clustering

Q.1 P1 (0.4, 0.53), P2 (0.22, 0.38), P3 (0.35, 0.32), P4 (0.26, 0.19) P5 (0.08, 041), P6 (0.45, 0.30).

A. Use Single link (MIN) and Complete Link (MAX) technique to group the points into clusters.

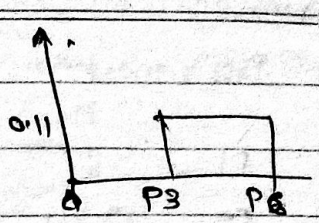
→ Proximity matrix based on Euclidian distance

	P1	P2	P3	P4	P5	P6
P1	0	0.24	0.22	0.37	0.34	0.23
P2	0.24	0	0.15	0.20	0.14	0.25
P3	0.22	0.15	0	0.15	0.28	0.11
P4	0.37	0.20	0.15	0	0.29	0.22
P5	0.34	0.14	0.28	0.29	0	0.39
P6	0.23	0.25	0.11	0.22	0.39	0

Q1 →

(A) Using MIN (Single Link) technique.

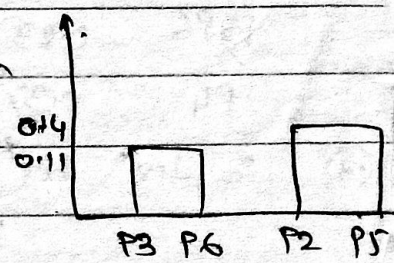
(1) As dist^n bet^n P3 & P6 is smallest select it & merge into first cluster



New dist^n matrix –

	P1	P2	P3/P6	P4	P5
P1	0	0.24	0.22	0.37	0.34
P2	0.24	0	0.15	0.20	0.14
P3/P6	0.22	0.15	0	0.15	0.28
P4	0.37	0.20	0.15	0	0.39
P5	0.34	0.14	0.28	0.39	0

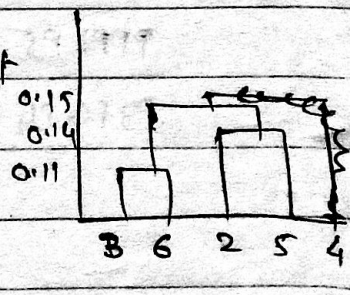
(2) As dist^n bet^n P2 & P5 is smallest, select it & merge into second cluster



New dist^n matrix –

	P1	P2/P5	P3/P6	P4
P1	0	0.24	0.22	0.37
P2/P5	0.24	0	0.15	0.20
P3/P6	0.22	0.15	0	0.15
P4	0.37	0.20	0.15	0

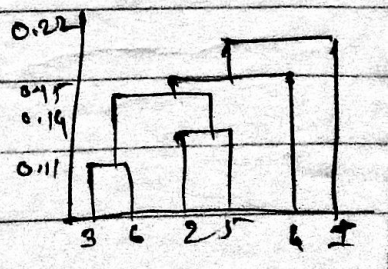
(3) As dist^n bet^n P2/P5 & P3/P6 is smallest, select them to merge into next cluster.



New matrix –

	P1	P2/P5 P3/P6	P4
P1	0	0.22	0.37
P2/P5 P3/P6	0.22	0	15
P4	0.37	0.15	0

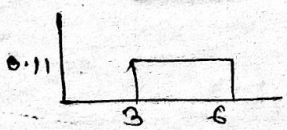
As dist^n bet^n {P2, P5, P3, P6} & {4} is smallest. Select them to form next cluster.



	P1	P2 P5 P3 P6	P4
P1	0	22	
P2, P5, P3, P6, P4	22	0	

(B) Using Complete Link (MAX) Technique

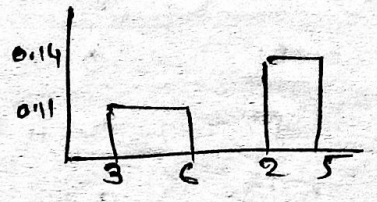
(1) As dist^n bet^n 3 & 6 is smallest merge them



Dist^n matrix.

