Name: Ayush Jain
SAP ID: 60004200132
Div: B
Computer Engineering

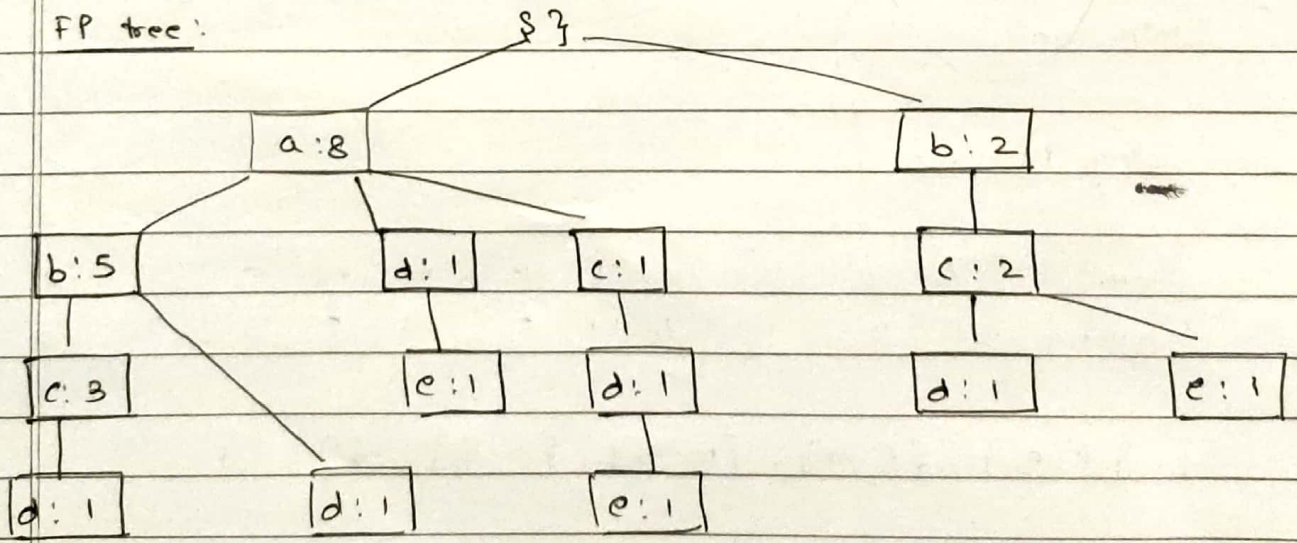
DMW - Assignment 3

1> Construct FP tree with support = 2

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

L = (a, 8), (b, 7), (c, 6), (d, 5), (e, 3)

FP tree:



ITEM	Conditional Pattern base	conditional FP tree	FP Generated
e	(a→c→d=1), (b→c=1), (a→d=1)	{a:2 d:2 c:2}	{a,e}, {d,e}, {c,e}, {a,d,e}
d	(a→b=1), (a→b→c=1), (a=1), (a→c=1), (b→c=1)	{a:4, b:2}, {a:4, c:2}, {b:2, c:2}	{a,d}, {b,d}, {c,d}, {a,b,d}, {a,c,d}, {b,c,d}
c	(a→b=3), (a=1), (b=2)	{a:4, b:5}	{a,b,c}, {a,c}, {b,c}
b	(a=5)	{a:5}	{a,b}

Q. 2) min_supp = 60%. Support = 0.6 × 5 = 3

min_conf = 80%.

Using FP tree:

