**CS 422**

**Due:**

**04/22/2025**

**Data Mining**

**Homework 3.0**

**Practicum Problems**

These problems will primarily reference the lecture materials, class

examples, the prescribed textbook, or the instructor manual (all available

on Canvas). For this assignment, you are required to type your responses

and submit them as a single PDF document via Canvas. Students are

encouraged to refer to the textbook or credible online resources to answer

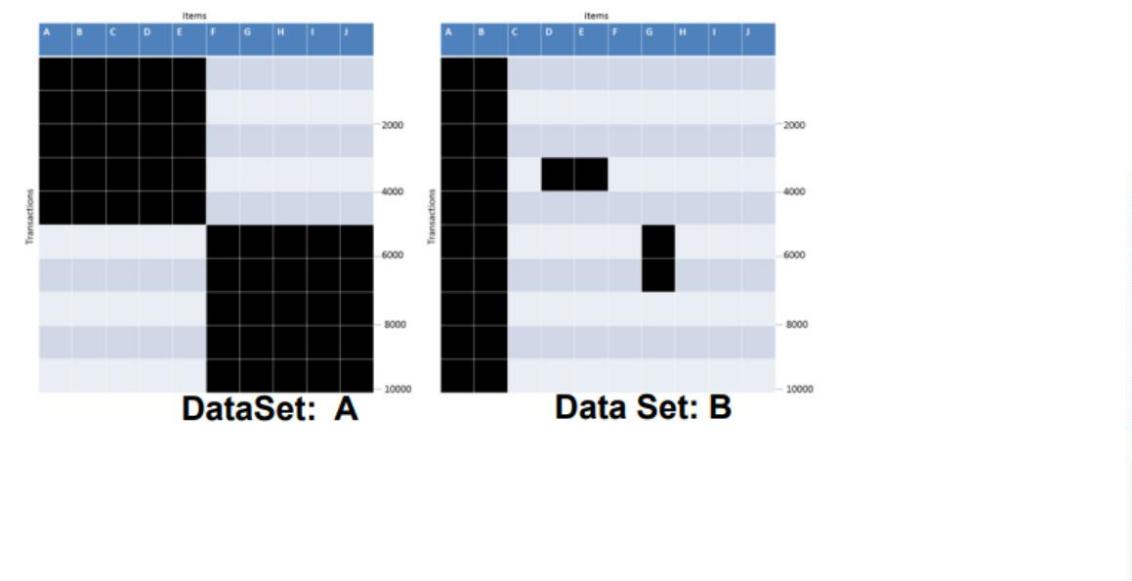
the questions accurately.

**~~Problem 1~~**

Given the following transaction data sets (dark cells indicate presence

of an item in a transaction) and a support threshold of 20%, answer the

following questions:



a. What is the number of frequent itemsets for each dataset?

Which dataset will produce the most number of frequent

itemsets?

b. Which dataset will produce the longest frequent itemset?

c. Which dataset will produce frequent itemsets with highest

maximum support?

d. Which dataset will produce frequent itemsets containing items

with widely varying support levels (i.e., itemsets containing

items with mixed support, ranging from 20% to more than

7

0%)?

e. What is the number of maximal frequent itemsets for each

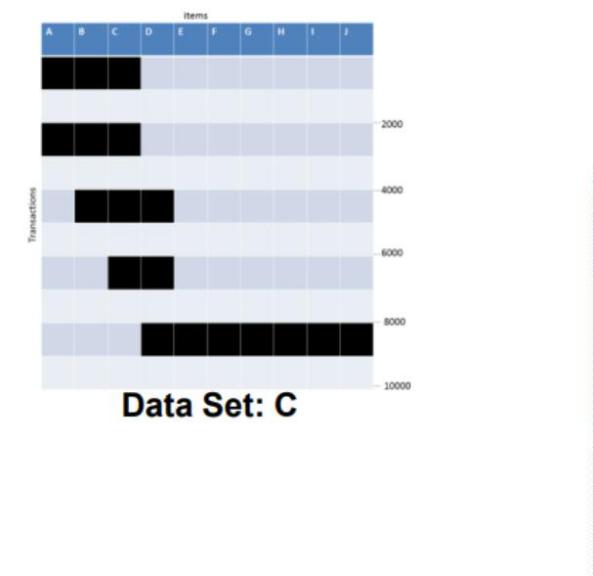
dataset? Which dataset will produce the most number of

maximal frequent itemsets?

f. What is the number of closed frequent itemsets for each

dataset? Which dataset will produce the most number of closed

frequent itemsets?



