

Reg. No.

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

M.E / M.TECH. DEGREE EXAMINATIONS, MAY 2023

Second Semester
CP22204 - BIG DATA ANALYTICS
 (Computer Science and Engineering)
 (Regulation 2022)

TIME: 3 HOURS

MAX. MARKS: 100

COURSE OUTCOMES	STATEMENT	RBT LEVEL
CO 1	Design algorithms by employing Map Reduce technique for solving Big Data problems.	3
CO 2	Design algorithms for Big Data by deciding on the apt Features set.	3
CO 3	Design algorithms for handling petabytes of datasets.	3
CO 4	Design algorithms and propose solutions for Big Data by optimizing main memory consumption.	3
CO 5	Design solutions for problems in Big Data by suggesting appropriate clustering techniques.	3

PART- A (20 x 2 = 40 Marks)
 (Answer all Questions)

Q. No.	QUESTION	CO	RBT LEVEL
1.	Define Bonferroni's principle.	1	1
2.	Define statistical model.	1	1
3.	What are the two approaches to modeling?	1	1
4.	Describe about GFS and HDFS.	1	2
5.	What is k-Shingles?	2	1
6.	Compute the Jaccard Similarity of each pair of the following three sets: {1,2,3,4}, {2,3,5,7}, {2,4,6}.	2	3
7.	Define Hamming Distance.	2	1
8.	How is an Efficient Minhashing achieved?	2	2
9.	What is Bloom Filters?	3	1
10.	What are the different queries used in data-stream management system?	3	1
11.	Define Decaying Windows.	3	1
12.	What is a sampling data in a data stream?	3	1
13.	What is meant by Page Rank?	4	1
14.	What is meant by spider traps?	4	1

Q. No.	QUESTION	Marks	CO	RBT LEVEL
15.	What is the purpose of Apriori Algorithm?	4	1	1
16.	Explain Association rule in mathematical notations.	4	2	2
17.	What is meant by curse of dimensionality?	5	1	1
18.	Differentiate centroids and clusteroids.	5	2	2
19.	Discuss about Collaborative and Content based Recommendation.	5	2	2
20.	Define the CURE algorithm.	5	1	1

PART- B (5 x 10 = 50 Marks)

Q. No.	QUESTION	Marks	CO	RBT LEVEL
21. (a)	Demonstrate any two computational approaches to modeling in detail.	(10)	1	3
	(OR)			
(b) (i)	Illustrate in detail about any two algorithms using MapReduce.	(7)	1	3
(b) (ii)	Apply Reduce Function for those two algorithms in MapReduce?	(3)	1	3
22. (a)	Illustrate about Shingling of Documents with a suitable example.	(10)	2	3
	(OR)			
(b) (i)	Demonstrate Matrix representation of sets with an example.	(4)	2	3
(b) (ii)	Explain Minhashing in detail and how Minhashing is used along with Jaccard Similarity.	(6)	2	3
23. (a)	Illustrate the architecture of data stream management system in detail.	(10)	3	3
	(OR)			
(b)	Explain briefly the methods applied for counting distinct elements.	(10)	3	3
24. (a) (i)	Demonstrate Early Search Engines and Term Spam in detail.	(5)	4	3
(a) (ii)	Illustrate structure of the web with diagram.	(5)	4	3
	(OR)			
(b)	Explain algorithms used for handling larger datasets in main memory.	(10)	4	3
25. (a)	Perform a hierarchial clustering of the one-dimensional set of points 1, 4, 9, 16, 25, 36, 49, 64, 81 assuming the clusters are represented by their centroid (average), and at each step the cluster with the closest centroids are merged.	(10)	5	4
	(OR)			
(b)	Compute the radius, in the sense used by the GRGPF Algorithm (square root	(10)	5	4

of the average square of the distance from the clustroid) for the cluster that is the five points in the lower right of given figure. Note that (11,4) is the clustroid.



PART- C (1 x 10 = 10 Marks)

(Q.No.26 is compulsory)

Marks	CO	RBT LEVEL
-------	----	-----------

26. Cluster the following eight points (with (x, y) representing locations) into three clusters:

(10)	5	5
------	---	---

A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9)

Initial cluster centers are: A1(2, 10), A4(5, 8) and A7(1, 2).

The distance function between two points a = (x1, y1) and b = (x2, y2) is defined as- $P(a, b) = |x2 - x1| + |y2 - y1|$

Use K-Means Algorithm to find the three cluster centers after the second iteration.
