PG/CSE/T/129F/2019

MASTER OF COMPUTER SC. & ENGG. EXAMINATION, 2019

(1st Year, 2nd 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 is meant by Data Scrubbing? Why is it required?

3

(ii) Explain what you mean by Edit Distance. Where can this kind of distance measure be used?

4

(iii) Explain the terms Variety and Veracity in the context of Big Data Systems.

4

(iv) What do you mean by Page Rank?

3

(v) What is the importance of Hash Algorithms in Big Data Processing?

3

(vi) Explain two major outcomes of Analytics?

3

2. Explain the architecture of Hadoop Distributed File System. What is the critical component in the HDFS for achieving speedup?

Explain how Fault Tolerance is achieved in HDFS.

Explain how the Map-Reduce programming paradigm can be mapped to the HDFS.

Explain how computational complexity of the Map-Reduce algorithms can be computed?

4+2+4+6+4

3. You are given a voluminous book and asked to prepare an index to be appended at the end of the book. Explain the approach you will take to get the job done. Hence develop an M-R algorithm for computing the index for a book.

Explain in detail, how two huge matrices can be multiplied using M-R algorithm, where the matrices do not fit into the memory of the data nodes. Give examples of some applications where huge matrix multiplication can be useful.

4+6+7+3

4. What are the two distinct methods used in Recommendation Systems? Explain their use cases.

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

Sometimes, there are some anomalous recommendations put forward to the customers. Explain why this may happen.

6+4+3+3+4

5. 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. Which kind of Datasets can be processed by the AVF algorithm?

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

What are the sources of speedup in the M-R implementation of AVF algorithm?

2+2+4+2+7+3

6. Management of a Retail Store wanted to organize the products it sells in a manner that the customers are tempted to purchase more of the displayed products. What strategies can the Management take? How can they derive their strategies based on data analysis? Explain.

In this context, discuss the use of Frequent Item Sets and their Monotonicity property.

Explain how the PCY algorithm improves the A-priori algorithm. What are the Multi-stage and Multi-hash extensions of the PCY algorithm? What are the benefits?

3+3+4+6+4

7. You are asked to download a huge number of webpages using a standard web crawler program. The task is to find similar webpages from the downloaded pages. Explain how you can proceed step by step to achieve your goal.

After finding sets of similar webpages, you are asked to rank the webpages so that their links can be displayed in an acceptable sequence through a Search Engine. Explain how you would rank the webpages.

12 + 8

8. Answer the following:

(i) What is the difference between Data Warehouse and Data Lake?

(ii) Discuss the importance of Visualization in Big Data Analytics.

(iii) Briefly explain how police records of crime data including location, time, date, nature of crime, etc. can be analyzed to help in predictive policing to reduce the number of crimes?

(iv) Massive Open Online Courses (MOOC) are taken by a huge number of students across the country. Discuss how Big Data Analytics can be used for the management and improvement of quality of the courses.

4+5+5+6