98

1 a)

Antecedents	Consequent	support	confidence
Beer	Diapers	3/5	1
Bread	Diapers	3/5	3/4
Milk	Diapers	3/5	3/4
Bread	Milk	3/5	3/4
Milk	Bread	3/5	3/4
Diapers	Milk	3/5	3/4
Diapers	Beer	3/5	3/4
Diapers	Bread	3/5	3/4

1 b) No 2 consequent Rules

1c)

Antecendent	Consequent	Support	confidence
Beer	Diapers	3/5	3/3

Scanned with CamScanner

TSK 6.2:

a) $\sigma(\{e\}) = \frac{8}{10} = 0.8$

$\sigma(\{b,d\}) = \frac{2}{10} = 0.2$

$\sigma(\{b,d,e\}) = \frac{2}{10} = 0.2$

b) Compute association rules $\{b,d\} \rightarrow \{e\}$ and $\{e\} \rightarrow \{b,d\}$. Is confidence a symmetric measure?

Sol:

$c(bd \rightarrow e) = \frac{0.2}{0.2} = 100\%$

$c(e \rightarrow bd) = \frac{0.2}{0.8} = 25\%$

No, confidence is not a symmetric measure.

c) Merging baskets by customer id, we get

Sol:

$s(\{e\}) = \frac{4}{5} = 0.8$

$s(\{b,d\}) = \frac{5}{5} = 1$

$s(\{b,d,e\}) = \frac{4}{5} = 0.8$

d) Now confidence for association rules,

Sol:

$c(bd \rightarrow e) = \frac{0.8}{1} = 80\%$

$c(e \rightarrow bd) = \frac{0.8}{0.8} = 100\%$

Scanned with CamScanner

(e) Is there any relation between s_1, s_2 or c_1, c_2 where s_1 and c_1 are support and confidence values of an association rule r when treating TID as market basket. s_2 and c_2 are support & confidence values of r when treating each customer ID as market basket.

Sol: No. There is no apparent relationships between s_1, s_2, c_1 & c_2

Scanned with CamScanner

TSK 6.6

a) what is the max no. of association rules that can be extracted from this data?

Sol: As there are 6 items in dataset,
Total no. of rules including rules that have zero support – 602

b) what is max size of frequent itemsets that can be extracted?

Sol: Assuming minsup > 0,
longest transaction contains 4 items
max size of frequent itemset is 4

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

Sol: $^6C_3 = 20$

d) It suffices to consider all 2-itemsets (why?)

Sol: we only need to consider the supports of 2-itemsets, since adding items will never grow the support. It either decrease or stay same.

An itemset with max support is

Sol: {Bread, Butter}.

Scanned with CamScanner

(e) $c(\{a\} \rightarrow \{b\}) = c(\{b\} \rightarrow \{a\}$ when

Sol:

when (a,b) is

(Beer, cookies)

or

(Bread, Butter)

Scanned with CamScanner

TSK 6.7

(a) List all candidate 4-itemsets obtained by a Candidate generation procedure using $F_{k-1} \times F_1$ merging strategy.

Sol:- {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}~~, ~~{2,4,5,6}~~

(b) Using $F_3 \times F_3$ method & ignoring duplicates, we get

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

(c) $f \in C_4$ survies pruning if all 3-subsets are frequent

		Pruning	subsets to check
{1,2,3,4}	—	No	{1,3,4} {2,3,4}
{1,2,3,5}	—	No	{1,3,5} {2,3,5}
{1,2,4,5}	—	Yes	{1,4,5} {2,4,5}
{1,3,4,5}	—	Yes	{1,4,5} {3,4,5}
{2,3,4,5}	—	Yes	{2,4,5} {3,4,5}