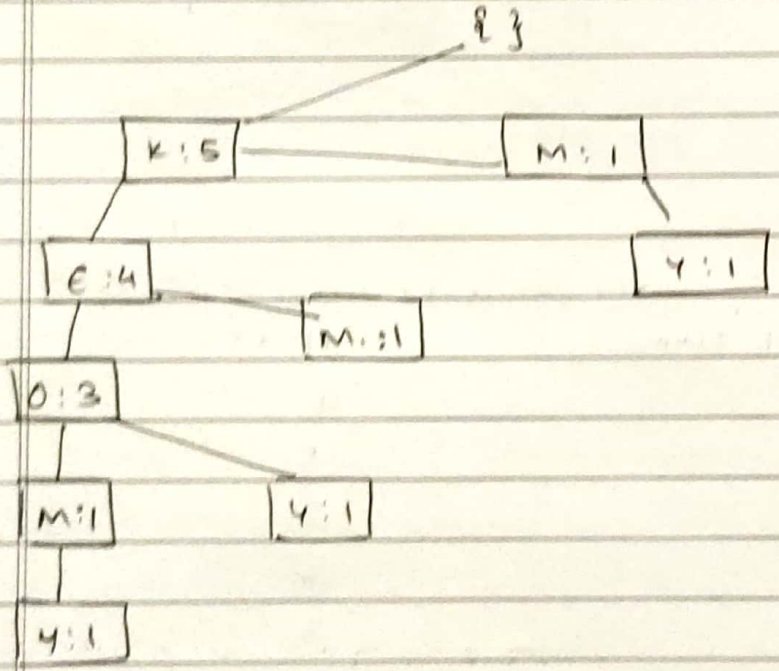
Item	A	C	D	E	I	K	M	N	O	U	V
Frequency	1	2	1	4	1	5	3	2	3	1	3

L: {K:5}, {E:4}, {M:3}, {O:3}, {Y:3}

Arranging transaction in order of L

Tid	Ordered Items
t100	{K, E, O, M, Y}
t200	{K, E, O, Y}
t300	{K, E, M}
t400	{K, M, Y}
t500	{K, E, O}

FP Tree:



Using Apriori :

C1 :

Item set	Support
A	1
C	2
D	1
E	4
I	1
K	5
M	3
N	2
O	3
U	1
Y	3

L1 :

Frequent-1 Items	Support
E	4
K	5
M	3
O	3
Y	3

C2 :

Freq-2 Items	(E,K)	(E,M)	(E,O)	(E,Y)	(K,M)	(K,O)
Support	4	2	3	2	3	3

(K,Y)	(M,O)	(M,Y)	(O,Y)
3	1	2	3

L_2 :

Frequent 2-Itemset	(E,K)	(E,O)	(K,M)	(K,O)	(K,Y)
Support	4	3	3	3	3

L_3 :

Frequent -3 Items	Support
(E,K,O)	3

∴ No more combinations can be formed.

∴ L = {E,K,O}

∴ Association rules for Apriori :

1	{K} → {O,E}	= 3/5	confidence = 60 %
2	{E} → {O,K}	= 3/4	75 %
3	{O} → {K,E}	= 3/3	100 %
4	{O,E} → {K}	= 3/3	100 %
5	{K,E} → {O}	= 3/4	75 %
6	{O,K} → {E}	= 3/3	100 %

∴ min-confidence = 80 %

∴ Strong Association rules are :

{O,E} → {K}

{O} → {K,E}

{O,K} → {E}

Strong association rules using FP tree are:

{4} → {k}
{m} → {k}
{o} → {k}
{o} → {e}
{o, e} → {k}
{o} → {k, e}
{o, k} → {e}
{k} → {e}
{e} → {k}

→(2b) The resulting FP ~~tree~~ are similar for both FP tree and the Apriori algorithm. However, in terms of overall algorithm for finding frequent patterns among dataset FP tree algorithm is better than the apriori algorithm as it doesn't require candidate generation thus saving time and space.

Q. 3) $A_1(2,10)$, $A_2(2,5)$, $A_3(8,4)$, $B_1(5,8)$, $B_2(7,5)$, $B_3(6,4)$, $C_1(1,2)$, $C_2(4,9)$

Let x_1, x_2, x_3 represent centroids.

