Reg. No.																

Question Paper Code	13841

B.E. / B.Tech. - DEGREE EXAMINATIONS, NOV / DEC 2025

Seventh Semester

Artificial Intelligence and Data Science

20AIEL707 - MINING MASSIVE DATASETS

Regulations - 2020

Duration: 3 Hours   Max. Marks: 100

PART - A (MCQ) (10 × 1 = 10 Marks)

Answer ALL Questions

		Marks	*K – Level*	*CO*
1.	Which of the following best describes the purpose of hash functions in data mining? (a) To reduce data dimensionality by clustering (b) To enable quick data retrieval and detect duplicates (c) To encrypt data for security (d) To visualize large datasets	*1*	*K1*	*CO1*
2.	Given a large set of documents, which MapReduce step is responsible for aggregating word counts across all documents? (a) Mapper (b) Combiner (c) Reducer (d) Partitioner	*1*	*K1*	*CO1*
3.	Why is Jaccard similarity preferred when comparing sets of shingles from two documents? (a) It calculates the cosine between two document vectors (b) It measures the proportion of common elements to total unique elements (c) It considers the frequency of each shingle in the documents (d) It calculates the Euclidean distance between shingle sets	*1*	*K1*	*CO2*
4.	Which of the following statements best describes the goal of similarity-preserving summaries like MinHash? (a) To compress documents without any data loss (b) To compute exact edit distances between documents (c) To create compact representations that approximate Jaccard similarity (d) To cluster documents based on keyword frequency	*1*	*K1*	*CO2*
5.	Which of the following is a method for estimating the number of distinct elements in a stream? (a) CURE (b) HyperLogLog (c) HITS (d) FP-Growth	*1*	*K1*	*CO3*
6.	Topic-sensitive PageRank differs from traditional PageRank by: (a) Ignoring link structures (b) Using user-specific teleport sets (c) Ranking only ads (d) Using a batch process	*1*	*K1*	*CO3*
7.	In clustering, which distance measure is typically used in Euclidean spaces? (a) Jaccard Distance (b) Cosine Similarity (c) Euclidean Distance (d) Manhattan Similarity	*1*	*K1*	*CO4*
8.	Hierarchical clustering can be of two types: (a) K-Means and CURE (b) Agglomerative and Divisive (c) Frequent and Infrequent (d) Euclidean and Non-Euclidean	*1*	*K1*	*CO4*
9.	Which similarity measure is most suitable for text document clustering? (a) Euclidean distance (b) Cosine similarity (c) Manhattan distance (d) Jaccard distance	*1*	*K1*	*CO5*
10.	Which classification algorithm is often used because it assumes feature independence? (a) Decision Trees (b) Naïve Bayes (c) SVM (d) Logistic Regression	*1*	*K1*	*CO6*

PART - B (12 × 2 = 24 Marks)

Answer ALL Questions

11.	A dataset is processed using MapReduce. The intermediate data output by mappers is extremely large. Suggest a technique to reduce communication overhead.	*2*	*K2*	*CO1*

12.		What is the role of a distributed file system in the MapReduce framework?	*2*	*K1*	*CO1*
13.		How shingling helps in detecting near-duplicate documents?	*2*	*K1*	*CO2*
14.		Given two sets A = {a, b, c, d} and B = {b, c, d, e}, compute the Jaccard similarity.	*2*	*K2*	*CO2*
15.		List the two techniques to prevent link spam.	*2*	*K1*	*CO3*
16.		What is a decaying window in stream mining?	*2*	*K1*	*CO3*
17.		State any two advantages of hierarchical clustering.	*2*	*K1*	*CO4*
18.		What is meant by limited-pass algorithms?	*2*	*K1*	*CO4*
19.		How does CURE deal with non-spherical clusters?	*2*	*K1*	*CO5*
20.		Define the objective function minimized by K-Means clustering.	*2*	*K1*	*CO5*
21.		How can word embeddings improve content-based recommendations?	*2*	*K1*	*CO6*
22.		What role do user-generated tags play in representing item features?	*2*	*K1*	*CO6*

PART - C (6 × 11 = 66 Marks)

Answer ALL Questions

23.	a)	List and briefly describe the different components of the MapReduce framework.	*11*	*K2*	*CO1*
		OR			
	b)	Explain the working of a distributed file system. How does it contribute to scalability and fault tolerance in big data systems?	*11*	*K2*	*CO1*
24.	a)	Discuss the communication cost model in distributed computing. Why is it important in designing MapReduce algorithms?	*11*	*K2*	*CO2*
		OR			
	b)	Explain the statistical limitations of data mining. How do overfitting and underfitting affect model performance? Provide suitable illustrations.	*11*	*K2*	*CO2*
25.	a)	Explain the concept of the Stream Data Model. Why is it important for processing large-scale data?	*11*	*K2*	*CO3*
		OR			
	b)	Explain the Page Rank algorithm and the significance of the teleportation factor.	*11*	*K2*	*CO3*
26.	a)	Explain the Apriori principle in frequent itemset mining. Why does it reduce the search space?	*11*	*K2*	*CO4*
		OR			
	b)	Describe how clustering algorithms can be adapted for data streams. Illustrate with an example of the STREAM algorithm.	*11*	*K2*	*CO4*
27.	a)	How does CluStream handle evolving data streams? Explain the role of micro-clusters and macro-clusters.	*11*	*K2*	*CO5*
		OR			
	b)	Discuss the role of parallelism in clustering algorithms. Illustrate with a MapReduce K-Means example.	*11*	*K2*	*CO5*
28.	a)	Apply how the item profiles are represented as vectors? Illustrate with an example from e-commerce products.	*11*	*K3*	*CO6*
		OR			
	b)	Choose With an example and show how a user profile is updated when new feedback (clicks, likes, purchases) arrives.	*11*	*K3*	*CO6*

K1 – Remember; K2 – Understand; K3 – Apply; K4 – Analyze; K5 – Evaluate; K6 – Create