Scanned with CamScanner

Candidate 4-itemsets that survive the pruning are {1,2,3,4}
{1,2,3,5}

Scanned with CamScanner

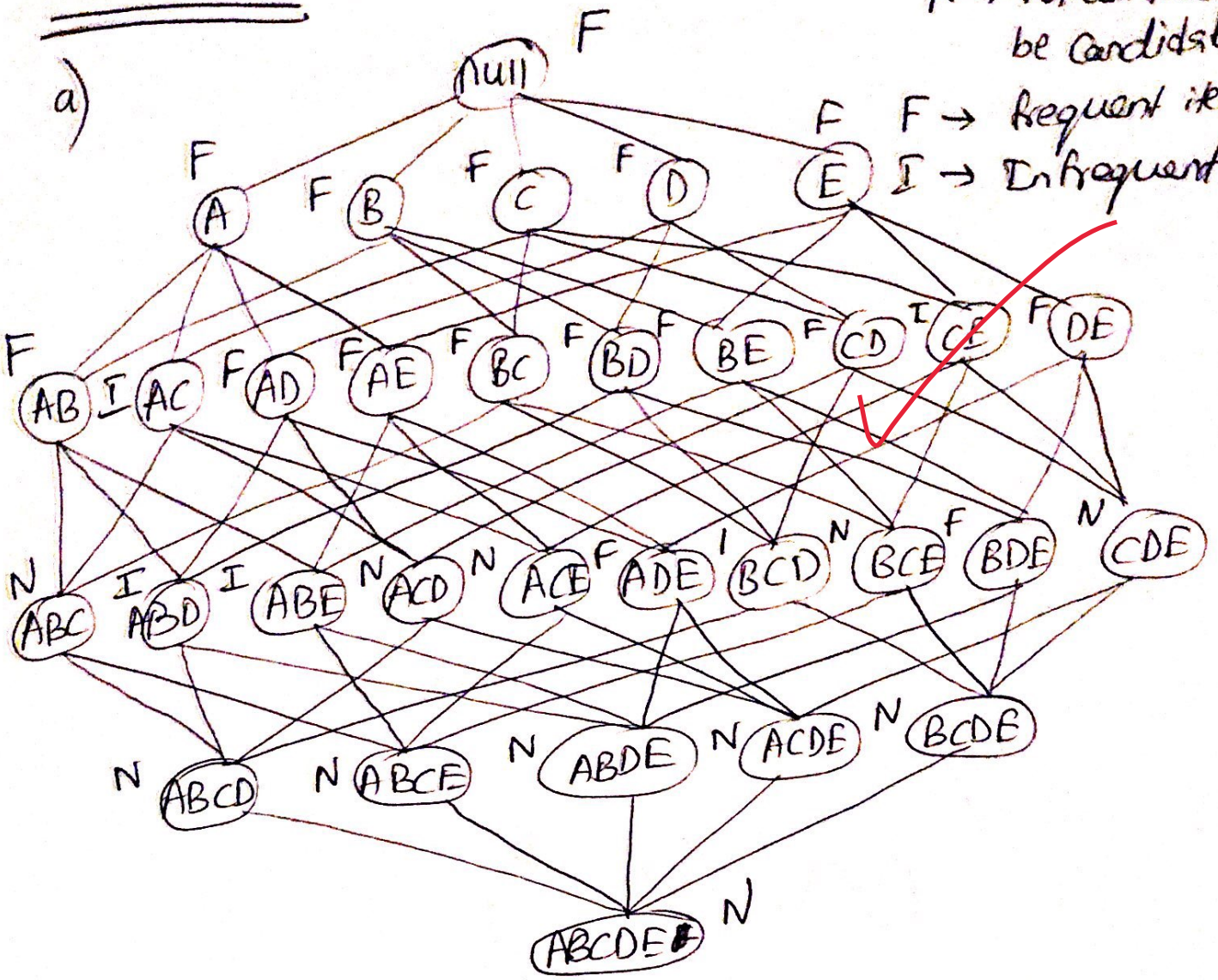
TSK : 6.8

N → not considered to be candidate items

F → frequent item

I → Infrequent

a)



b) what is % of frequent itemsets? (including null set)

Sol: 16/32 = 50%

-1

c) what is pruning ratio of Apriori algorithm?

Sol: Pruning ratio is ratio of N to tsl no. of itemsets.

∵ count of N = 11, pruning ratio is 11/32

= 34.4%

-1

Scanned with CamScanner

Tsk 6-9

a) Leaf nodes visited are

L1	L3	L5	L9	L11
{1,4,5}	{1,2,5}	{1,6,8}	{3,4,6}	{3,5,6}
{1,7,8}	{1,5,8}		{3,7,9}	{6,8,9}
	{4,5,8}		{6,7,8}	

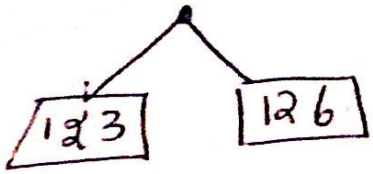
b)