∴ Initial centroids are as follows:

$x_1 : A_1(2,10)$, $x_2 = B_1(5,8)$, $x_3 = C_1(1,2)$

First Iteration

	A1	A2	A3	B1	B2	B3	C1	C2
X1	0	5	8.48	3.61	7.07	7.21	8.06	2.24
X2	3.61	4.24	5	0	3.61	4.12	7.21	1.41
X3	8.06	3.16	7.28	7.21	6.71	5.39	0	7.62

The three clusters are:

cluster 1 = { A1 (2,10) }

cluster 2 = { B1 (5,8), A3 (8,4), B2 (7,5), B3 (6,4), C2 (4,9) }

cluster 3 = { C1 (1,2), A2 (2,5) }

Centroids after first iterations:

X1 = (2,10)

X2 = [(8+5+7+6+4)/5 , (8+4+5+4+9)/5] = (6,6)

X3 = (1.5, 3.5)

Second Iteration

	A1	A2	A3	B1	B2	B3	C1	C2
X1	0	5	8.48	3.61	7.07	7.21	8.06	2.4
X2	5.66	4.12	2.83	2.24	1.41	2	6.4	3.61
X3	6.52	1.58	6.52	5.7	5.7	4.52	1.58	6.04

Centroids after second iteration:

X1 = (3, 4.5)

X2 = (6.5, 5.25)

X3 = (1.5, 3.5)

Third Iteration

	A1	A2	A3	B1	B2	B3	C1	C2
x1	1.12	2.35	7.43	2.5	6.02	6.26	7.76	1.12
x2	6.54	4.51	1.95	3.13	0.56	1.35	6.38	7.68
x3	6.52	1.58	6.52	5.7	5.7	4.52	1.58	6.04

Centroids after 3rd iteration.

$x_1 = (3.67, 9)$

$x_2 = (7, 4.33)$

$x_3 = (1.5, 3.5)$

Fourth Iteration:

	A1	A2	A3	B1	B2	B3	C1	C2
x1	1.95	4.33	6.61	1.66	5.2	5.52	7.49	0.33
x2	6.01	5.04	1.05	4.17	0.67	1.05	6.44	5.55
x3	6.52	1.58	6.52	5.7	5.7	4.52	1.58	6.04

By comparing clusters from 3rd and 4th iteration, we see that both clusters have same points so we can say that K-means reached its stability.

Final Clustere groups are:

Cluster 1 : {A1, B1, C2}

Cluster 2 : {A3, B2, B3}

Cluster 3 : {A2, C1}

Q. 4)

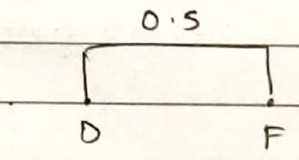
→ Single link :

At level 0, there are 6 clusters
{A}, {B}, {C}, {D}, {E}, {F}

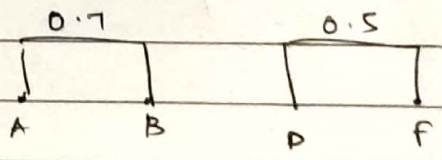
In matrix, item F and D are closest to each other.
i.e. min-dist = 0.5

Distance matrix

Item	A	B	C	D,F	E
A	0				
B	0.7	0			
C	5.66	4.95	0		
D,F	3.2	2.5	2.5	0	
E	4.24	3.54	1.41	1	0



Item	A,B	C	D,F	E
A,B	0			
C	4.95	0		
D,F	2.5	2.5	0	
E	3.54	1.41	1	0



d(D,F,E), (A,B) = 2.5

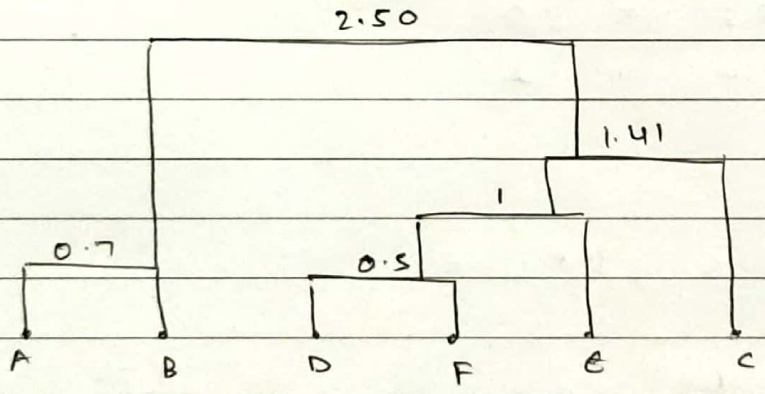
d (D,E,F), (C) = 1.41

Items	A,B	C	D,E,F
A,B	0		
C	4.95	0	
D,E,F	2.5	1.41	0

$d[(D,E,F,C),(A,B)] = 2.5$

Items	A,B	D,E,F,C
AB	0	
(D,E,F	2.5	0

∴ Dendrogram:



