**Assignment 3: Frequent Itemsets and Association Rules**

This assignment uses arules package of R. Download and install R from <https://www.r-project.org/> and RStudio from <https://rstudio.com/>. R has a GUI, but working on RStudio is recommended.

To submit, put the questions and your answers in a PDF file and submit to e-learning.

**Question 1:**

Consider the following set of frequent 3-itemsets:

*{*1*,* 2*,* 3*}, {*1*,* 2*,* 4*}, {*1*,* 2*,* 5*}, {*1*,* 3*,* 4*}, {*1*,* 3*,* 5*}, {*2*,* 3*,* 4*}, {*2*,* 3*,* 5*}, {*3*,* 4*,* 5*}.*

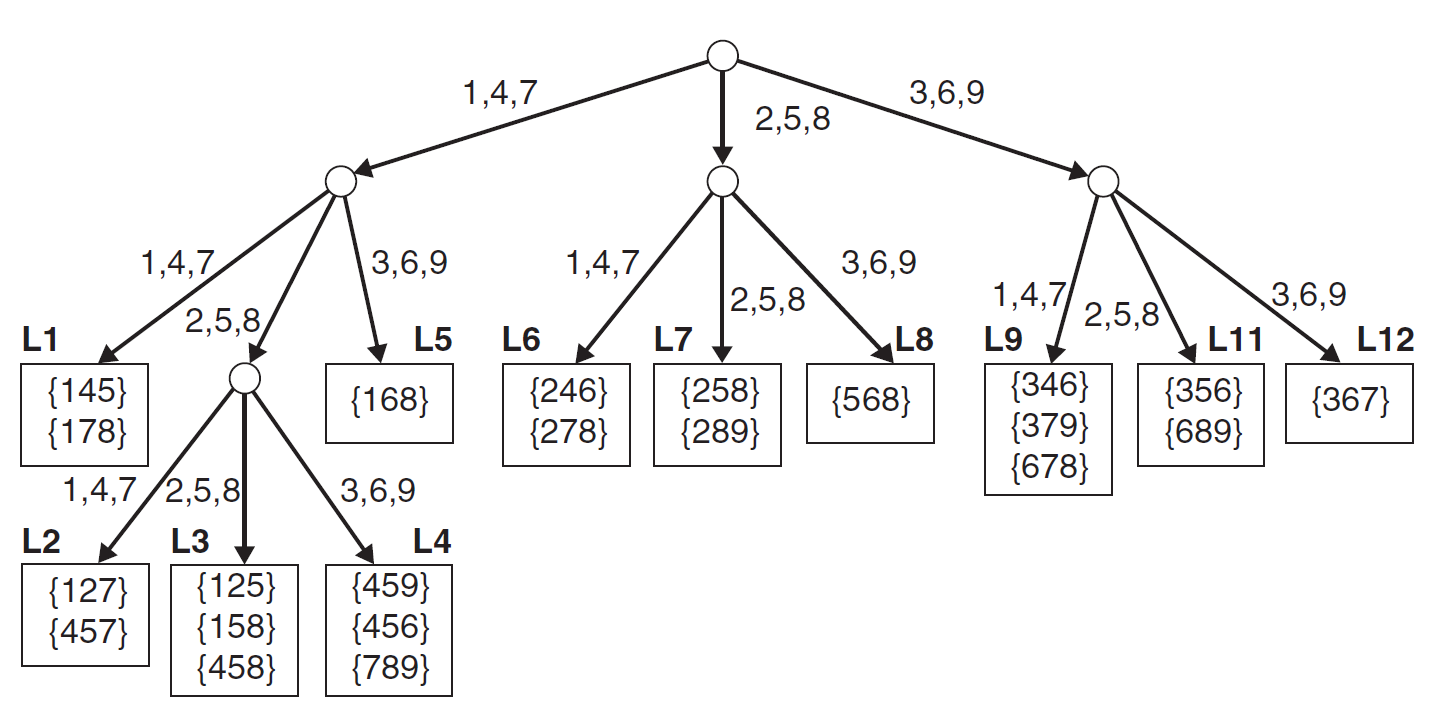
Assume that there are only five items {1,2,3,4,5} in the data set.

1. List all candidate 4-itemsets obtained by the candidate generation procedure in *Apriori*.
2. List all candidate 4-itemsets that survive the candidate pruning step of the *Apriori* algorithm.

**Question 2:**

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 shown in the following figure.

1. 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?
2. Use the visited leaf nodes in part (a) to determine the candidate itemsets that are contained in the transaction *{*1*,* 3*,* 4*,* 5*,* 8*}*.



**Question 3:**

Suppose the *Apriori* algorithm is applied to the data set shown in the following table with *minsup* = 30%, i.e., any itemset occurring in less than 3 transactions is considered to be infrequent.

|  |  |
| --- | --- |
| 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}* |

1. Draw an itemset lattice representing the data set given in the table. Label each node in the lattice with the following letter(s):
   1. **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.
   2. **F**: If the candidate itemset is found to be frequent by the *Apriori* algorithm.
   3. **I**: If the candidate itemset is found to be infrequent after support counting.
2. Use the arules package in R, generate all frequent itemses when the minimum support is 30%. List them by support in decreasing order.
3. Use the arules package in R, generate all association rules when the minimum support is 30% and the minimum confidence is 50%. List them by lift in decreasing order.