{1,4,5}
{1,5,8}
{4,5,8}

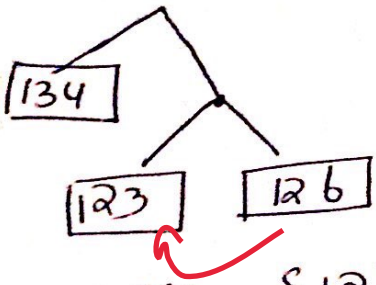
Scanned with CamScanner

TSK 6-10

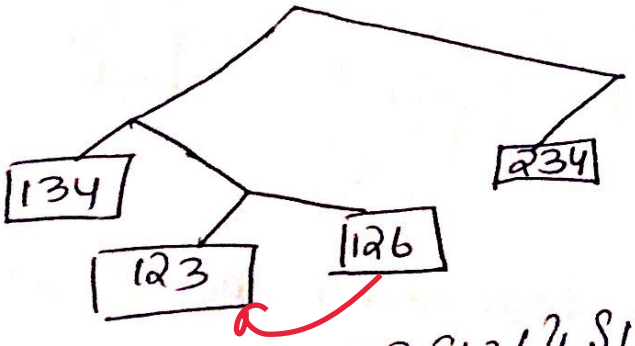
a) Hash tree after {1,2,3}, {1,2,6}



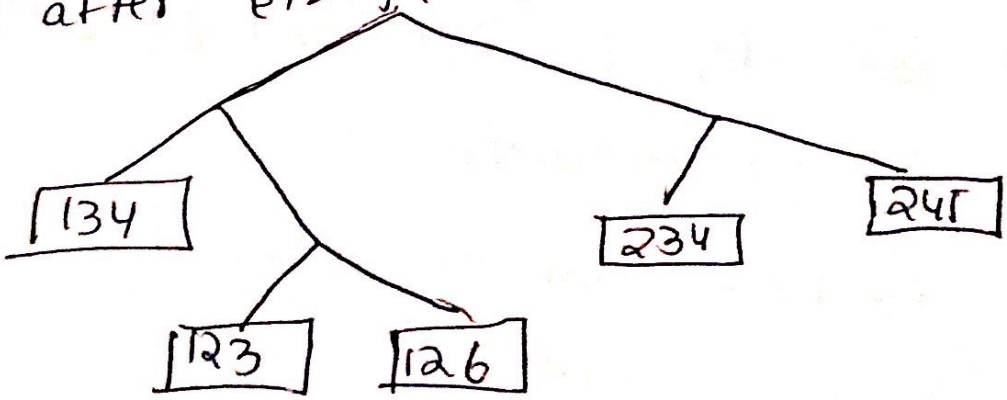
hash tree after {123} {126} {134}



hash tree after {123} {126} {134} {234}

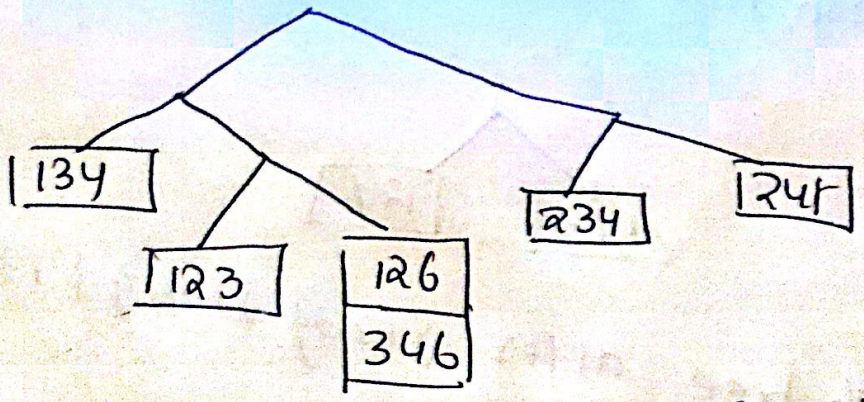


hash tree after {123} {126} {134} {234} {245}

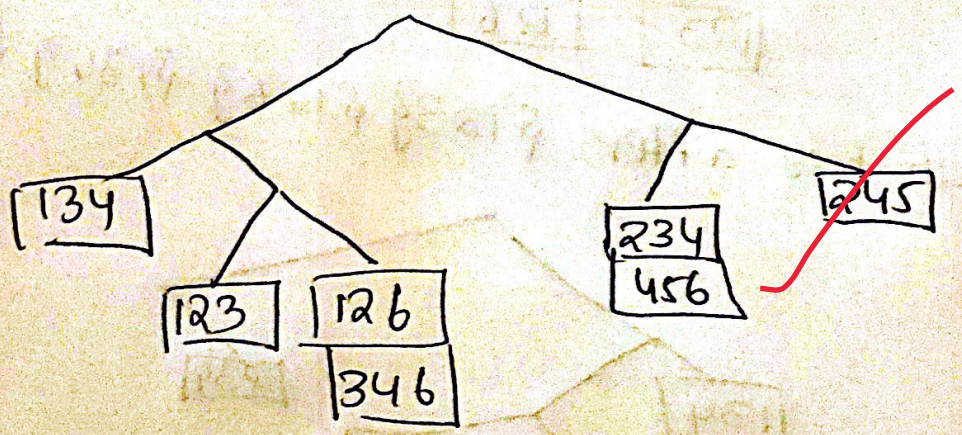


Scanned with CamScanner

hash tree after {123} {126} {134} {234} {245} {346}



hash tree after {123} {126} {134} {234} {245} {346} {456}



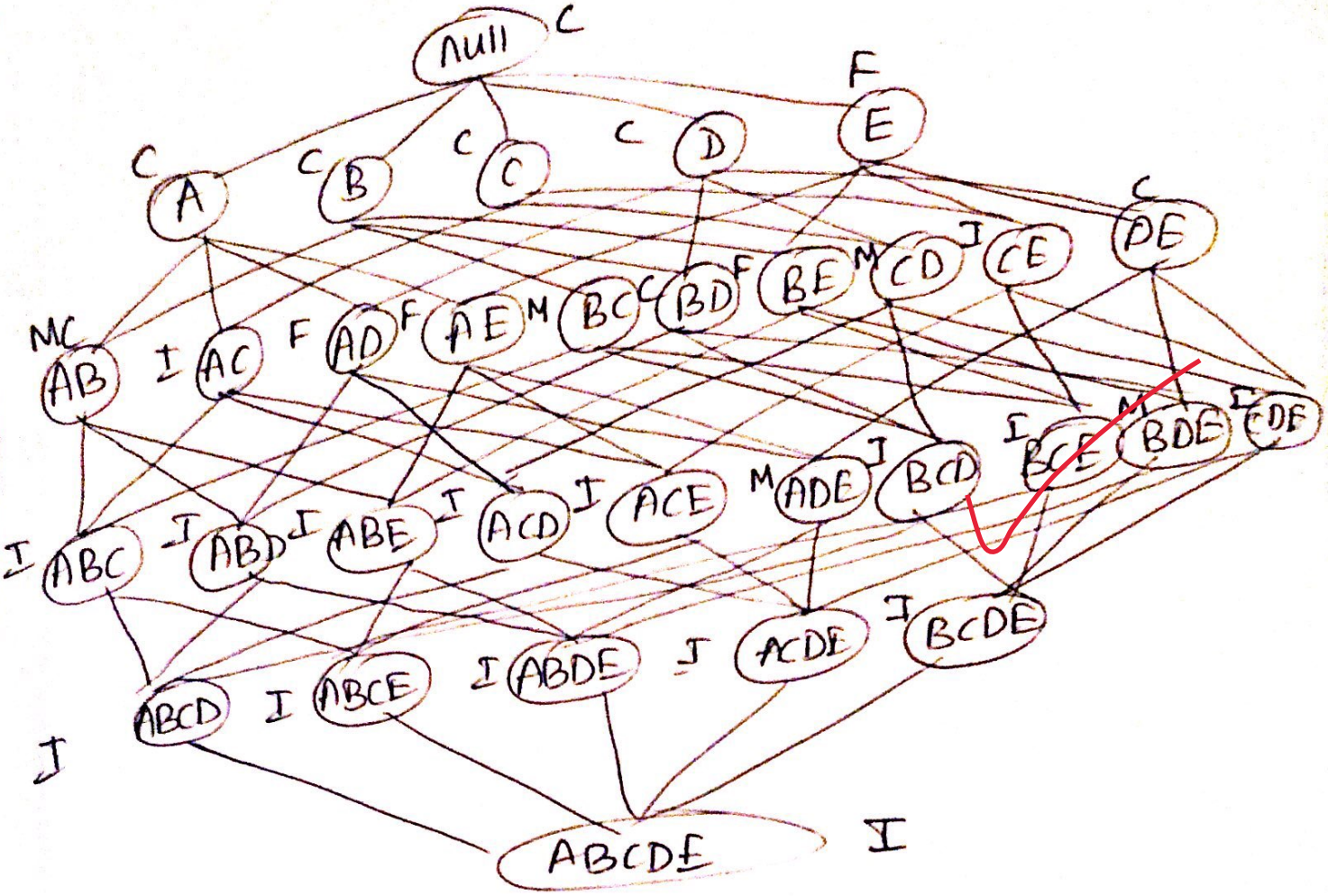
b) There are 5 leaf nodes and 4 internal nodes