Ansewr：

**a**.The total number of transactions is N=10000. Due to the support threshold of 20%, the support count threshold is 10000 \* 20%=200

Data Set A:The number of frequent itemsets in a single item set is 10, and the number of frequent itemsets in a binomial set is 2\*C52=20,the number of frequent itemsets in the three itemsets is 2\*C35=20,the number of frequent itemsets in the four itemsets is 2\*C45=10,the number of frequent itemsets in the five itemsets is 2;So the total number of frequent itemsets =10+20+20+10+2=68

Data Set B:The number of frequent itemsets in a single item set is 3,and the number of frequent itemsets in a binomial set is 3,the number of frequent itemsets in the three itemsets is 1;So the total number of frequent itemsets=3+3+1=7

Data Set C: The number of frequent itemsets in a single item set is 4,and the number of frequent itemsets in a binomial set is 4,the number of frequent itemsets in the three itemsets is 1;So the total number of frequent itemsets=4+4+1=9

So DataSetA generates the highest number of frequent datasets.

**b**.The longest frequent itemset of DataSetA is 5;The longest frequent itemset of DataSetB is 4;The longest frequent itemset of DataSetC is 7;So DataSetC generates the longest frequent itemset dataset.

**c**.

DataSet A

|  |  |
| --- | --- |
| item | Frequent itemset support |
| A | 50% |
| B | 50% |
| C | 50% |
| D | 50% |
| E | 50% |
| F | 50% |
| G | 50% |
| H | 50% |
| I | 50% |
| J | 50% |

The maximum support for frequent itemsets in dataset A is 50%.

DataSet B

|  |  |
| --- | --- |
| item | Frequent itemset support |
| A | 100% |
| B | 100% |
| C | 0% |
| D | 10% |
| E | 10% |
| F | 0% |
| G | 20% |
| H | 0% |
| I | 0% |
| J | 0% |

The maximum support for frequent itemsets in dataset B is 100%.

DataSet C

|  |  |
| --- | --- |
| item | Frequent itemset support |
| A | 20% |
| B | 30% |
| C | 40% |
| D | 30% |
| E | 10% |
| F | 10% |
| G | 10% |
| H | 10% |
| I | 10% |
| J | 10% |

The maximum support for frequent itemsets in dataset B is 40%.

So Dataset B is the dataset that generates frequent itemsets with the highest maximum support

**d**.

The frequent itemset of the item with the highest support difference in Dataset A is 0;

The frequent itemset of the item with the highest support difference in Dataset B is 100%;

The frequent itemset of the item with the highest support difference in Dataset A is 40%;

So DataSetB is a dataset that generates frequent itemsets containing items with significant differences in support

**e**.

The highly frequent items in Dataset A are:{A,B,C,D,E},{F,G,H,I,J}

The highly frequent items in Dataset B are:{A,B,G}

The highly frequent items in Dataset C are:{A,B,C},{C,D}

Dataset A and Dataset C are datasets with the highest number of frequent itemsets and the most frequent itemsets.

**f**.

The closed frequent itemsets of Dataset A are:{A,B,C,D,E},{F,G,H,I,J}

The closed frequent itemsets of Dataset B are:{A,B},{A,B,G}

The closed frequent itemsets of Dataset C are:{A,B,C}.{C,D}

The number of closed frequent itemsets for the three datasets is the same, all of which are 3.