(b) Average linkage

At level 0, {A}, {B}, {C}, {D}, {E}, {F},

At level 1, avg: 1

dis (D,F) ≤ 1, merge (D,F)

Item	A	B	C	D,F	E
A	0				
B	0.71	0			
C	5.06	4.95	0		
D,F	3.4	2.71	2.37	0	
E	3.2	2.5	1.41	0.75	0

dist [(D,F), A] = 3.4 , dist[(D,F), C] = 2.37

dist [(D,F), B] = 2.71 , dist [(D,F), E)] = 0.75

dist (A,B) ≤ 1 , merge (A,B)

∴ clusters = {A,B}, {C}, {D,F}, {E}

Items	A,B	C	D,F	E
A,B	0			
C	5.31	0		
D,F	3.06	2.37	0	
E	3.89	1.41	0.75	0

At level 2, Avg = 2

∴ avg (dist (D,F, E) ≤ 2

Merge (DF, E)

Items	A,B	C	D,E,F
A,B	0		
C	5.31	0	
D,E,F	3.335	2.05	0

At level 3, Avg = 3

∴ avg [dist (DEF, C)] ≤ 3

Merge [DEF, C]

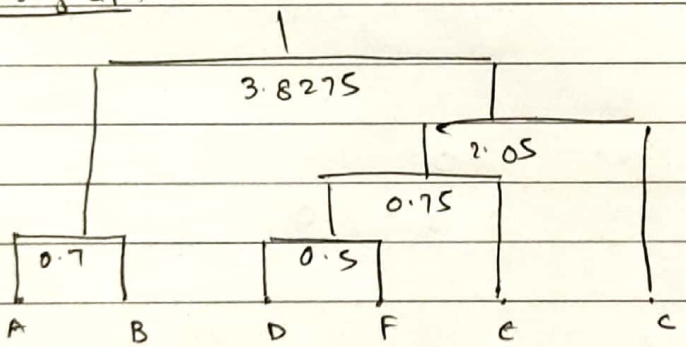
Item	A, B	C, D, E, F
A, B	0	
C, D, E, F	3.8275	0

At level 4, Avg = 4

∴ Avg dist (AB, C, D, EF) ≤ 4

∴ cluster : {A, B, C, D, E, F}

Dendrogram :



→ (Q. 5)

Item	A	B	C	D
A	0			
B	1	0		
C	4	2	0	
D	5	6	3	0

Let Assume two mediods as A, B

mediods : {A, B}

Non-mediods : {C, D}

Replace A → C

$TC_{AC} = 1 + 0 + (-2) + (-2) = -3$

Reduced by +3

Replace A → D

$TC_{AD} = 1 + 0 + 0 + (-5) = -4$

Reduced by +4

Replace B → C

$TC_{BC} = 0 + 1 + (-2) + (-2) = -3$

Reduced by 3

Replace B → D

$TC_{BD} = 0 + 1 + 1 + (-5) = -3$

Reduce by 3

∴ Cost is reduced by replacing D with A

∴ New mediods : {D, B}

Non-mediods : {A, C}

Replace D → A

$TC_{DA} = -1 + 0 + 0 + 5 = 4$

Increase by 4

Replace by D → C

TCBC = 0 + 0 + (-2) + 3 = 1

Increase by 1

Replace B → A

TCBA = (-1) + 1 + 1 + 0 = 1

Increase by 1

Replace B → C

TCBC = 3 + 2 + (-2) + 0 = 3

Increase by 3

∴ None reduced cost, original medoids {B, D} remains same.

Final clusters:

Medoids: {D}, {B}

Non-medoids: {A}, {C}