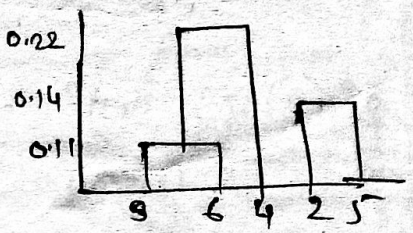
	P1	P2	P3P6	P4	P5
P1	0	0.24	0.23	0.37	0.34
P2	0.24	0	0.25	0.2	0.14
P3P6	0.23	0.25	0	0.22	0.39
P4	0.37	0.2	0.22	0	0.29
P5	0.34	0.14	0.39	0.29	0

(2) As dist^n bet^n P2 & P5 is smallest merge them



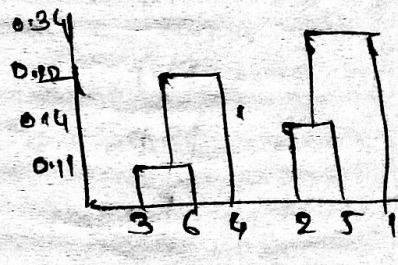
	P1	P2P5	P3P6	P4
P1	0	0.34	0.23	0.37
P2P5	0.34	0	0.39	0.29
P3P6	0.23	0.39	0	0.22
P4	0.37	0.29	0.22	0

(3) As dist^n bet^n P3P6 & P4 is smallest, merge them

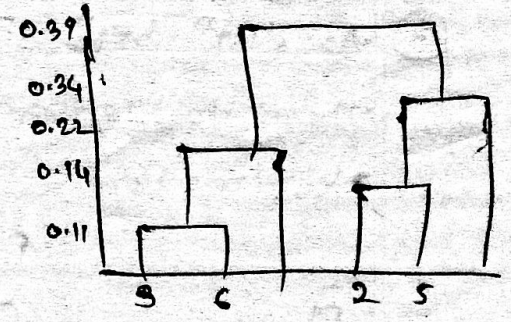


	P1	P2P5	P3P6P4
P1	0	0.34	0.37
P2P5	0.34	0	0.39
P3P6P4	0.37	0.39	0

(4) As dist^n bet^n P1 & P2P5 is smallest merge them



	P1 P2 P5	P3 P6 P4
P1P2P5	0	0.39
P3P6P4	0.39	0



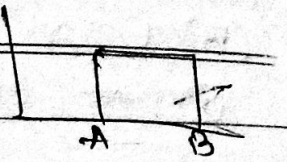
Q.2 (A) Use single link & complete link to cluster the data

	A	B	C	D	E
A	0	1	4	5	7
B	1	0	2	6	8
C	4	2	0	3	4
D	5	6	3	0	4
E	7	8	4	4	0

Q.2(B) Use 2-means to cluster above data points using A, B as initial centroids.

Q.2 (A) Using Single link technique -

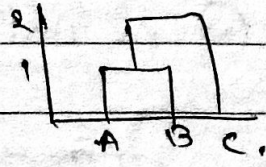
(1) As dist bet^n A&B is smallest, merge them



Dist matrix -

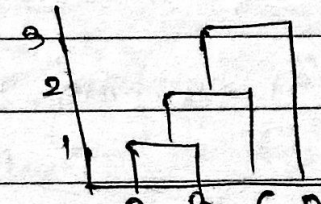
	AB	C	D	E
AB	0	2	5	7
C	2	0	3	4
D	5	3	0	4
E	7	4	4	0

(2) As dist bet^n ABC is smallest merge them



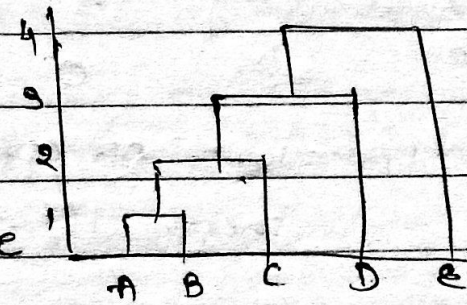
	ABC	D	E
ABC	0	3	4
D	3	0	4
E	4	4	0

(3) As dist bet^n ABC & D is smallest merge th



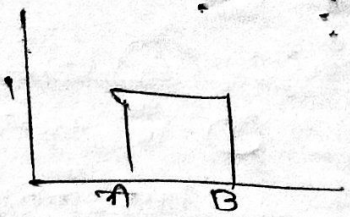
	ABCD	E
ABCD	0	4
E	4	0

(4) Merge the final remaining node with clusters formed.



