Ex/CSE/T/424C/2019

B. E. COMPUTER SCIENCE & ENGINEERING EXAMINATION, 2019

(Fourth Year, Second Semester)

BIG DATA ANALYTICS

Time : Three Hours | Full Marks : 100

Answer *question no. 1* and any *four* from the rest
Special credit will be given to brief and to-the-point answers

1. (i) What are the characteristics of Big Data? Explain.

4

(ii) Explain what you mean by Data Munging.

4

(iii) Explain Cosine distance. Prove that it is indeed a distance function.

5

(iv) Show with a diagram, how a block of data is written into a file stored in the Hadoop Distributed File System.

4

(v) What do you mean by Analytics?

3

2. What is meant by an Outlier? What are the challenges in the Outlier detection in Large Data Sets?

Explain how you can detect outlier by the use of Clustering techniques.

Explain the AVF algorithm for Outlier detection. What are its use cases?

How can you implement the AVF algorithm in the Map-Reduce framework?

2+2+4+7+2+3

3. Explain in detail how the Map-Reduce Programming paradigm works.

What do you mean by Inverted Index of a text file? Explain how you can find the inverted index of a very large text file using Map Reduce paradigm.

How can you determine the complexity of a Map Reduce Algorithm?

8 + 2 + 6 + 4

[Turn over

4. Explain how a search engine provides the search query results by using Page Rank? How can the naive Page Rank methodology be fooled? Explain with examples.

What is Topic Sensitive Page Rank? How do you propose to compute Topic Sensitive Page Ranks?

Explain the utility of Hubs and Authorities?

$6 + 4 + 2 + 4 + 4$

5. How are the web advertisements different from newspaper advertisements?

Develop a model for the Adwords problem.

Explain how the clickstreams can be processed for immediate response to user queries.

$4+8+8$

6. An online store recommends movies when a user searches for books. What could be the flaw in the implementation of the recommendation system of the store?

Explain clearly the difference between Content-based recommendation system and Collaborative Filtering.

Explain in detail, how a system for recommending websites to web-surfers can be designed.

$4 + 6 + 10$

7. Explain how Association Rules are formulated from a Frequent Item Set?
What do you mean by Confidence and Interest of an Association Rule? What are their significances?
What is the most memory-consuming part of finding Frequent Item Sets? Explain with reference to practical situations.

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?

$3 + 3 + 2 + 3 + 5 + 4$

8. Answer any four from the following: $4 \text{ X } 5 = 20$

(i) Briefly explain how the Law Enforcement Agencies can identify suspicious groups using Call Data Records of Mobile Phones?

(ii) What are the differences between Data Lake and Data Warehouse?

(iii) What is a Bloom Filter? Find out the optimum number of Hash Functions required to assure a particular rate of False Positives.
(iv) What do you mean by Euclidean Space and Non-Euclidean Space?
(v) Considering the present technology trends and data generation trends, in your opinion, which are the directions of growth of Big Data Analytics?

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