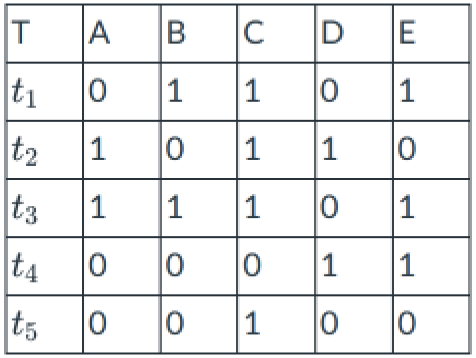
# Assignment 1

UNIVERSITY OF MEMPHIS

S. Parisa Daj. U00743495

COMP8118-Data Mining

09/06/2022



| T | A | B | C | D | E |
|---|---|---|---|---|---|
| $t_1$ | 0 | 1 | 1 | 0 | 1 |
| $t_2$ | 1 | 0 | 1 | 1 | 0 |
| $t_3$ | 1 | 1 | 1 | 0 | 1 |
| $t_4$ | 0 | 0 | 0 | 1 | 1 |
| $t_5$ | 0 | 0 | 1 | 0 | 0 |

| T | Items |
|---|---|
| T1 | BCE |
| T2 | ACD |
| T3 | ABCE |
| T4 | DE |
| T5 | C |

| Item | COUNT |
|---|---|
| A | 2 |
| B | 2 |
| C | 4 |
| D | 2 |
| E | 3 |

## Q1

1. The first scan: Count for all items that are greater or equal than the support so each item itself is a large itemset as well. Then the first candidate is generated by the join rule. (As with all items are large itemsets, there is no pruning in here.)
    a. $L_1$ = {A, B, C, D, E}
    b. $C_2$ = {AB, AC, AD, AE, BC, BD, BE, CD, CE, DE}
2. The second scan: L2 includes the itemsets that have the support of 2. Joining the itemsets in L2, results in four itemsets as in C3 that after pruning, only one of them remains based on property 2 (If an itemset S is not large, then any proper superset of S must not be large.)
    a. $L_2$ = {AC, BC, BE, CE}
    b. $C_3$ = {~~ABC, ABE, ACE~~, BCE}
3. The third scan: The only remaining itemset has a support of 2, so can be in L3.
    a. $L_3$ = {BCE}
4. Finally, L is the union of all Lis found above.
    a. L = {A, B, C, D, E, AC, BC, BE, CE, BCE}

| Itemset | COUNT[1] |
|---|---|
| AB | 1 |
| AC | 2 |
| AD | 1 |
| AE | 1 |
| BC | 2 |
| BD | 0 |
| BE | 2 |
| CD | 1 |
| CE | 2 |
| DE | 1 |

