Your name:

B565 HW4 (Fall 2022)

Submission instructions

Submit a PDF file to canvas. The PDF file includes your answers to all the problems, including results from your implementation of the Apriori algorithm. In addition, submit your code/notebook (for problem 5) to github.iu under HW4 folder in your B565 repository.

Questions

1. (15 pts) Consider the market basket transactions shown in Table 1 below,

Table 1: Market basket transactions.

Transaction ID	Items
1	Milk, Beer, Diapers
2	Bread, Butter, Milk
3	Milk, Diapers, Cookies
4	Bread, Butter, Cookies
5	Beer, Cookies, Diapers
6	Milk, Diapers, Bread, Butter
7	Bread, Butter, Diapers
8	Beer, Diapers
9	Milk, Diapers, Bread, Butter
10	Beer, Cookies

(a) What is the maximum number of association rules that can be extracted from this data (including rules that have zero support, but not rules involving empty itemsets)?

(b) What is the maximum size of frequent items that can be extracted (assuming $minsup > 0$)?

(c) Write an expression for the maximum number of 3-itemsets that can be derived from this data set.

(d) Find an itemset (of size 2 or larger) that has the largest support.

(e) Find a pair of items, a and b, such that rules $\{a\} \rightarrow \{b\}$ and $\{b\} \rightarrow \{a\}$ have the same confidence.

2. (15 pts) The Apriori algorithm uses hash tree data structure to store and count support of candidate itemsets. Consider the hash tree for candidate 3-itemsets shown in the Figure 1,

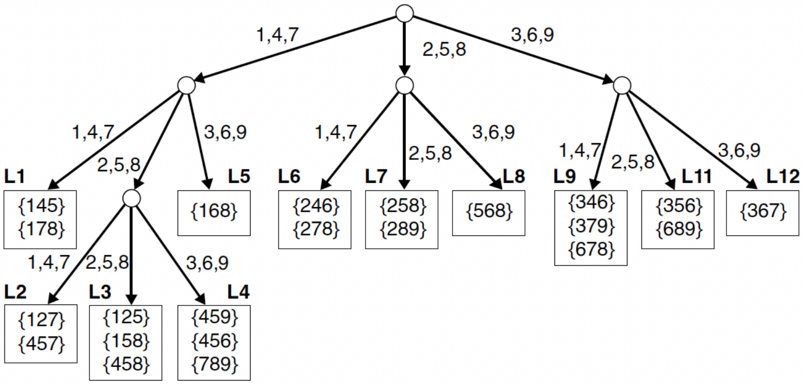
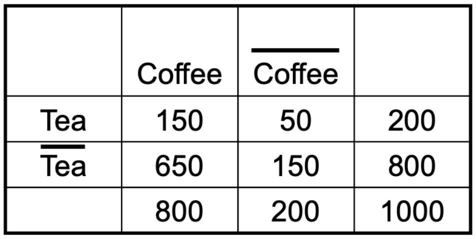


Figure 1: A hash tree of 3-itemsets.



	Coffee	$\overline{\text{Coffee}}$	
Tea	150	50	200
$\overline{\text{Tea}}$	650	150	800
	800	200	1000

Figure 2: A hash tree of 3-itemsets.

(a) Given a transaction $\{1, 3, 6, 8, 9\}$, which of the hash tree leaf nodes ($L1$ to $L12$) will be visited when finding the candidates of the transaction?

(b) Use the visited leaf nodes in part (a) to determine the candidate itemsets that are supported by the given transaction.

3. (10 pts) Given a contingency table for Tea and Coffee shown in Figure 2, compute the chi-squared statistic. Show the steps (including the expected counts assuming these two variables are independent to each other). You are welcome to use online tools to check if your answer is correct.

4. (10 pts) Use an example to explain Simpson's Paradox. If you get your idea from somewhere, please make sure that you include proper citations. Use your own words to describe. Don't copy and paste.

5. (50 pts) Apriori algorithm (implementation)

   Implement the Apriori algorithm by first determining frequent itemsets and then proceeding to identify association rules.

   (a) Implement both $F_{k-1} \times F_1$ and $F_{k-1} \times F_{k-1}$ methods. Allow in your code to track the number of generated candidate itemsets as well as the total number of frequent itemsets. There are tons of implementations of the Apriori algorithm on the web. Make sure that you implement your own for this assignment. You don't need to implement the hash tree for counting the support. Consider using lists or dictionaries for this problem.

(b) Use this transaction dataset on Kaggle to test your program. Compare the two candidate generation methods on the given dataset for three different meaningful levels of the minimum support threshold and minimum confidence (the thresholds should allow you to properly compare different methods and make useful conclusions). Provide the numbers of candidate itemsets considered in a table and discuss the observed savings that one of these methods achieves.