PSG COLLEGE OF ARTS & SCIENCE
(AUTONOMOUS)

BSc DEGREE EXAMINATION MAY 2022
(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 x 1 = 10)

1 Which theorem of statistics gives a statistically sound way to avoid most of these bogus positive responses to a search through the data?
(i) Bonferroni correction (ii) Alon-Matias principle
(iii) Datar-Gionis similarities (iv)Matias-Szegedy correction

2 The key-value pairs from each Map task are collected by a _____ and sorted by Key.
(i) chunks (ii) master node
(iii) cloud store (iv)master controller

3 _____ is the continuous flow of data generated by various sources.
(i) database (ii) dataflow
(iii) data stream (iv)data sequence

4 Find out the formula to compute the sum of the stream for exponentially decaying window.
(i) $\sum_{i=0}^{t-1} a_{t-i}(1-c)^i$ (ii) $\frac{2^{j-1}-1}{1+(r-1)(2^j-1)}$
(iii) $\sum_{i=1}^{n}(2a(i)-1)$ (iv) $(\sum_{i=1}^{n}|x_i-y_i|^r)^{1/r}$

5 _____ is a function that assigns a real number to each page in the Web.
(i) Link Spam (ii) PageRank (iii) TrustRank (iv)Spam Mass

6 Which algorithms make their decision in response to each input element by maximizing some function of the input element?
(i) greedy algorithm (ii) balance algorithm
(iii) Stream-Clustering Algorithm (iv)Toivonen's Algorithm

7 _____ focus on properties of items.
(i) Product-based systems (ii) Content-Based systems
(iii) Collaborative-Filtering systems (iv)All the above

8 _____ is a technique for taking a dataset consisting of a set of tuples representing points in a high-dimensional space and finding the directions along which the tuples line up best.
(i) Eigenvectors (ii) Singular-Value Decomposition
(iii) CUR Decomposition (iv)Principal-component analysis

9 A _____ is a website that allows people with similar interests to come together and share information, photos and videos.
(i) Simrank (ii) Link Spam (iii) social network (iv)chunks

10 SCC stands for _____.
(i) Strongly Connected Component (ii) Strongly Connected Communities
(iii) Social Connected Communities (iv)Smart Connected Component

Cont...

SECTION - B (25 Marks)

Answer **ALL** questions

ALL questions carry **EQUAL** Marks (5 x 5 = 25)

11 a Bring out the most important directions of modeling in data mining.

OR

b Describe in detail the workflow system and its functions.

12 a With neat sketch, express the Data-Stream-Management System and list the sources of data stream.

OR

b State about the Bloom filter technique and show the analysis of Bloom filtering.

13 a Outline the motivation and usage of Topic-Sensitive PageRank.

OR

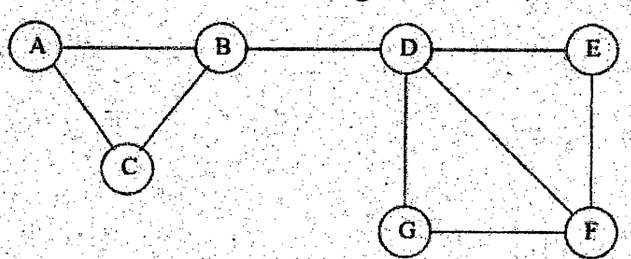
b Explain the balance algorithm and show how it works among many bidders.

14 a Narrate the steps to build a complete UV-Decomposition algorithm of dimensionality reduction.

OR

b Show the Eigenvalues and Eigenvectors of Symmetric Matrices with example.

15 a Construct the adjacency matrix and the degree matrix for the following directed graph.



(OR)

b Write notes on Simrank in detail.

SECTION -C (40 Marks)

Answer **ALL** questions

ALL questions carry **EQUAL** Marks (5 x 8 = 40)

16 a Summarize the different types of Algorithms using Map Reduce.

OR

b Enumerate the Complexity Theory for Map Reduce

17 a Discuss about the counting of distinct elements in a stream.

OR

b Point out the Datar-Gionis-Indyk-Motwani (DGIM) algorithm and how to maintain the DGIM conditions.

18 a Elucidate the architecture and analysis of a spam farm.

OR

b Enumerate the Ad-words Implementation to process a document using matching algorithm with hash table technique.

19 a Discuss about a model for recommendation systems.

OR

b Give an overview of singular-value decomposition.

20 a Explain in detail the Direct Discovery of Communities.

OR

b Summarize the Neighborhood Properties of Graphs.

Z-Z-Z END