| Itemset | COUNT |
|---|---|
| AC | 2 |
| BC | 2 |
| BE | 2 |
| CE | 2 |

| Itemset | COUNT |
|---|---|
| BCE | 2 |

[1] orange rows are eliminated and green ones stay

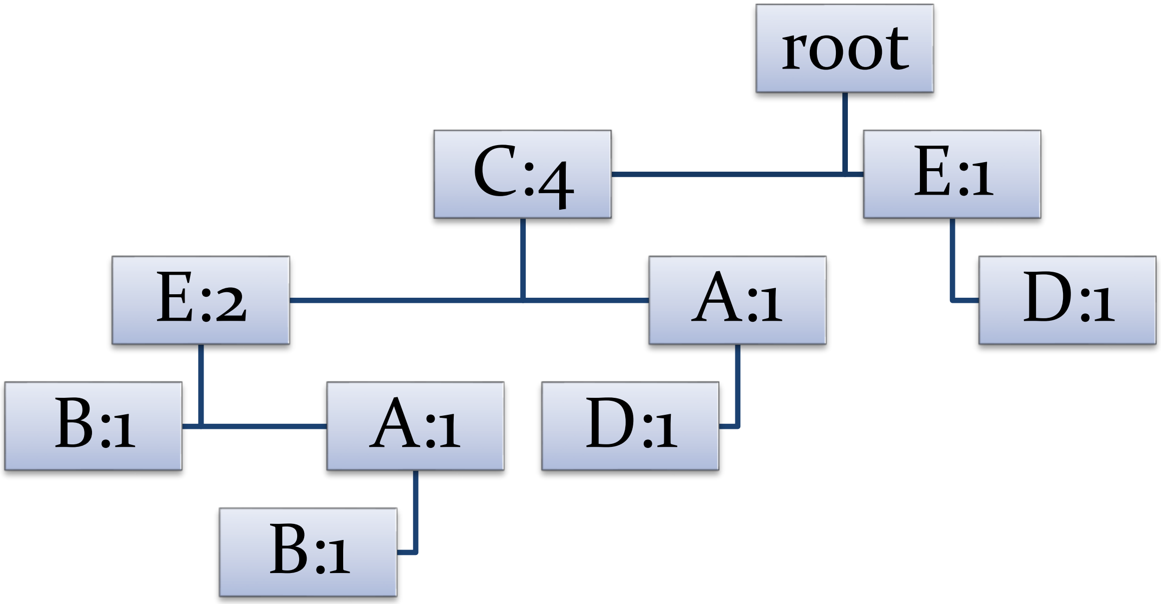
## Q2

1. Deduce the ordered frequent items. For items with the same frequency, the order is given by the alphabetical order

| Item | COUNT |
|---|---|
| C | 4 |
| E | 3 |
| A | 2 |
| B | 2 |
| D | 2 |

| T | Items | Ordered frequent items |
|---|---|---|
| T1 | BCE | CEB |
| T2 | ACD | CAD |
| T3 | ABCE | CEAB |
| T4 | DE | ED |
| T5 | C | C |

2. Construct the FP-tree using the ordered frequent items



3. Conditional FP-tree
   (I later realized there was no need to include them, so I decided to keep them here)
   a. On D: {{C1, A1, D1}, {E1, D1}}
   b. On B: {{C1, E1, B1}, {C1, E1, A1, B1}}
   c. On A: {{C1, E1, A1}, {C1, A1}}
   d. On E: {{C2, E2}, {E1}}
   e. On C: {{C4}}

| Item | Freq. |
|---|---|
| C | 1 |
| E | 1 |
| A | 1 |
| B | 0 |
| D | 2 |

| Item | Freq. |
|---|---|
| C | 2 |
| E | 2 |
| A | 1 |
| B | 2 |
| D | 0 |

| Item | Freq. |
|---|---|
| C | 2 |
| E | 1 |
| A | 2 |
| B | 0 |
| D | 0 |

| Item | Freq. |
|---|---|
| C | 1 |
| E | 2 |
| A | 0 |
| B | 0 |
| D | 0 |

| Item | Freq. |
|---|---|
| C | 4 |
| E | 0 |
| A | 0 |
| B | 0 |
| D | 0 |

## Q3

L = {A, B, C, D, E, AC, BC, BE, CE, BCE}

1. C(A -> C) = P(C|A) = support(A U C) / support(A) = 2 / 2 = 1
2. C(C -> A) = P(A|C) = support(A U C)) / support(C) = 2 / 4 = 0.5
3. C(B -> CE) = P(CE|B) = support(B U CE) / support(B) = 2 / 2 = 1
4. C(CE -> B) = P(B|CE) = support(B U CE)) / support(CE) = 2 / 2 = 1

## Q5

In the association rule, we learned to find frequent itemsets given a support value. If we consider the base string as a regular text, and we are searching for words separated by space, then each word is an item and at the same time a transaction as well. _It is also possible to make fewer transactions, but that can be more complicated_. The next step is to filter them by eliminating the items that have a length of greater than k. After finding the right items, we only need to count them and calculate the support for each item by counting their replications.

For instance, given the following string for this problem, with K = 3 and θ = 5 we first put the items in the table. Then eliminate long strings.
String: Can you can the can? Yes, I can can the can with the can-opener

By applying association rule to this example, we can clearly se
e that the word "can" is the most frequent item, and probably the most important word in this example.

| T | Items |
|---|---|
| T1 | can |
| T2 | you |
| T3 | can |
| T4 | the |
| T5 | can |
| T6 | yes |
| T7 | I |
| T8 | can |
| T9 | can |
| T10 | the |
| T11 | can |
| T12 | with |
| T13 | the |
| T14 | Can-opener |

| Item | COUNT |
|---|---|
| can | 6 |
| you | 1 |
| yes | 1 |
| I | 1 |
| the | 3 |