**~~Problem 2:~~**

Consider the following set of candidate 3-itemsets:

{

1, 2, 4}, {1, 3, 5}, {1, 4, 6}, {2, 3, 5}, {2, 5, 6}, {3, 4, 5}, {3, 5, 6}, {2, 4, 6}

a. Construct a hash tree for the above candidate 3-itemsets. Assume the tree uses a

hash function where all odd-numbered items are hashed to the left child of a

node, while the even-numbered items are hashed to the right child. A candidate

k-itemset is inserted into the tree by hashing on each successive item in the

candidate and then following the appropriate branch of the tree according to the

hash value. Once a leaf node is reached, the candidate is inserted based on one of

the following conditions:

i. Condition 1: If the depth of the leaf node is equal to k (the root is

assumed to be at depth 0), then the candidate is inserted regardless of

the number of itemsets already stored at the node.

ii. Condition 2: If the depth of the leaf node is less than k, then the

candidate can be inserted as long as the number of itemsets stored at

the node is less than maxsize. Assume maxsize=2 for this question.

iii. Condition 3: If the depth of the leaf node is less than k and the

number of itemsets stored at the node is equal to maxsize, then the

leaf node is converted into an internal node. New leaf nodes are

created as children of the old leaf node. Candidate itemsets

previously stored in the old leaf node are distributed to the children

based on their hash values. The new candidate is also hashed to its

appropriate leaf node.

b. How many leaf nodes are there in the candidate hash tree? How many internal

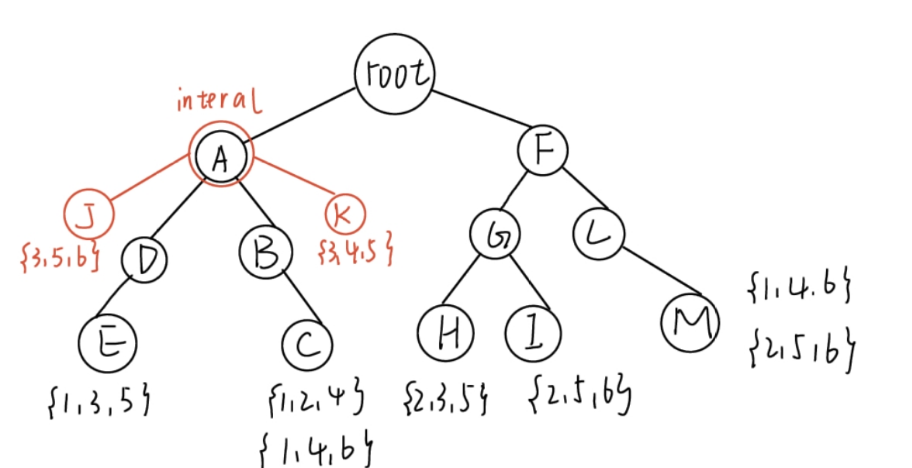
nodes are there?

c. Consider a transaction that contains the following items: {1, 2, 3, 5, 6}. Using

the hash tree constructed in part (a), which leaf nodes will be checked against the

transaction? What are the candidate 3-itemsets contained in the transaction?

**a**.



**b**.

There are 6 leaf nodes.

There are 2 internal nodes.

**c**.

The leaf nodes checked for transaction {1, 2, 3, 5, 6} include the leaf nodes storing {3, 5, 6} and other leaf nodes storing itemsets related to {1, 2, 3, 5, 6}.

The three candidate itemsets included in the transaction are {1, 3, 5}, {2, 3, 5}, {2, 5, 6}, {3, 5, 6}.



**~~Problem 3~~**

The Apriori algorithm uses a generate-and-count strategy for deriving frequent

itemsets. Candidate itemsets of size k+1 are created by joining a pair of frequent

itemsets of size k (this is known as the candidate generation step). A candidate is

discarded if any one of its subsets is found to be infrequent during the candidate

pruning step. Suppose the Apriori algorithm is applied to the data set shown in Table

2

.0 with minsup=30% , i.e., any itemset occurring in less than 3 transactions is

considered to be infrequent.