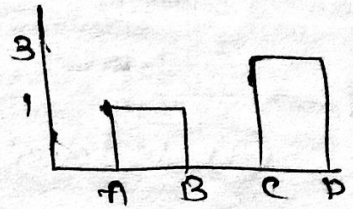
(B) Using MAX (Complete link technique)

(1) As dist bet^n A & B is smallest, merge the.



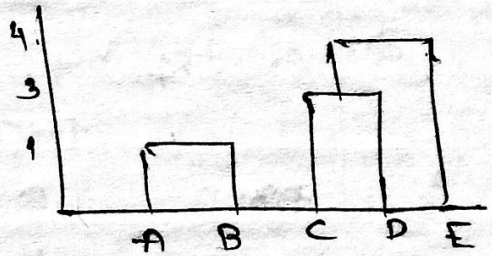
	AB	C	D	E
AB	0	4	6	8
C	4	0	3	4
D	6	3	0	4
E	8	4	4	0

(2) As dist^n bet^n C & D is smallest merge them



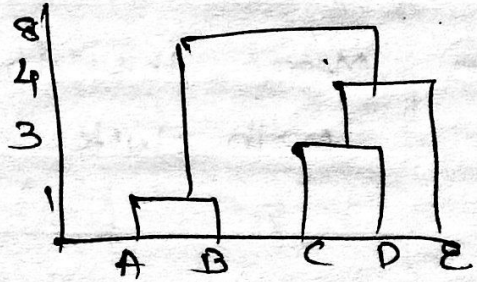
	AB	CD	E
AB	0	6	8
CD	6	0	4
E	8	4	0

(3) As dist^n bet^n E & CD is smallest merge them



	AB	CDE
AB	0	8
CDE	8	0

(4) As dist^n bet^n AB, CDE is smallest merge the

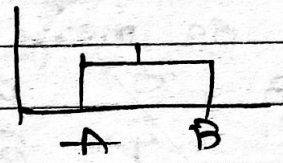


Q.2 Apply Average link algo for clustering

	A	B	C	D	E
A	0	1	4	5	7
B	1	0	2	6	8
C	4	2	0	3	4
D	5	6	3	0	4
E	7	8	4	4	0

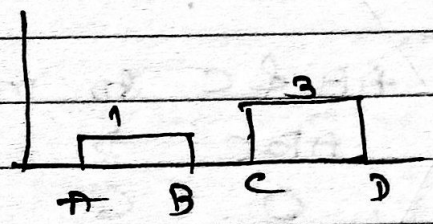
→

Sol.1 Smallest dist is 1 bet' A & B so merge them



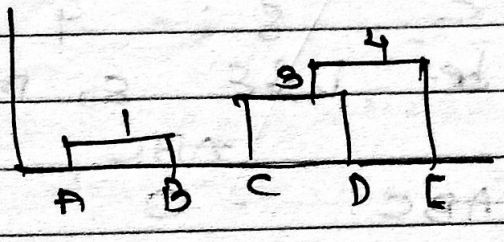
	AB	C	D	E
AB	0	3	5.5	7.5
C	3	0	3	4
D	5.5	3	0	4
E	7.5	4	4	0

Sol.2 Smallest dist is 3 bet' C & D so merge them.



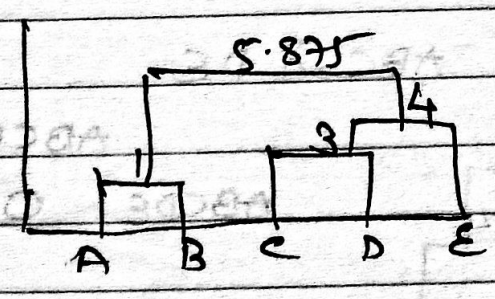
	AB	CD	E
AB	0	4.25	7.5
CD	4.25	0	4
E	7.5	4	0

Sol.3 Smallest dist is 4 bet' CD & E so merge them.



	AB	CDE
AB	0	5.875
CDE	5.875	0

Sol.4 Last step so fuse AB & CDE together



	ABCDE
ABCDE	0

cluster 1 → A, B

cluster 2 → C, D, E.

Use same data for Q4 & use K-means method to cluster the data points. Assume $C_1(0.2, 0.3)$ & $C_2(0.4, 0.4)$ as initial centroids.

K-means Q4.