c) Leaf nodes L1, L2, L3 and L4 will be checked against transaction

Candidate itemsets contained in transaction are {123} {126}

Scanned with CamScanner

TSC 6.11



Scanned with CamScanner

FP-Growth Exercise

In this problem, we will begin with the dataset to the right, and develop all frequent itemsets and their support counts. The minimum support count for frequent itemsets is 4 = minsup. We will use FPGrowth

Transaction Data Set

TID	Items
1	{a,b}
2	{b,c,d}
3	{a,c,d,e}
4	{a,d,e}
5	{a,b,c}
6	{a,b,c,d}
7	{a}
8	{a,b,c}
9	{a,b,d}
10	{b,c,e}

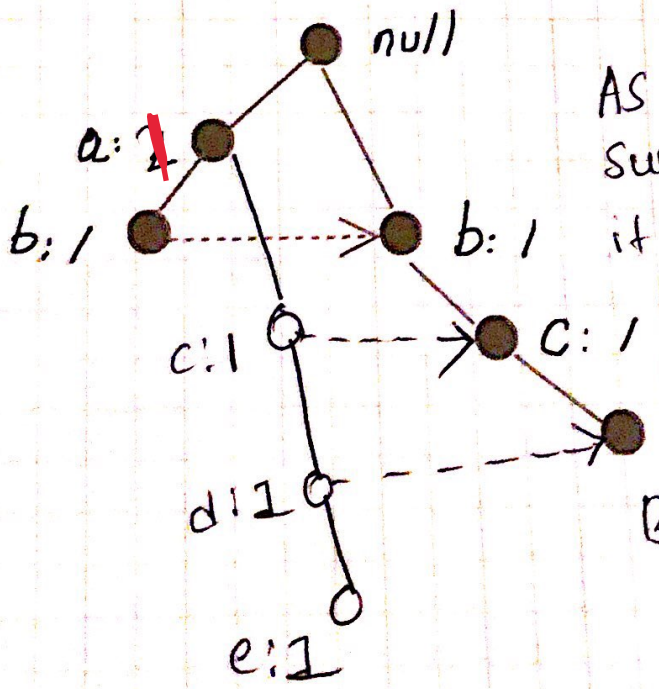
(a) Find the supports of all 1-itemsets, in descending order of supports. Retain the frequent itemsets.

