Midterm exam

Question 1 (10 points):
Calculate city block, Euclidean and supremum distances for the below data.
(a). x = (1, -1, 10, 3, 4), y = (10, -1, 4, 5, 2)
(b). x1 = (5, 4), x2 = (-2, 3)

(a). city block: 19
Supremum: 9
Euclidean = \sqrt(9^2+6^2+2^2+2^2) = 11.18

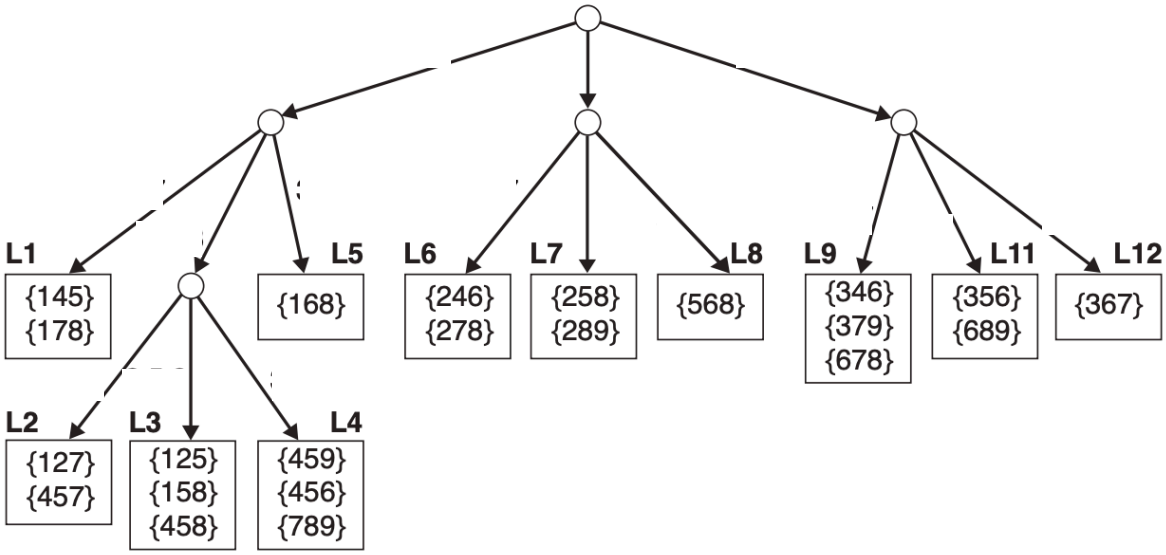
(b). city block: 8
Supremum = 7
Euclidean = \sqrt(7^2+1^1) = 7.07

Question 2 (15 points):
Discuss the advantages and disadvantages of using sampling to reduce the number of data objects. Would simple random sampling (without replacement) be a good approach to sampling? Why or why not? What kind of sampling method that you would like to use?

Answer: Simple random sampling is not the best approach since it will eliminate most of the points in sparse regions. It is better to under sample the regions where data objects are too dense while keeping most or all of the data objects from sparse regions.

Question 3 (10 points):
The Apriori algorithm uses a hash tree data structure to efficiently count the support of candidate itemsets. Consider the hash tree for candidate 3-itemsets as show in the below figure.