→ Dist^n of points from $C_1(0.2, 0.3)$.

P1 (0.4, 0.53) → $d = \sqrt{(0.4-0.2)^2 + (0.53-0.3)^2} = 0.304$

P2 (0.22, 0.38) → $d = \sqrt{(0.22-0.2)^2 + (0.38-0.3)^2} = 0.082$

P3 (0.35, 0.32) → $d = \sqrt{(0.35-0.2)^2 + (0.32-0.3)^2} = 0.151$

P4 (0.26, 0.19) → $d = \sqrt{(0.26-0.2)^2 + (0.19-0.3)^2} = 0.125$

P5 (0.08, 0.41) → $d = \sqrt{(0.08-0.2)^2 + (0.41-0.3)^2} = 0.162$

P6 (0.45, 0.30) → $d = \sqrt{(0.45-0.2)^2 + (0.30-0.3)^2} = 0.25$

Dist of points from $C_2(0.4, 0.4)$

P1 (0.4, 0.53) → $d = \sqrt{(0.4-0.4)^2 + (0.53-0.4)^2} = 0.13$

P2 (0.22, 0.38) → $d = \sqrt{(0.22-0.4)^2 + (0.38-0.4)^2} = 0.181$

P3 (0.35, 0.32) → $d = \sqrt{(0.35-0.4)^2 + (0.32-0.4)^2} = 0.094$

P4 (0.26, 0.19) → $d = \sqrt{(0.26-0.4)^2 + (0.19-0.4)^2} = 0.241$

P5 (0.08, 0.41) → $d = \sqrt{(0.08-0.4)^2 + (0.41-0.4)^2} = 0.320$

P6 (0.45, 0.3) → $d = \sqrt{(0.45-0.4)^2 + (0.3-0.4)^2} = 0.111$

Points in cluster 1 with centroid C_1 are P2, P4, P5,

Points in cluster 2 with centroid C_2 are P1, P3, P6.

Need to check whether centroid of cluster changes or not

New centroid of cluster 1: $C_{1\,new_x} = \frac{1}{3}(0.22 + 0.26 + 0.08) = 0.186$

$C_{1\,new_y} = \frac{1}{3}(0.38 + 0.19 + 0.41) = 0.326$

$C_{1\,new} = (0.186, 0.326)$

New centroid for cluster 2: $C_{2new_x} = \frac{1}{3}(0.4 + 0.35 + 0.45) = 0.4$

$C_{2\,new_y} = \frac{1}{3}(0.53 + 0.32 + 0.3) = 0.383$

$C_{2new} = (0.4, 0.303)$

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} dist(C_i, x)^2$$

where dist is standard Euclidian dist^n bet^n two objects.

C_i is centroid of cluster i

$= (0.082)^2 + (0.125)^2 + (0.162)^2 + (0.13)^2 + (0.094)^2 + (0.111)^2 = 0.08$.

Q.2 B

K=2, so need to find out 2 clusters. using 3 means.

dist of points from centroid A →

d(A,B) → dist =

d(A,C) → dist = 4

d(A,E) → dist = 7

d(A,D) → dist = 5

K means Q2

dist of points from centroid B →

d(B,C) = 2

d(B,D) = 6

d(B,E) = 8

from above distances, points in first cluster with centroid

A are → A, E, D.

B are → B, C

New centroid for cluster 1 =

Apply Minimum Spanning Tree (MST) clustering technique to find out 3 clusters in dataset

दि.: / /
पृष्ठ सं.:

	A	B	C	D	E
A	0	1	4	5	7
B	1	0	2	6	8
C	4	2	0	3	4
D	5	6	3	0	4
E	7	8	4	4	0

→ Using Kruskal's algorithm find M.S.T.

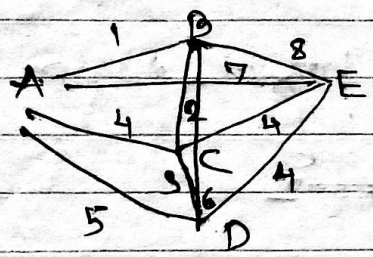
A MST of graph is subgraph that.

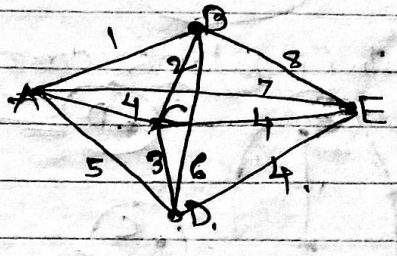
(a) has no cycles i.e. is a tree.