Table: 2.0

**Items** Bought

**T**ransaction

**ID**

**T1**

**T2**

**T3**

**T4**

**T5**

**T6**

**T7**

**T8**

**T9**

**T10**

a, b, x, y

b, x, y

a, y, z

a, b, x, z

x, y

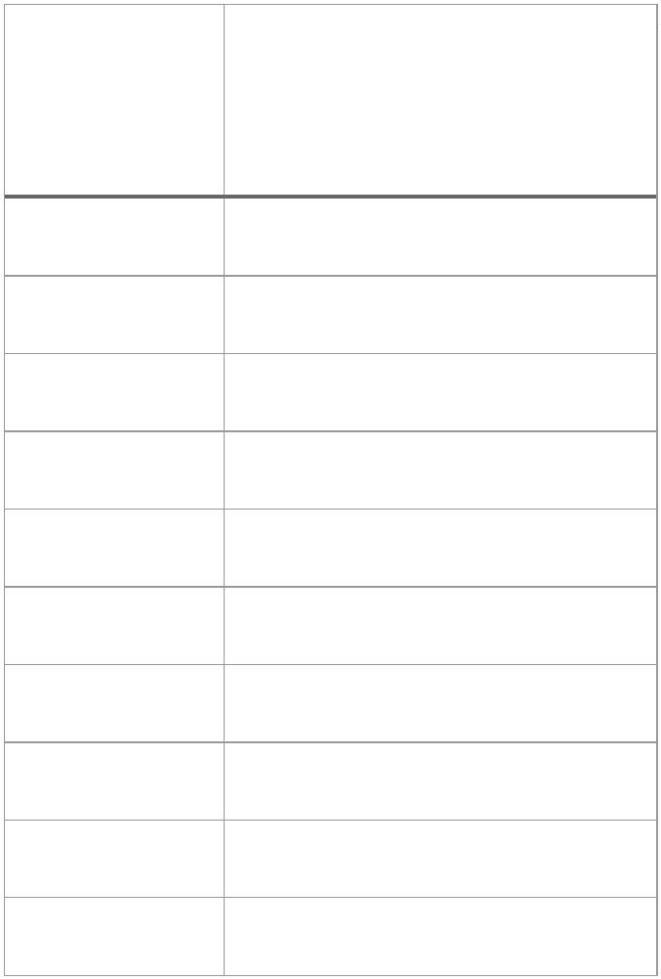
b, z

a, x, y, z

a, b

b, y, z

a, b, x, y



a. Draw an itemset lattice representing the data set given in Table 2.0 . Label each

node in the lattice with the following letter(s):

i. N: If the itemset is not considered to be a candidate itemset by the Apriori

algorithm. There are two reasons for an itemset not to be considered as a

candidate itemset: (1) it is not generated at all during the candidate

generation step, or (2) it is generated during the candidate generation step but

is subsequently removed during the candidate pruning step because one of its

subsets is found to be infrequent.

ii. F: If the candidate itemset is found to be frequent by the Apriori algorithm.

iii. I: If the candidate itemset is found to be infrequent after support counting.

b. What is the percentage of frequent itemsets (with respect to all itemsets in the

lattice)?

c. What is the pruning ratio of the Apriori algorithm on this data set? (Pruning ratio is

defined as the percentage of itemsets not considered to be a candidate because (1)

they are not generated during candidate generation or (2) they are pruned during the

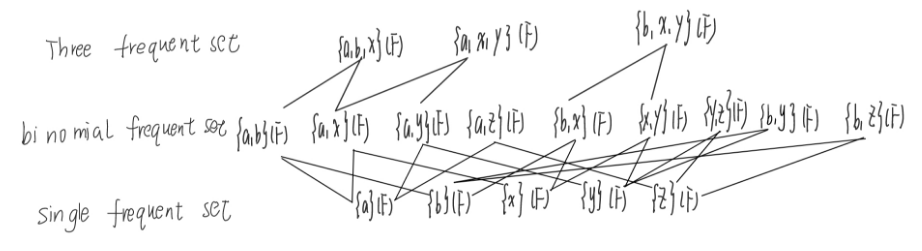
candidate pruning step.)

d. What is the false alarm rate (i.e., percentage of candidate itemsets that are found to

be infrequent after performing support counting)?

**E.N.D**

1. The total number of transactions is n=10, and the minimum support is 10 \* 30%=3.



Other itemsets are denoted as N.

**b**.

The total number of itemsets is 10+5+10+5+1=31

The number of frequent item sets is 5+10+7+1=23

The percentage of frequent items is 23/31 ≈ 74.2%

**c**.

Number of non candidate sets: 31 − 23=8

Pruning rate: 8/31 ≈ 25.8%

**d**.0