

**PSG COLLEGE OF ARTS & SCIENCE
(AUTONOMOUS)**

**BSc DEGREE EXAMINATION MAY 2026
(Sixth Semester)**

Branch – COMPUTER SCIENCE WITH DATA ANALYTICS

MINING OF MASSIVE DATA

Time: Three Hours

Maximum: 75 Marks

SECTION-A (10 Marks)

Answer ALL questions

ALL questions carry EQUAL marks

(10 × 1 = 10)

Module No.	Question No.	Question	K Level	CO
1	1	Choose the term that describes a model of a problem where we look for rare events in huge amounts of data. (i) Clustering (ii) Bonferroni's Principle (iii) Hashing (iv) PageRank	K1	CO1
	2	why the "Communication Cost" is a primary bottleneck in MapReduce algorithms. (i) High CPU usage (ii) Disk storage limits (iii) Data movement between nodes (iv) RAM capacity	K2	CO1
2	3	Find the Jaccard Similarity of sets $A=\{1,2,3\}$ and $B=\{2,3,4\}$. (i) 0.2 (ii) 0.5 (iii) 0.75 (iv) 1.0	K1	CO2
	4	Compare shingling and minhashing in terms of their role in document similarity (i) Both are the same (ii) Shingling creates sets; Minhashing creates signatures (iii) Shingling ranks pages (iv) Minhashing compresses images	K2	CO2
3	5	Label the technique used to select a subset of a stream such that each element has an equal probability of being chosen. (i) Bloom Filtering (ii) Reservoir Sampling (iii) DGIM (iv) PageRank	K1	CO3
	6	Interpret the result of a "False Positive" in a Bloom Filter membership test. (i) Item is definitely in the set (ii) Item is definitely not in the set (iii) Item might be in the set (iv) The filter is broken	K2	CO3
4	7	Match the "Support Count" definition to its correct usage in market basket analysis. (i) Number of clusters (ii) Frequency of an itemset (iii) Number of web links (iv) Page loading time	K1	CO4
	8	Contrast the "Point Assignment" vs. "Hierarchical" clustering approaches. (i) One uses labels; the other doesn't builds a tree (ii) One is for streams; the other for files (iii) One assigns to centroids; the other (iv) They are identical	K2	CO4
5	9	Name the vector that corresponds to the steady-state distribution of a random surfer in PageRank. (i) Zero vector (ii) Identity vector (iii) Principal Eigenvector (iv) Null vector	K1	CO5
	10	Describe the main advantage of CUR decomposition over Singular Value Decomposition (SVD). (i) Higher accuracy (ii) Uses actual rows and columns of the matrix (iii) Faster for small data (iv) Requires no memory	K2	CO5

Cont...

SECTION - B (35 Marks)

Answer ALL questions

ALL questions carry EQUAL Marks

(5 × 7 = 35)

Module No.	Question No.	Question	K Level	CO
1	11.a.	Outline the various components of the HDFS (Hadoop Distributed File System)	K1	CO1
		(OR)		
	11.b.	Relate the Map task and Reduce task in a word count algorithm.		
2	12.a.	Summarize the theory of Locality-Sensitive Hashing (LSH) for finding similar pairs.	K2	CO2
		(OR)		
	12.b.	Rephrase the distance measures for non-Euclidean spaces like Cosine and Edit distance		
3	13.a.	Utilize the DGIM algorithm to estimate the number of 1's in a binary stream window.	K3	CO3
		(OR)		
	13.b.	Solve a problem to determine the probability of a false positive in a Bloom Filter.		
4	14.a.	Analyze the PCY algorithm and how it improves upon the Apriori algorithm using buckets.	K4	CO4
		(OR)		
	14.b.	Dissect the BFR algorithm for clustering massive datasets in Euclidean space.		
5	15.a.	Make use of the power iteration method to compute PageRank for a small web graph.	K3	CO5
		(OR)		
	15.b.	Apply Singular Value Decomposition (SVD) to reduce the dimensionality of a matrix.		

SECTION -C (30 Marks)

Answer ANY THREE questions

ALL questions carry EQUAL Marks

(3 × 10 = 30)

Module No.	Question No.	Question	K Level	CO
1	16	Evaluate the efficiency of MapReduce for performing a matrix-matrix multiplication versus a standard approach.	K5	CO1
2	17	Discuss the construction of minhash signatures and prove why they preserve Jaccard similarity.	K6	CO2
3	18	Organize a stream processing pipeline that handles decaying windows to find the most frequent recent elements.	K3	CO3
4	19	Examine the "Market Basket" model and investigate its application in modern recommendation engines	K4	CO4
5	20	Determine the impact of "Taxation" (damping factor) on solving the spider trap problem in PageRank.	K5	CO5