σ(a) = 8/10
σ(b) = 7/10
σ(c) = 6/10
σ(d) = 5/10
σ(e) = 3/10

set

Freq itemset = {a, b, c, d}

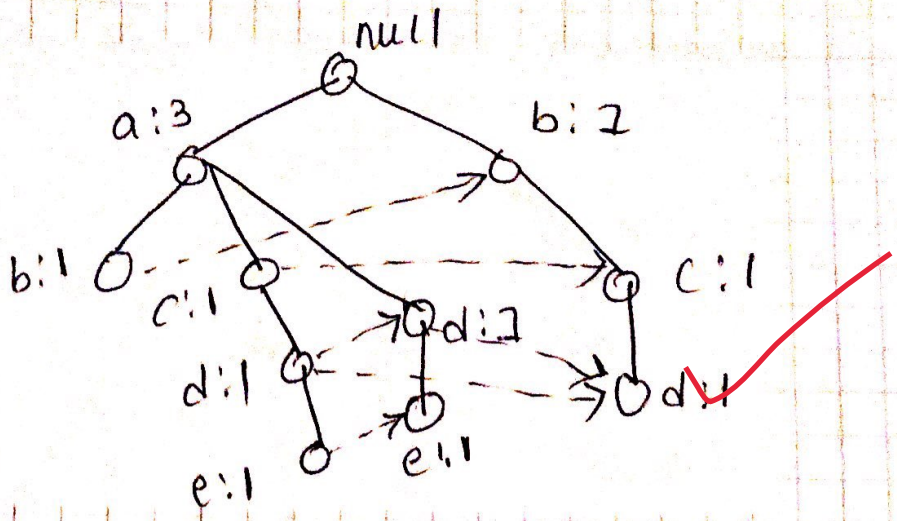
(b) Draw the F-P tree after the first 2 transactions. Why can we ignore "e" in the baskets?



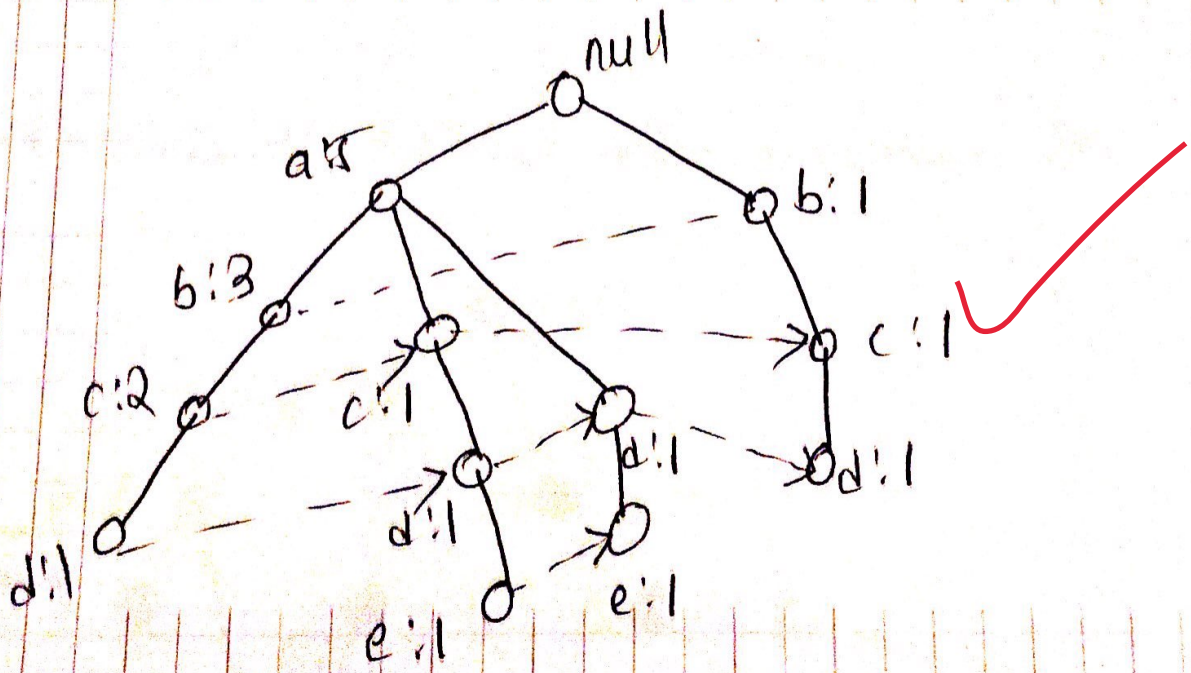
As minsup = 4 and support of e < 4, it is infrequent itemset. ∴ we can ignore it.

But im drawing just for my understanding

(c.) Draw the FP-tree after transaction 4.

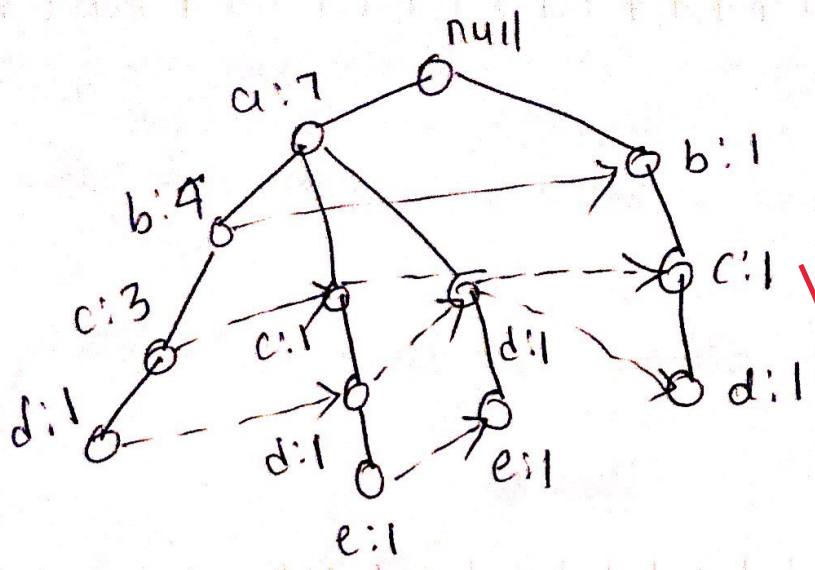


(d) Draw the FP-tree after transaction 6.