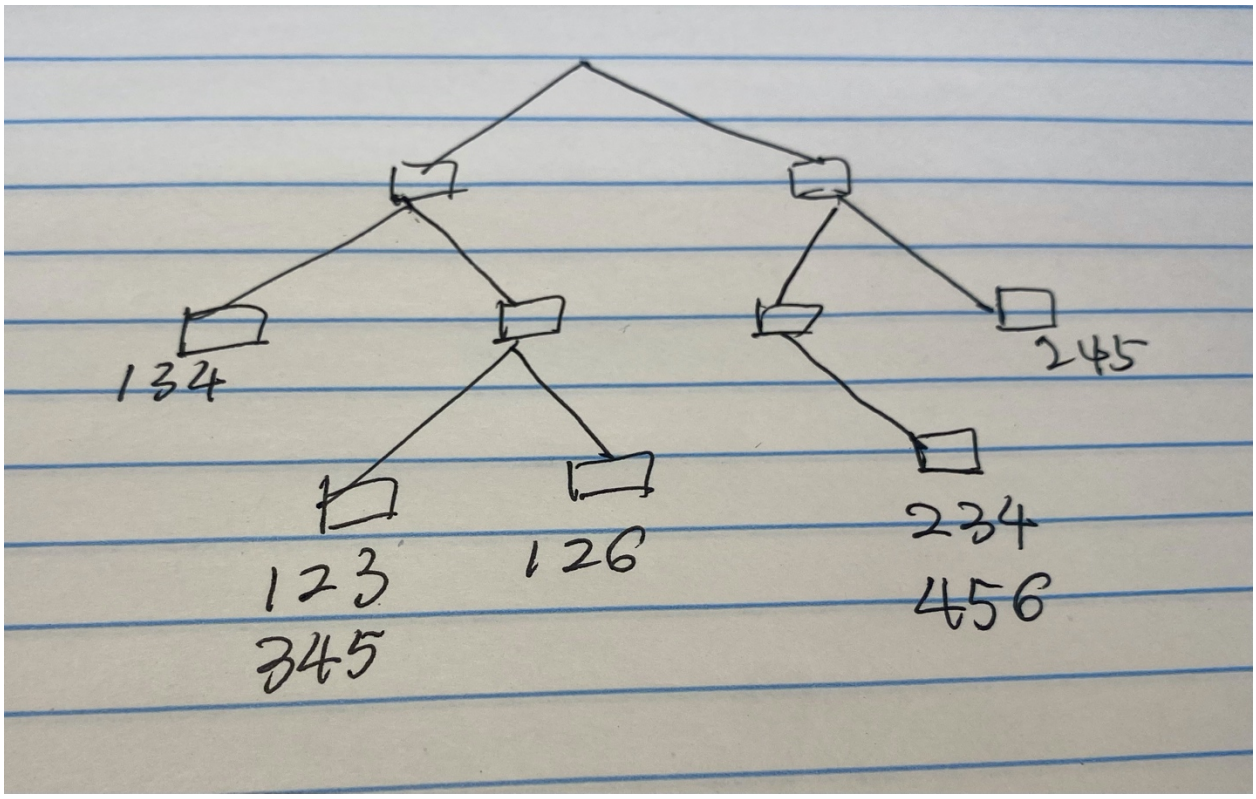
Given a transaction that contains items {1,3,4,5,8}, which of the hash tree leaf nodes will be visited when finding the candidates of the transaction?



Answer: The leaf nodes visited are L1, L3, L5, L9, and L11.

Question 4 (15 points):
Consider the following set of candidate 3-itemsets:
{1,2,3}, {1,2,6}, {1,3,4}, {2,3,4}, {2,4,5}, {3,4,5}, {4,5,6}

