Ex/PG/CSE/T/129F/2018

M. E. COMPUTER SCIENCE & ENGINEERING EXAMINATION, 2018
(First 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.

3

(ii) Define the Cosine distance $d(x, y)$ between two vectors x and y; show that the function d defined by you conforms to properties of distance functions.

5

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

4

(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? Explain how Analytics can be useful in Recommendation System for Book Industry.

2+2

2. Explain how the Search Engines show the top few web pages relevant to a query.

Explain two methods by which a search engine can be fooled by a web designer. How can a Search Engine overcome these problems.

Define Trustrank. How is Trustrank determined? What is it used for?

What do you mean by Hubs and Authorities? What are their uses?

3 + 4 + 4 + 5 + 4

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

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.

8 + 6 + 6

4. In an online store, a customer faces the following two situations:
i) When (s)he tries to purchase a book of poems by Shankha Ghosh, the system responds "You may also try the following books of poems by Shankha, Sunil and Shakti".

ii) When (s)he tries to purchase Diaper for the kid, the system responds "Those who have purchased Diaper have also purchased Beer".
Explain in detail how the Recommendation System of the online store works in the above two situations.

What is the Long-tail phenomenon in Retail Marketing? Explain how this can be handled by All-digital retailers.

What is the complexity of Collaborative Filtering? State how to manage the complexity.

You are advising an online store for recommendation system. Explain to them the pros and cons of Collaborative Filtering with huge data and simple algorithm.

4 + 4 + 4 + 4 + 4

5. 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

6. Explain how Web Advertising works from the point of view of Advertisers and Search Engine Service Providers. In this context explain the terms Click-Through-Ratio and Advertisers' Bid.

Hence Model the Adwords Problem. Explain the rationale of the BALANCE Algorithm for solving the Adwords Problem.

3 + 4 + 5 + 8

7. There are 100,000 webpages with an average size of 4 KB each. After 5-shingling, it was found that the total number unique shingles are 10,000. Explain the procedure of forming the Input Matrix using shingling. Calculate the percent reduction of the size of data to be handled by the above case of shingling for similarity detection.

Explain how the Input Matrix may be further compressed by using Minhashing without compromising the similarity properties of the files.

If there are 50 hash functions used in the Minhash algorithm, what is the percent reduction of the data to store the Sketches of the file?

6 + 4 + 7 + 3

8. Answer the following: 4 X 5 = 20

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

(ii) State and explain Bonferroni's principle.

(iii) Briefly explain the Attribute-Value-Frequency (AVF) Algorithm for finding out the outliers.

(iv) Considering the present technology trends and data generation trends, in your opinion, which are the directions of growth of Big Data Analytics?

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