

THE KAVERY ENGINEERING COLLEGE
(An Autonomous Institution)

M.E DEGREE END SEMESTER THEORY EXAMINATIONS, NOV /DEC 2024

Second Semester

Computer Science and Engineering

BD4251 - Big Data Mining and Analytics

Regulations – 2021

Duration : 3 Hrs

Maximum : 100 Marks

PART – A (ANSWER ALL QUESTIONS) (Marks 10*2=20)

Q.No	Questions	BLT	CO	Marks
1.	Outline the significance of statistical modeling in data analysis.	UN	1	2
2.	What is the role of cluster computing in handling large datasets?	RE	1	2
3.	Define document shingling.	RE	2	2
4.	Why locality-sensitive functions are important?	UN	2	2
5.	Define filtering in the context of data streams.	RE	3	2
6.	What is the advantage of using a decaying window in dynamic data environments?	RE	3	2
7.	Write a short note on market basket model in data mining.	RE	4	2
8.	Brief on the importance of counting frequent item sets in data mining.	RE	4	2
9.	Differentiate K-means clustering and hierarchical clustering algorithms.	UN	5	2
10.	List down real-world applications of clustering in advertising.	RE	5	2

PART – B (ANSWER ALL QUESTIONS) (Marks 5*13 = 65)

Q.No	Questions	BLT	CO	Marks
11.	a Explain the various computational approaches used in machine learning for modeling data. Discuss the advantages and challenges of each approach with examples.	UN	1	13
	Or			
	b Describe the MapReduce framework and its importance in processing big data. Outline the steps involved in a MapReduce job and provide an example of its application.	UN	1	13
12.	a Explain how locality-sensitive hashing optimizes search and similarity operations in high-dimensional spaces.	UN	2	13
	Or			
	b Discuss in detail the methods for high-degree similarity detection in multimedia data using LSH.	UN	2	13
13.	a Develop a method for counting distinct elements in a high-volume stream using approximate counting techniques.	AP	3	13
	Or			
	b Apply the algorithm to count the number of ones in a sliding window over a binary data stream.	AP	3	13

14.	a	Apply the PageRank algorithm to a small web graph of five nodes and calculate the PageRank of each node after three iterations.	AP	4	13
		Or			
	b	Apply the A-priori algorithm to find frequent item sets in a sample dataset of grocery transactions.	AP	4	13
15.	a	Demonstrate the steps involved in implementing the CURE clustering algorithm for a large, non-euclidean dataset.	UN	5	13
		Or			
	b	Illustrate how parallelism could enhance the efficiency of clustering in large datasets for an online recommendation system.	UN	5	13

*PART - C (Marks 1*15 = 15)*

<i>Q.No</i>		<i>Questions</i>	<i>BLT</i>	<i>CO</i>	<i>Marks</i>
16.	a i	Analyze the trade-offs between precision and memory efficiency when using the limited-pass algorithm for large datasets.	AN	4	8
	ii	Examine the market basket model's application in e-commerce.	AN	4	7
		Or			
	b	Examine how hierarchical clustering and CURE differ in handling outliers within a recommendation system for web-based advertisements.	AN	5	15