VIT

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

CAT-II, Winter Semester 2018

Course Code: CSE3019		**Branch**	**: CSE**
Course Title: Data Mining		**Slot**	**: G1**
Duration: 90 Minutes		**Max. Marks**	**: 50**

Answer all questions

1. a) Compute the Minkowski distance (of order 3) and supremum distance for the three data points A= (2,5), B=(1,6), C=(3,4). Provide the answers in matrix format. **6 marks**

b) Compute the cosine distance for the following two data entries.
A= (3, 5, 0, 2, 4) and B= (1, 2, 2, 3, 3) **2 marks**

c) Compute the edit distance for the following two data entries.
A=abce, B=acebd **2 marks**

2. Generate the signature matrix for the following two statements. Use a k=8 size shingling, and the shingling tuples should be sorted in descending order, based on the number of non –zero entries that the tuples have. Use only the top eleven shingling tuples and then use three hashing functions to generate the signature entries for each of the statements. The three hashing functions and the two statements are listed below:

h1(x) = (2x+1) mod 11, **h2**(x)=(3x-1) mod 11, **h3**(x) = (3x+1) mod 1

Statement 1: China–Pakistan Economic Corridor, originally valued at $46 billion, was worth $61 billion last year.
Statement 2: India's trade deficit with China narrowed marginally to $51 billion last year. **10 marks**

3. Explain diagrammatically how data streams can be used to dynamically demarcate low emission zones and car free zones in a metropolitan scenario for maintaining acceptable levels of clean air for the pedestrians. **10 marks**

4. The aim is to group the following points (represented by their x and y coordinates) into proper clusters by agglomerative clustering techniques. Generate two separate hierarchical clusters and their corresponding dendograms, by using two evaluation parameters, single link (minimum) and centroid. Compare the two dendograms and justify your choice for the better value of k (the optimal number of clusters). The points are as follows:
A=(1,1), B=(4,1), C=(5,2), D=(6,2), E=(2,3) **10 marks**

5. Let us suppose huge number of data points (N) need to be clustered, and these points cannot be accommodated inside the main memory. Therefore, apply a linear algorithm (O(N)) which uses a 3 tuple (n, LS, SS) to represent a cluster, having n (the number of points in a cluster), LS (Linear sum of the points), SS (squared sum of the points). The algorithm also uses a height balanced tree to place the clusters as leaf nodes. Ultimately, the entire data points (N) are represented as m clusters inside the main memory. Now, considering the tree has the following properties: B=2 (branching factor of internal nodes), L=3 (branching factor of leaf nodes), T=4 (threshold of diameter of the cluster(s)), construct a clustering tree with the following clusters coming in order (only construction of tree is needed, global clustering and merging up are not required) .

	n	LS	SS
Cluster 1	5	(10, 18)	(50, 80)
Cluster 2	5	(14, 22)	(60,90)
Cluster 3	10	(20, 20)	(90,80)
Cluster 4	5	(30, 30)	(215, 210)
Cluster 5	5	(40, 40)	(370, 400)

Important Note: the threshold T stands for the largest of the diameter values (if the data points are having more than one dimension). Also, the clusters **do not** merge with any other clusters in the leaf node. **10 marks**