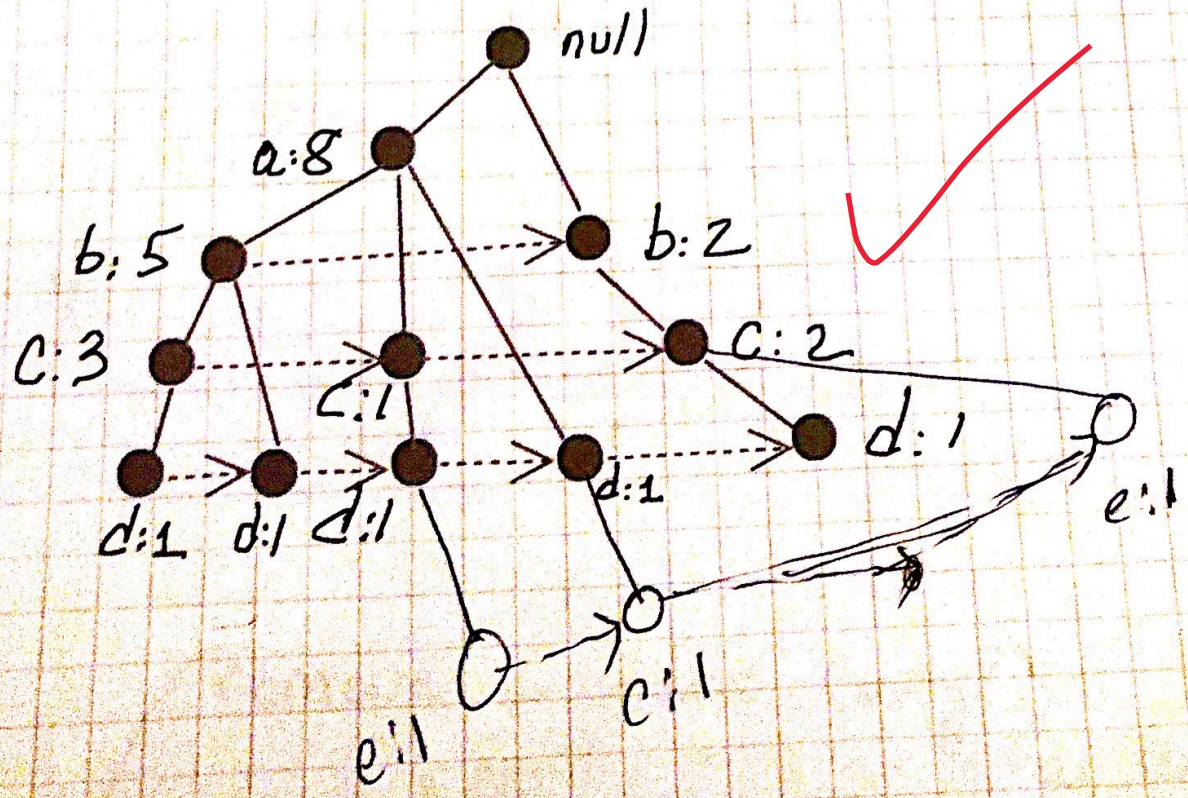


Scanned with CamScanner

(e) Draw the FP-tree after transaction 8.



(f) Draw the FP-tree after transaction 10.



Scanned with CamScanner

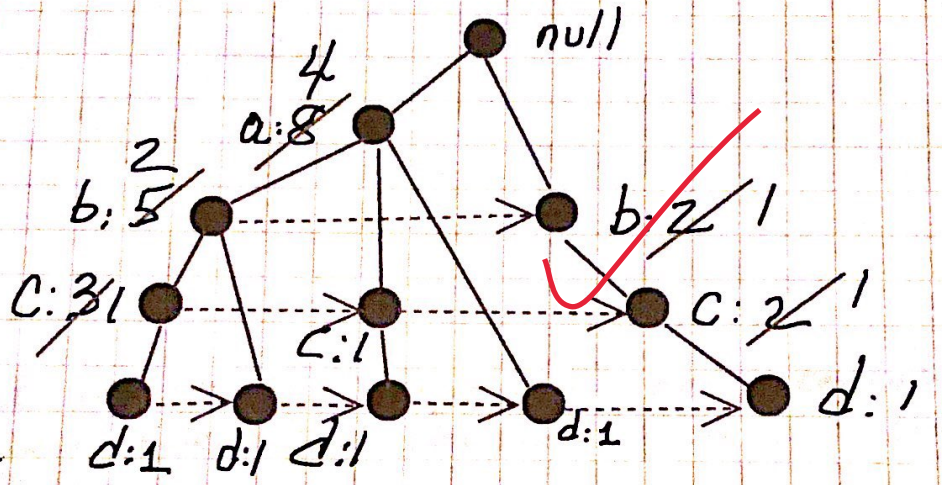
(g) From (f), explain how to compute the support of {d}. Why is {d} frequent?