(b) contains all the nodes of graph

(c) has minimum total edge weight of all possible trees.

Step 1 → First construct graph for the points given.





Step 2 → Arrange edges of graphs in increasing order of weight

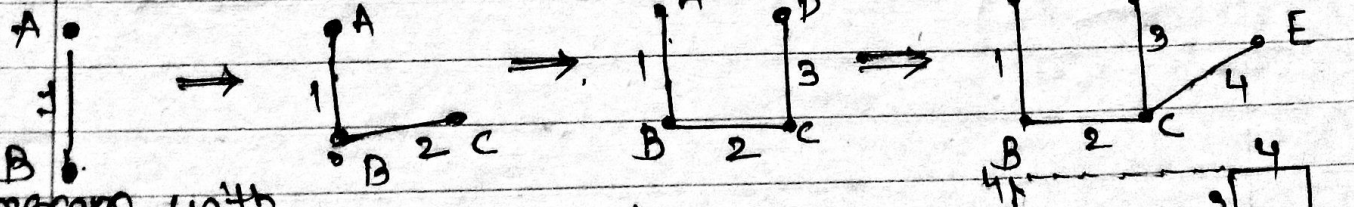
AB - 1 BC - 2 CD - 3 CE, AC, DE - 4 AD - 5 BD - 6 AE - 7 BE - 8.

Step 3 → Apply Kruskal's algorithm to build MST.

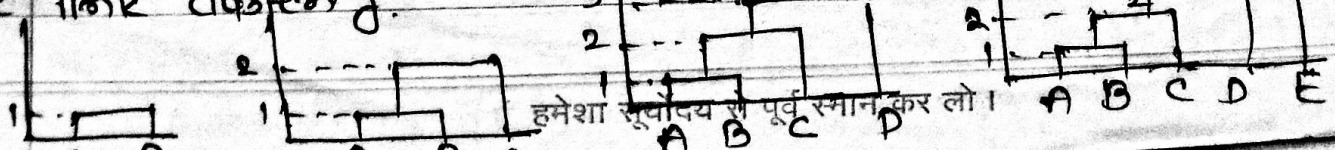
(a) Sort all edges in increasing order of weight.

(b) Pick smallest edge. Check if it forms cycle with S.T. formed so far. If cycle is not formed, include edge. Else discard it.

(c) Repeat #b until there are V-1 edges in tree. where V is number of vertices in graph.



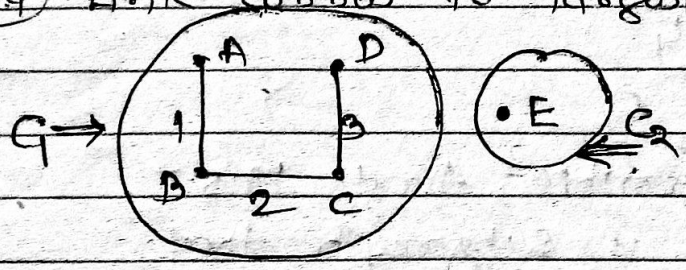
Dendrogram with single link clustering.



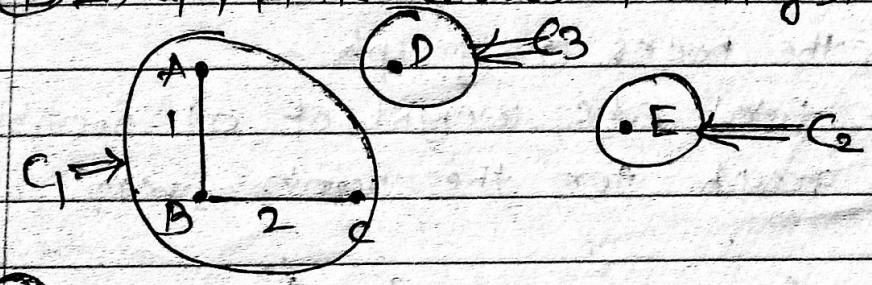
Repeat

Step 4 Create new cluster by breaking the link corresponding to the largest dissimilarity Until only singleton cluster remain

