PG/CSE/PE/T/115B/2025

MASTER OF COMPUTER SC. & ENGG. EXAMINATION, 2025

(1st Year, 1st Semester)

BIG DATA ANALYTICS

Time : Three Hours Full Marks : 100

Answer All Questions

Special credit will be given to brief and to-the-point answers

1. (i) What is meant by Data Munging?

2

(ii) Explain Locality Sensitive Hashing.

4

(iii) Explain Bonferroni's principle with an example.

4

(iv) Give three examples of applications of Outlier Detection.

2

(v) Explain how Mahalanobis Distance is defined. What are its assumptions?

4

(vi) What do you mean by Analytics? What are the types of Analytics?

4

2. Answer any two:

(a) Explain the architecture of Hadoop Distributed File System. What is the critical component in the HDFS for achieving speedup? Why?

6+1+3

(b) Explain the mechanism of Map-Reduce Programming Framework.
Show in detail, how you will find the Natural Join of two relations R(A,B) and S(B,C) using M-R technique.

6+4

(c) What do you mean by Euclidean and Non-Euclidean Space?
Define Cosine Distance and Edit Distance. What are their application domains?

4+4+2

3. Answer any two:

(a) Explain the difference between Content based Recommendation and Collaborative Filtering.
Explain the role of Distance functions in these methods?

6+4

(b) State and explain the Monotonicity property of Frequent Item Sets.
Explain in detail, how this property is used in the A-priori algorithm for finding out the frequent item sets. What are the challenges of its implementation in the Map-Reduce paradigm?

2+6+2

[Turn over

(c) What is meant by an Outlier? What are the challenges in the Outlier detection in Large Data Sets?
Explain the AVF algorithm for Outlier detection.

1+2+7

4. Answer any two:
(a) What are the factors affecting the efficiency of M-R algorithms?
How are the failures of Map or Reduce functions managed in the M-R framework?

6+4

(b) What do you mean by Page Rank? How would you avoid Spider Traps while computing Page Rank?
What is Topic Sensitive Page Rank? How do you propose to compute Topic Sensitive Page Ranks?

2+3+2+3

(c) Give the rationale of K-means Clustering.
Show how Map-Reduce paradigm can be used to implement K-means Clustering for a massive set of data points.

4+6

5. Answer any two:
(a) One million text files are to be analyzed to come up with a manageable set of candidate files for detailed matching to find out similar files. Explain the broad steps to get an acceptable solution.

10

(b) Formulate the Adwords problem in the context of Online Web Adverftising.
Briefly explain the rationale of the BALANCE Algorithm.

5+5

(c) Explain the terms Confidence and Interest in the context of an Association Rule.
Considering the present technology trends and data generation trends, in your opinion, which are the directions of growth of Big Data Analytics?

4+6

----------X----------