To compute the support of d, count d nodes in FP tree and sum up the frequencies of all d nodes.

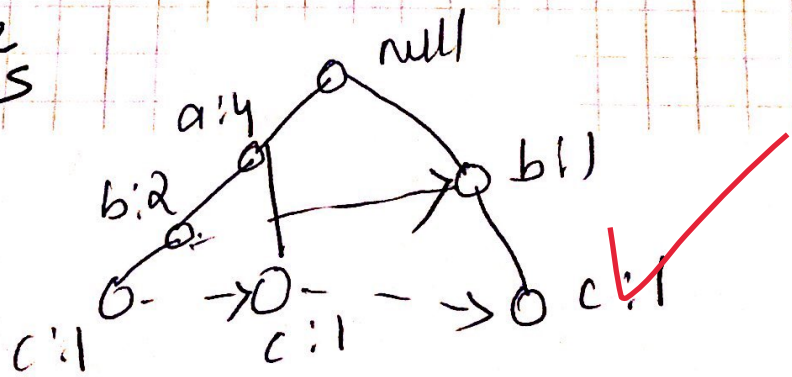
$\sigma(d) = 5 > 4$ (minsup). $\therefore$ it is frequent.

(h) Use TSK p349, 3.) to find the conditional FP tree of d.

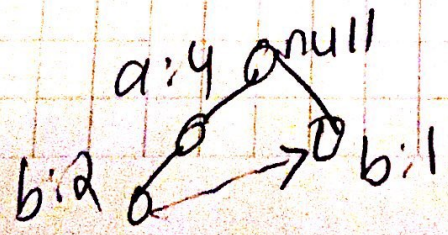
adjust supports for paths ending in d.



remove d-nodes



remove c-nodes (why?)

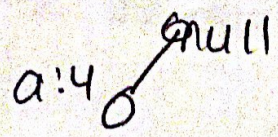


$\therefore$ it has low support count than minsup

Scanned with CamScanner

remove b. nodes (why?)

∵ the support count of b is 3 which is less than minsup. ∴ we remove b nodes

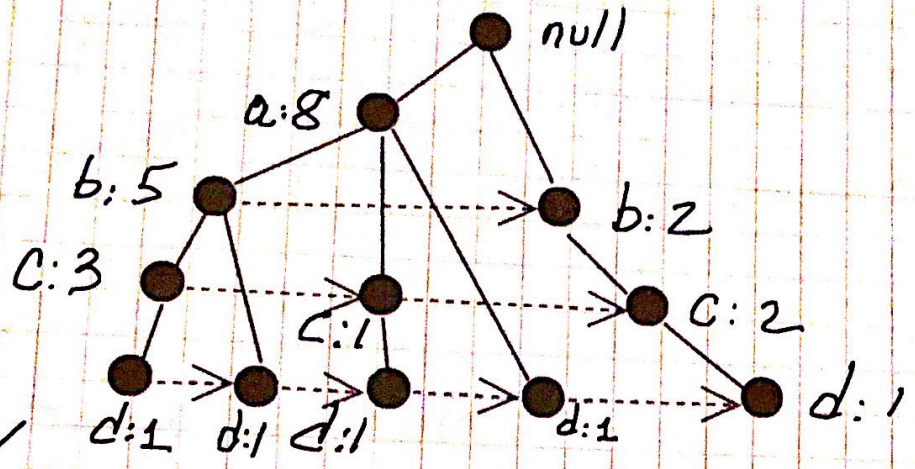


(i) What are all the frequent itemsets containing d and other items obtained from this conditional FP tree?

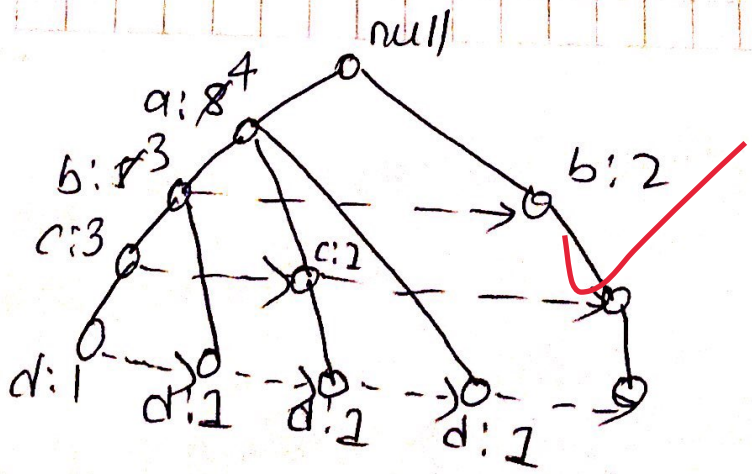
~~'d'~~ , 'ad'

Scanned with CamScanner

(j) Find the conditional FP tree for c.