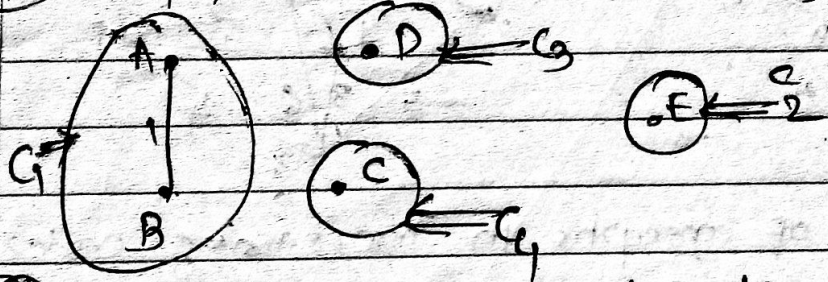
(4a) Link corresp. to largest dissimilarity CE = 4



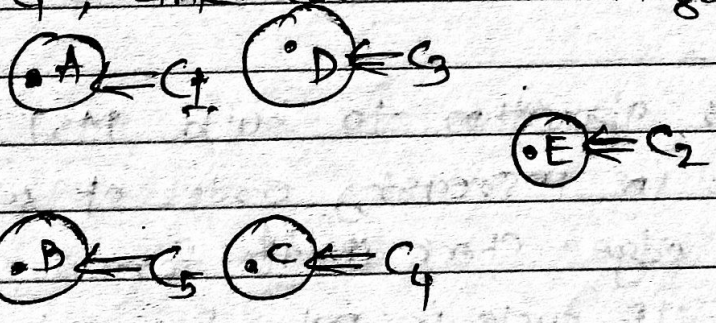
(4b) In C_1, Link corresp. to largest dissimilarity DC = 3



(4c) In C_1, Link corresp. to largest dissimilarity BC = 2



(4d) In C_1, Link corresp. to largest dissimilarity AB = 1.



(4e) All clusters are singletons, so no more splitting req'd

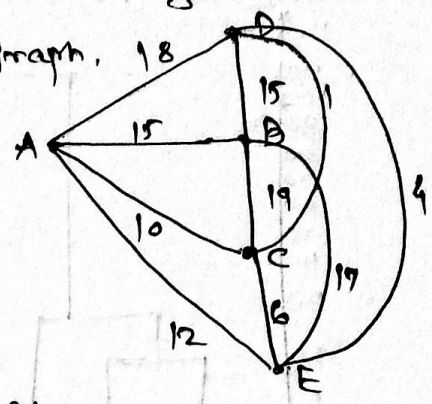
Q. Construct a graph showing all edges of following data, find MST for this graph, Is MST single link hierarchical clustering the same as that found using the traditional of single link algorithm.?

→

Item	A	B	C	D	E
A	0	15	10	18	12
B	15	0	19	15	17
C	10	19	0	1	6
D	18	15	1	0	4
E	12	17	6	4	0

⟹

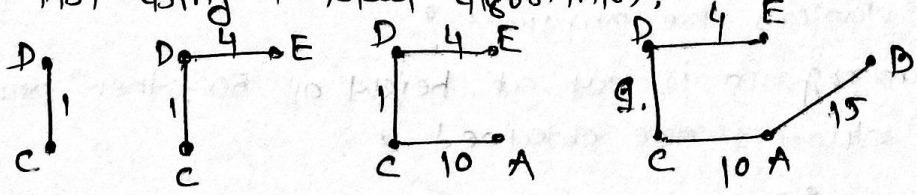
Graph. 18



Edges of graph in ascending order.

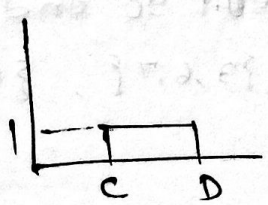
DC=1 DE=4 CE=6 AC=10 AE=12 AB=15 BE=17 AD=18
BC=19

MST using Kruskal algorithm,-

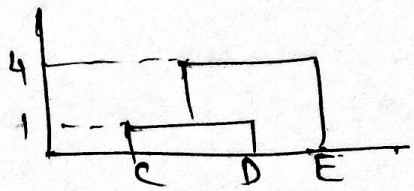


Traditional single link Algorithm -.

