PSG COLLEGE OF ARTS & SCIENCE
(AUTONOMOUS)

BSc DEGREE EXAMINATION MAY 2025
(Sixth Semester)

Branch – **COMPUTER SCIENCE**

MAJOR ELECTIVE COURSE – II: MINING OF MASSIVE DATA

Time: Three Hours Maximum: 50 Marks

SECTION-A (5 Marks)
Answer **ALL** questions
ALL questions carry **EQUAL** marks (5 x 1 = 5)

1. What is the primary purpose of Hash Functions in Data Mining?
   (i) Encrypt data (ii) Compress data
   (iii) Index and search data efficiently (iv)) Generate random numbers
2. Which method is used for finding similarity between large sets of items?
   (i) K-Means Clustering (ii) Locality Sensitive Hashing (LSH)
   (iii) Decision Trees (iv) Linear Regression
3. Specify the primary challenge in data stream mining?
   (i) Processing infinite amounts of data in limited memory
   (ii) Sorting small datasets
   (iii) Storing data in relational databases
   (iv) Visualizing data in real-time
4. The clustering technique which is suitable for non-Euclidean spaces?
   (i) K-Means Clustering (ii) Hierarchical Clustering
   (iii) CURE Algorithm (iv) Decision Trees
5. The technique is which commonly used for dimensionality reduction?
   (i) Principal Component Analysis (PCA) (i) Naïve Bayes
   (i) Random Forest (i) Gradient Boosting

SECTION - B (15 Marks)
Answer **ALL** Questions
ALL Questions Carry **EQUAL** Marks (5 x 3 = 15)

6 a Suppose there is a repository of ten million documents. What (to the nearest integer) is the IDF for a word that appears in (i) 40 documents (ii) 10,000 documents?

OR

b Let us consider there is a repository of ten million documents, and word w appears in 320 of them. In a particular document d, the maximum number of occurrences of a word is 15. Approximately what is the TF.IDF score for w if that word appears (i) once (ii) five times?

7 a Evaluate the S-curve $1 - (1 - s^r)^b$ for $s = 0.1, 0.2, ..., 0.9$, for the following values of r and b: • $r = 3$ and $b = 10$

OR

b On the space of nonnegative integers, which of the following functions are distance measures? If so, prove it; if not, prove that it fails to satisfy one or more of the axioms. (i) max(x, y) = the larger of x and y. (ii) diff(x, y) = $|x - y|$ (the absolute magnitude of the difference between x and y). (iii) sum(x, y) = $x + y$

8 a Describe counting distinct elements in a stream.

OR

b Explain Decaying windows.

9 a A database has five transactions. Let min sup = 60% and min conf = 80%.

Transaction id	items
T100	{N,P,O,L,F,Z}
T200	{E, P, O, L, F, Z}
T300	{N, B, L, F}
T400	{N, V, D, L, F, Z}
T500	{D, P, L, J, F}

(a) Find all frequent item sets using Apriori algorithm.
(b) List all the strong association rules.

Cont...

OR

9 b Form 3 clusters using K means algorithm for the following data set A1(4,6), A2(2,5), A3(9,3), A4(5,5), A5(10,4), A6(11,3), A7(10,10), A8(2,2). Consider A1, A2, A3 as the seeds of the 3 clusters.

10 a Discuss the concept of Cur Decomposition.

OR

b Elucidate the concept of Clustering of social network graphs.

SECTION -C (30 Marks)

Answer **ALL** questions

ALL questions carry **EQUAL** Marks (5 x 6 = 30)

11 a A company wants to analyze its web server logs to determine the most frequently occurring HTTP response codes (e.g., 200, 404, 500). The logs are stored in large files distributed across multiple machines.

Each log entry has the following format: Analyze logs using Map Reduce.

```
<IP_Address> - - [Timestamp] "GET /index.html HTTP/1.1" <Response_Code> <Bytes_Sent>
```

```
192.168.1.1 - - [10/Mar/2025:10:00:01] "GET /home.html HTTP/1.1" 200 1024
192.168.1.2 - - [10/Mar/2025:10:00:02] "GET /about.html HTTP/1.1" 404 512
192.168.1.3 - - [10/Mar/2025:10:00:03] "GET /index.html HTTP/1.1" 200 2048
192.168.1.4 - - [10/Mar/2025:10:00:04] "GET /contact.html HTTP/1.1" 500 256
```

OR

b Suppose hash-keys are drawn from the population of all nonnegative integers that are multiples of some constant c, and hash function h(x) is x mod 15. For what values of c will h be a suitable hash function, i.e., a large random choice of hash-keys will be divided roughly equally into buckets?

12 a Find the Jaccard distances between the following pairs of sets: (i) {1, 2, 3, 4} and {2, 3, 4, 5}. (ii) {1, 2, 3} and {4, 5, 6}.

OR

b Infer the methods for high degrees of similarities in mining of massive datasets.

13 a What is Page Rank and summarize the efficient computation of Page Rank.

OR

b Analyze sampling data in a stream.

14 a Here is a collection of twelve baskets. Each contains three of the six items 1 through 6. {1, 2, 3} {2, 3, 4} {3, 4, 5} {4, 5, 6} {1, 3, 5} {2, 4, 6} {1, 3, 4} {2, 4, 5} {3, 5, 6} {1, 2, 4} {2, 3, 5} {3, 4, 6} Suppose the support threshold is 4. On the first pass of the PCY Algorithm we use a hash table with 11 buckets, and the set {i, j} is hashed to bucket i × j mod 11.

(i) By any method, compute the support for each item and each pair of items.
(ii) Which pairs hash to which buckets?
(iii)Which buckets are frequent?
(iv) Which pairs are counted on the second pass of the PCY Algorithm?

OR

b There are eight items, A, B, . . . , H, and the following are the maximal frequent itemsets: {A, B}, {B, C}, {A, C}, {A, D}, {E}, and {F}. Find the negative border.

15 a Let M be the matrix of data points[(1, 1),(2, 4),(3, 9),(4, 16)]

(i) What are M^TM and MM^T?
(ii) Compute the eigenpairs for M^TM.
(iii) What do you expect to be the eigenvalues of MM^T?
(iv) Find the eigenvectors of MM^T, using your eigenvalues from part (iii).

OR

b With the help of an example explain Singular Value Decomposition.

Z-Z-Z

END