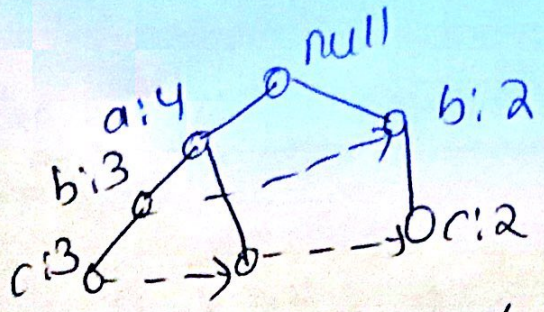


adjust supports for paths ending in c



remove paths not ending in c

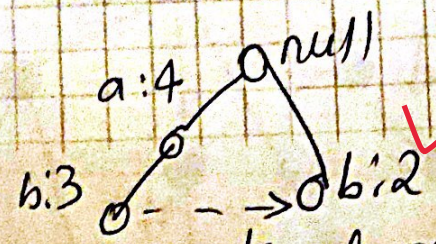
Scanned with CamScanner



support count of c = 6 > minsup = 4

c is frequent.

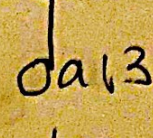
remove c nodes



By adding frequency counts of node 'b', we obtain support count of {b, c}. {b, c} is frequent itemset since its support count = 5

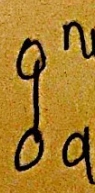
Next, FP tree for 'bc'. After updating support count

Conditional FP tree for 'bc' is



∵ a = 3 < minsup = 4 'abc' is not freq item set

Next algo will construct 'ac'



∵ a = 4 ≥ minsup = 4

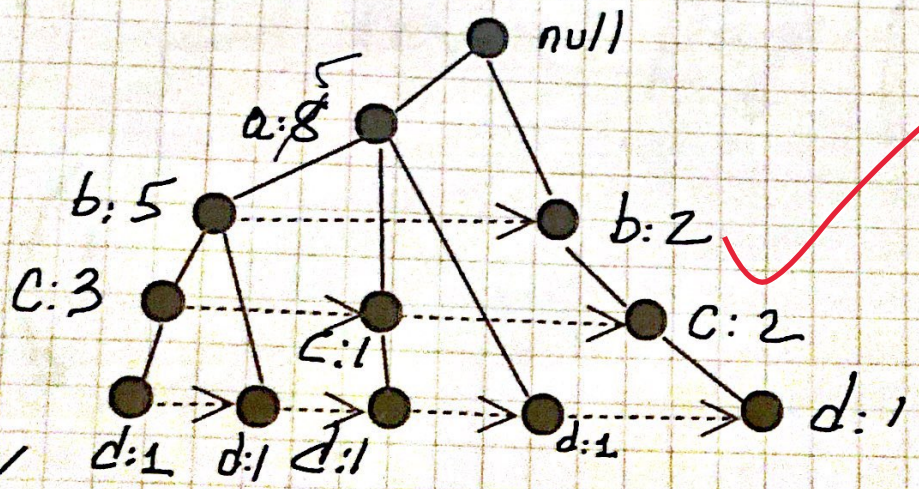
∴ 'ac' is frequent.

Scanned with CamScanner

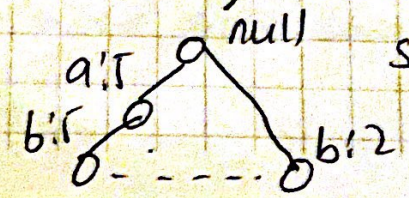
k) What are all the frequent itemsets containing c and other items obtained from this conditional FP tree?

~~'c'~~, 'bc', 'ac'

l) Find the conditional FP-Tree for b.

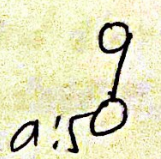


adjust supports for paths ending in b.



support count of b=7 > minsup=4

Removing b nodes,



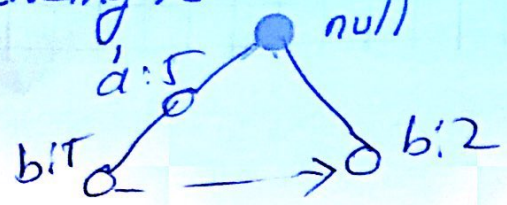
b

c:3

support count for a=5 > minsup=4

∴ 'ab' is a frequent itemset

remove paths not ending in b



support count of b = 7
→ mistake

remove b nodes

Previous page

m) What are all the frequent itemsets containing b and other items obtained from this conditional FP tree?

{ab} ~~{b}~~

n) List all frequent itemsets and their supports in lexicographical order.

itemset	support
~~a~~	~~0.8~~
ab	0.5
ac	0.4
ad	0.4
~~b~~	~~0.7~~
bc	0.4
~~c~~	~~0.6~~
~~d~~	~~0.5~~

Scanned with CamScanner

o.) List all association rules with minimum support of 4 in descending order of confidence

association rule	confidence
$c \rightarrow b$	0.83
$d \rightarrow a$	0.8
$b \rightarrow a$	0.71
$b \rightarrow c$	0.71
$c \rightarrow a$	0.67
$a \rightarrow b$	0.62
$a \rightarrow c$	0.5
$a \rightarrow d$	0.5

Scanned with CamScanner