,b} is always a frequent itemset.

b. Given that {a,b}, {b,c} and {a,c} are frequent itemsets, {a,b,c} is always frequent.

c. Given that the support of {a,b} is 20 and the support of {b,c} is 30, the support of {b} is larger than 20 but smaller than 30.

d. Given that the support of {b,c} is 30 and {b} is a maximal frequent itemset, the count of minsup is equal to 30.

e. Given that the support of {a,b,c} is 20 and the support of {b,c} is 30, {b,c} is a closed itemset.

Question 3

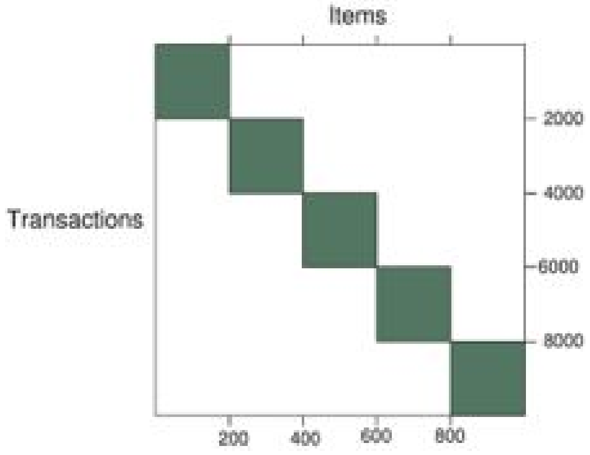
Consider the market basket transactions shown in the table below:

TID	Items
1	Bread, Butter
2	Bread, Milk, Beer
3	Milk, Apple, Beer, Coke
4	Bread, Apple, Beer, Butter
5	Bread, Milk, Apple, Coke

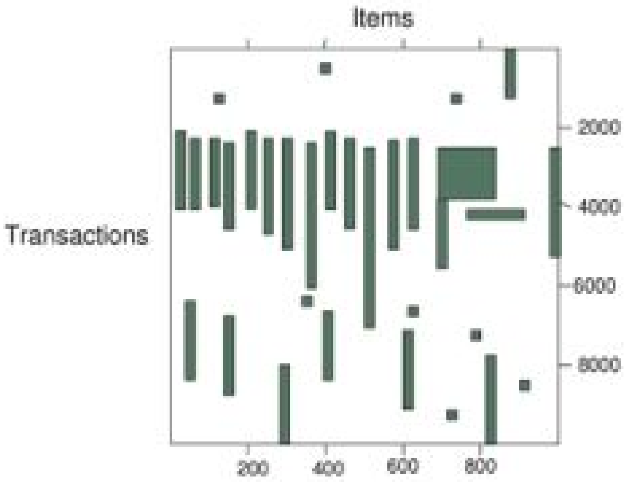
a) What is the maximum number of association rules that can be extracted from this data (including rules that have zero support)?
b) What is the maximum size of frequent itemsets that can be extracted (assuming minsup > 0)?
c) Write an expression for the maximum number of size-3 itemsets that can be derived from this data set.
d) What is the support of {Bread},{Milk},{Bread, Milk}?
e) What is the confidence of the rule {Bread} -> {Milk} and {Milk}->{Bread}?
f) Comment on the situation where the rules {a} –> {b} and {b} –> {a} have the same confidence and the situation where the rules {a} –> {b} and {b} –> {a} have the different confidence.

Question 4

Answer the following questions based on the data sets shown in the figure below. Note that each data set contains 1000 items and 10000 transactions. Dark cells indicate the presence of items and white cells indicate the absence of items. We will apply the Apriori algorithm to extract frequent itemsets with minsup = 10% (i.e, itemsets must contain at least 1000 transactions). For ease of comparison, you can assume that the vertical bars in data set B are approximately of width 10 items each and the square block centered around item 800 is of the same size as each of the 5 square blocks in data set A.



Data set A



Data set B

1. Which data set will produce the most number of frequent item sets?

2. Assume that the minimum support threshold is equal to 10%. How many closed frequent itemsets will be discovered from data set 1?