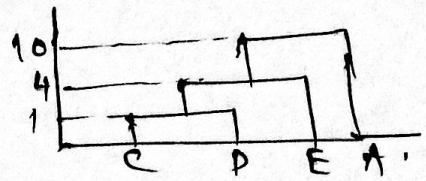
	A	B	CD	E
A	0	15	10	12
B	15	0	15	17
CD	10	15	0	4
E	12	17	4	0



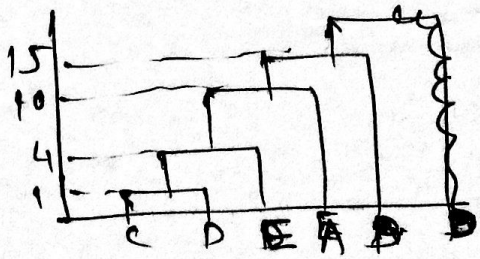
	AB	CDE
A	0	10
B	15	15
CDE	10	15



	CDEA	B
CDEA	0	15
B	15	0

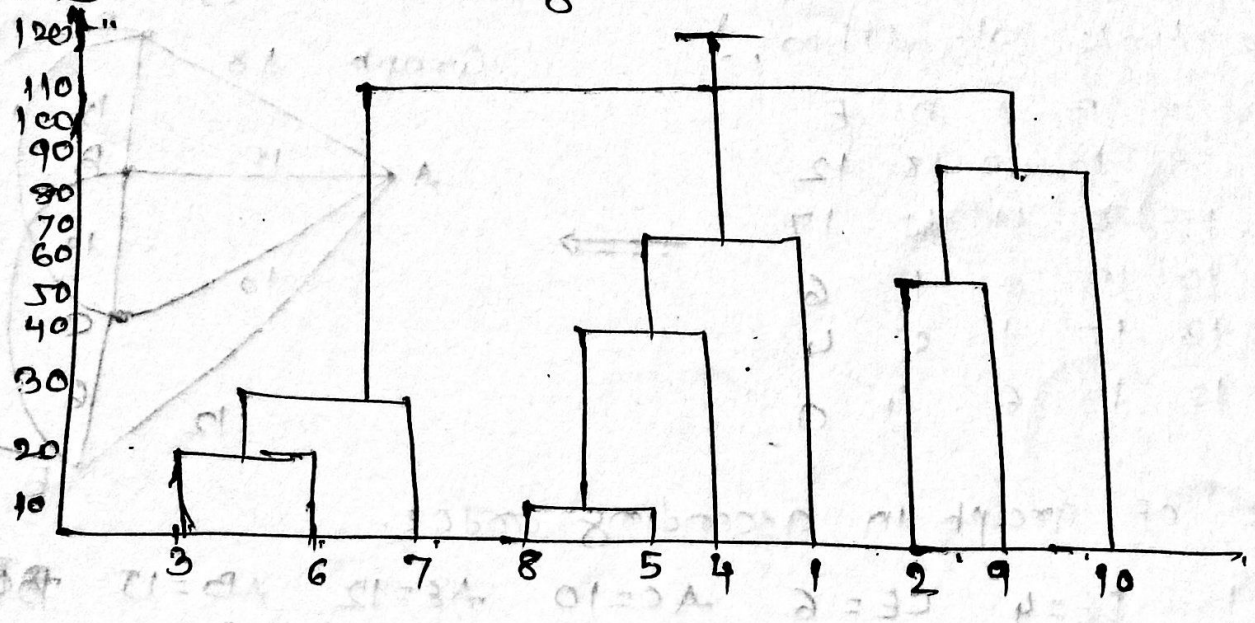


	CDEAB
CDEAB	0



MST single link clustering produces same clusters as traditional single link clustering algorithm as the edges of MST are added in same order as clusters are merged

Q. Consider the dendrogram obtained using hierarchical agglomerative clustering algorithm. Answer the following questions w.r.t. dendrogram.



(a) How many natural clusters are there in dataset.

(b) At what height the dendrogram must be cut, if three clusters are required?

(c) If dendrogram is cut at height of 50, then how many clusters are obtained?

→ (a) 3 $\{3,6,7\}$, $\{8,5,4,1\}$ & $\{2,9,10\}$.

(b) between 90 and 110. excluding both values

(c) 6 $\{3,6,7\}$, $\{8,5,4\}$, $\{1\}$, $\{2\}$, $\{9\}$, $\{10\}$