Reg.No.:												

VIVEKANANDHA COLLEGE OF ENGINEERING FOR WOMEN
[AUTONOMOUS INSTITUTION AFFILIATED TO ANNA UNIVERSITY, CHENNAI]
Elayampalayam – 637 205, Tiruchengode, Namakkal Dt., Tamil Nadu.

Question Paper Code: 5031

M.E. / M.Tech. DEGREE END-SEMESTER EXAMINATIONS – JUNE / JULY 2024
Second Semester
Information Technology
P23CSE24 – MINING MASSIVE DATASETS
(Regulation 2023)

Time: Three Hours Maximum: 100 Marks

Answer ALL the questions

Knowledge Levels (KL)	K1 – Remembering	K3 – Applying	K5 - Evaluating
	K2 – Understanding	K4 – Analyzing	K6 - Creating

PART – A

(10 x 2 = 20 Marks)

Q.No.	Questions	Marks	KL	CO
1.	Explicate Index.	2	K1	CO1
2.	Write about distributed file system.	2	K2	CO1
3.	Interpret Minhashing.	2	K2	CO2
4.	Explicate Prefix Indexing.	2	K3	CO2
5.	Mention the function of a working store in a data-stream management system.	2	K2	CO3
6.	Define Page Rank.	2	K1	CO3
7.	Annotate social graph.	2	K2	CO4
8.	Interpret the Betweenness in standard clustering methods.	2	K1	CO4
9.	Construe about long-tail in a recommendation system.	2	K3	CO5
10.	Elucidate the Collaborative-Filtering.	2	K2	CO5

PART – B

(5 x 13 = 65 Marks)

Q.No.			Questions	Marks	KL	CO
11.	a)	i.	Explain Bonferroni's Principle. Give an example to explain the same.	8	K2	CO1
		ii.	Explain how MapReduce computation executes.	5	K3	CO1
			(OR)			
	b)	i.	Explain Matrix-Vector Multiplication by MapReduce.	5	K3	CO1
		ii.	Design MapReduce algorithms to take a very large file of integers and produce as output: a. The largest integer. b. The average of all the integers. c. The same set of integers, but with each integer appearing only once. d. The count of the number of distinct integers in the input.	8	K6	CO1
12.	a)	i.	Compute the Jaccard similarities of each pair of the following three sets: {1, 2, 3, 4}, {2, 3, 5, 7}, and {2, 4, 6}.	5	K2	CO2
		ii.	Using an example, explain the entity resolution.	8	K2	
			(OR)			
	b)	i.	Explain LSH Family for Fingerprint Matching.	6	K2	CO2
		ii.	Explain methods for High Degrees of Similarity.	7	K1	
13.	a)	i.	Explain the ways in which stream data arises naturally. What are two ways that queries get asked about streams?	7	K3	CO3
		ii.	Explain the Bloom Filter. What is the Count-Distinct Problem?	6	K2	
			(OR)			
	b)	i.	How is the efficient computation of PageRank done?	5	K5	CO3
		ii.	What are Biased Random Walks? What must be done to integrate topic-sensitive PageRank into a search engine?	8	K3	
14.	a)	i.	Show how Betweenness can be used to find Communities. Why is it necessary to find a large clique?	7	K3	CO4
		ii.	Why complete bipartite graphs must exist? What Makes a Good Partition?	6	K3	
			(OR)			

	b)	i.	Explain Maximum-Likelihood Estimation. What is a directed graph?	6	K3	CO4
		ii.	Explain the approach to analyzing social-network graphs called Simrank. Why Count Triangles?	7	K2	
15.	a)	i.	What are the issues for Display Ads? What are greedy algorithms?	6	K3	CO5
		ii.	What is the Matching Problem? Explain the Greedy Algorithm for Maximal Matching.	7	K2	
			(OR)			
	b)	i.	What are the applications of Recommendation Systems? What is collaborative filtering?	7	K3	CO5
		ii.	What are the methods for measuring similarity? Explain UV-decomposition as an instance of singular-value decomposition.	6	K2	

PART – C

(1 x 15 = 15 Marks)

Q.No.			Questions	Marks	KL	CO
16.	a)	i.	Suppose we execute the word-count MapReduce program on a large repository such as a copy of the Web. We shall use 100 Map tasks and some number of Reduce tasks. a. Suppose we do not use a combiner at the Map tasks. Do you expect there to be significant skew in the times taken by the various reducers to process their value list? Why or why not? b. If we combine the reducers into a small number of Reduce tasks, say 10 tasks, at random, do you expect the skew to be significant? What if we instead combine the reducers into 10,000 Reduce tasks? c. Suppose we do use a combiner at the 100 Map tasks. Do you expect skew to be significant? Why or why not?	9	K6	CO1
		ii.	Suppose we have a stream of tuples with the schema *Grades(university, courseID, studentID, grade)* Assume universities are unique, but a courseID is unique only within a university (i.e., different universities may have different courses with the same ID, e.g., "CS101") and likewise, studentID's are unique only within a university (different universities may assign the same ID to different students). Suppose we want to answer certain queries approximately from a	6	K5	CO3

1/20th sample of the data. For each of the queries below, indicate how you would construct the sample. That is, tell what the key attributes should be.

a. For each university, estimate the average number of students in a course.
b. Estimate the fraction of students who have a GPA of 3.5 or more.
c. Estimate the fraction of courses where at least half the students got "A."

(OR)

b) 15 K2 CO2

	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*
A	4	5		5	1		3	2
B		3	4	3	1	2	1	
C	2		1	3		4	5	3

Figure above is a utility matrix, representing the ratings, on a 1–5 star scale, of eight items, a through h, by three users A, B, and C. Compute the following from the data of this matrix.

a. Treating the utility matrix as boolean, compute the Jaccard distance between each pair of users.
b. Repeat Part (a), but use the cosine distance.
c. Treat ratings of 3, 4, and 5 as 1 and 1, 2, and blank as 0. Compute the Jaccard distance between each pair of users.
d. Repeat Part (c), but use the cosine distance.
e. Normalize the matrix by subtracting from each nonblank entry the average value for its user.
f. Using the normalized matrix from Part (e), compute the cosine distance between each pair of users.