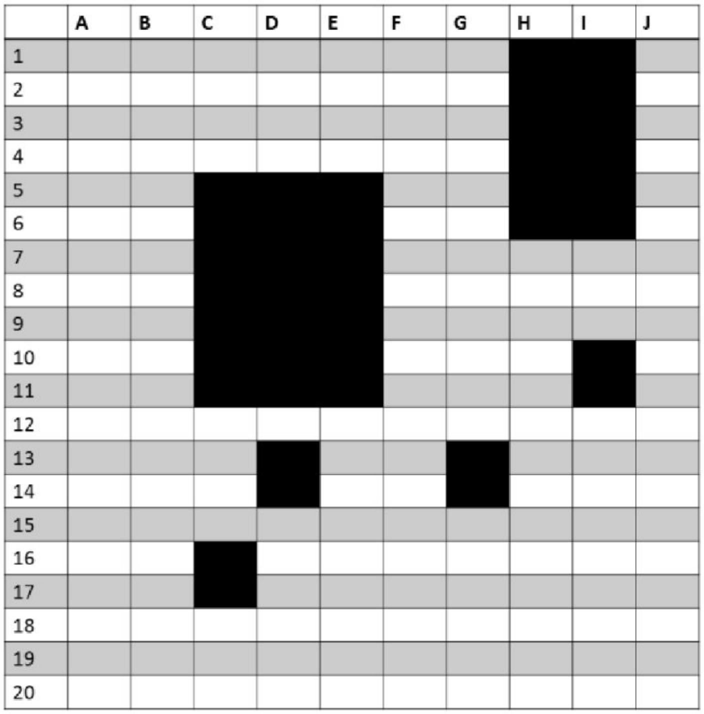
3. Which data set will produce the longest frequent itemset?

4. Which data set will produce frequent itemset with highest support?

5. Which data set will produce the most number of closed frequent itemsets?

Question 5

The figure below depicts a transaction matrix with 10 items and 20 transactions. Dark cells indicate the presence of items and white (or grey) cells indicate the absence of items. We apply the Apriori algorithm to extract frequent itemsets with minsup=20% (i.e. itemsets must be contained in at least 4 transactions). Answer the following questions:



	A	B	C	D	E	F	G	H	I	J
1								■	■	
2								■	■	
3								■	■	
4								■	■	
5			■	■	■			■	■	
6			■	■	■			■	■	
7			■	■	■					
8			■	■	■					
9			■	■	■					
10			■	■	■				■	
11			■	■	■				■	
12										
13				■			■			
14				■			■			
15										
16			■							
17			■							
18										
19										
20										

a) List all the **maximal itemsets** in the dataset.
b) List all the frequent itemsets in the dataset.
c) List all the closed frequent itemsets in the dataset.

Practice Problems

Question 1

Answer the following questions for the data set in the given table.

Transaction ID	Items bought
1	{A,B,D,E}
2	{B,C,D}
3	{A,B,D,E}
4	{A,C,D,E}
5	{B,C,D,E}
6	{B,D,E}
7	{C,D}
8	{A,B,C}
9	{A,D,E}
10	{B,D}

A. What is the maximum number of association rules that can be extracted from this dataset (including rules that have zero support)?
   What is the maximum size of frequent itemsets that can be extracted (assuming minsup >0)?
B. Calculate the maximum number of size-3 itemsets that can be derived from this dataset.
C. Find an itemset (of size 2 or larger) that has the largest support.
D. Find a pair of items, say x and y, such that the rules {x} → {y} and {y} → {x} have the same confidence.

We ***strongly*** encourage you to do Questions:

12, 13, 14, 17, 20

from the textbook "Introduction to Data Mining, Second Edition, P.N. Tan, M. Steinbach, A. Karpatne, V. Kumar.", Chapter 5. It's also freely available below:
https://www-users.cs.umn.edu/~kumar001/dmbook/ch5_association_analysis.pdf