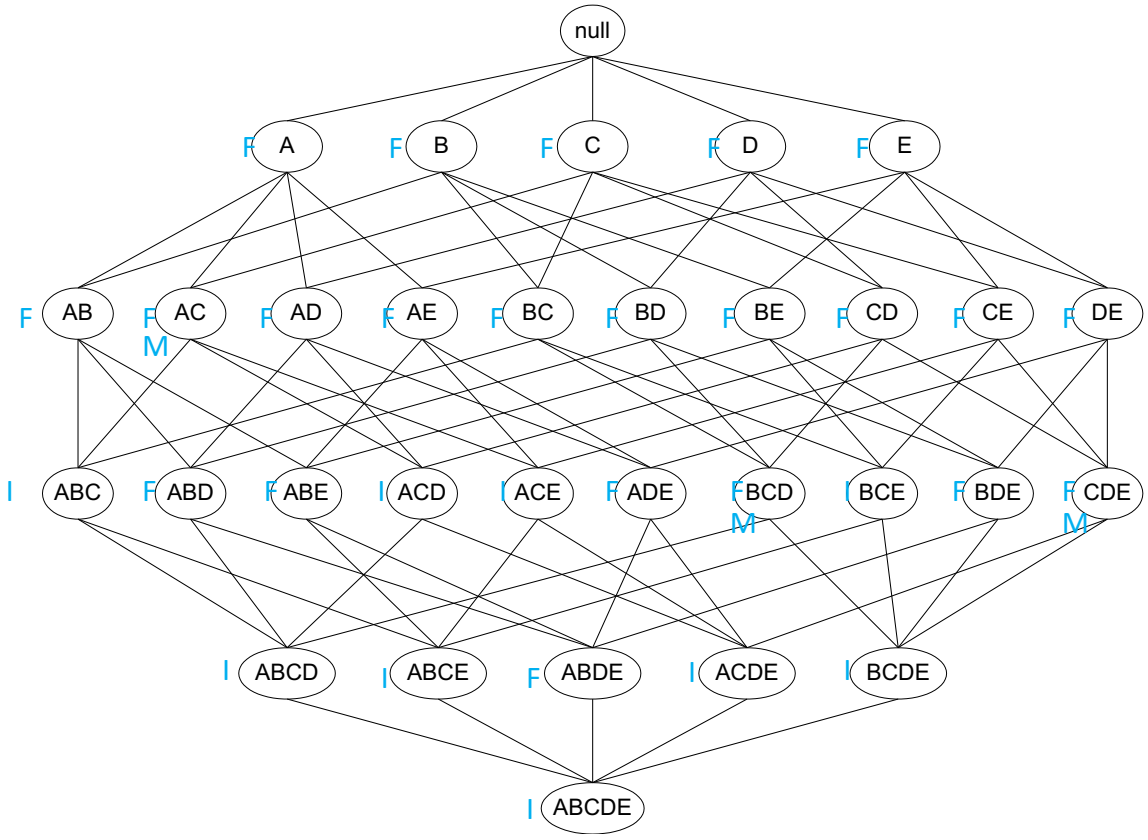
Construct a hash tree for the above candidate 3-itemsets. Assume the tree uses a hash function where:
1. all odd-numbered items are hashed to the left child of a node
2. the even-numbered items are hashed to the right child.



Question 5 (30 points)

Given the lattice structure in the below picture and the transactions given in the below table, label each node with the following letter(s): I if it is infrequent; F if it is frequent; M if the node is a maximal frequent itemset. Assume that the support threshold (minimum support) is 20%.

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	$\{b, d\}$



Question 6 (20 points)
Consider the dataset in the below table. A rule is considered to be strong if its support exceeds 10%. The dataset in the Table supports the following two strong rules:
(a). $\{(1 \leq A \leq 2), B = 1\} \rightarrow \{C = 1\}$
(b). $\{(5 \leq A \leq 8), B = 1\} \rightarrow \{C = 1\}$

Compute the support and confidence for both rules.

A	B	C
1	1	1
2	1	1
3	1	0
4	1	0
5	1	1
6	0	1
7	0	0
8	1	1
9	0	0
10	0	0
11	0	0
12	0	1

Answer:

s({(1 ≤ A ≤ 2), B = 1}→{C = 1})=1/6
c({(1 ≤ A ≤ 2), B = 1}→{C = 1})=1
s({(5 ≤ A ≤ 8), B = 1}→{C = 1})=1/6
c({(5 ≤ A ≤ 8), B = 1}→{C = 1})=1

Extra points (10 points)

Consider the following set of frequent 3-itemsets:

Assume that there are only 5 items in the dataset.
(a). List all candidate 4-itemsets obtained by the candidate generation procedure in Apriori.
(b). List all the candidate 4-itemsets that survive the candidate pruning step of the Apriori algorithm.

$$\{1,2,3\},\{1,2,4\},\{1,2,5\},\{1,3,4\},\{1,3,5\},\{2,3,4\},\{2,3,5\},\{3,4,5\}.$$

(a) $\{1, 2, 3, 4\},\{1, 2, 3, 5\},\{1, 2, 3, 6\}.\ \{1, 2, 4, 5\},\{1, 2, 4, 6\},\{1, 2, 5, 6\}$

$\{1, 3, 4, 5\},\{1, 3, 4, 6\},\{2, 3, 4, 5\}.\ \{2, 3, 4, 6\},\{2, 3, 5, 6\}.$

(b). {1, 2, 3, 4}, {1, 2, 3, 5}, {1, 2, 4, 5}, {2, 3, 4, 5}, {2, 3, 4, 6}.