PG/CSE/T/129F/2017

MASTER OF COMPUTER SC. & ENGG. EXAMINATION, 2017

(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 Munging? 3

(ii) Explain Locality Sensitive Hashing. 4

(iii) Explain the concepts of the CURE algorithm. 5

(iv) Give three examples of applications of Outlier Detection. 3

(v) What are the major components of Apache STORM? 3

(vi) What do you mean by Analytics? 2

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

Detail out how Replication Management is done in HDFS.

Show with figures, how the data reads and writes are executed in the File System.

What has been done to avoid single point failure of the File System?

4+3+3+6+4

3. 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.

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+8+3+3

4. What do you mean by Euclidean and Non-Euclidean Space?

Define Cosine Distance and Edit Distance. What are their application domains?

Explain the difference between Content based Recommendation and Collaborative Filtering.

Explain the role of Distance functions in these methods?

2+4+10+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.

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+9+3

6. 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?

What are Hubs and Authorities? What are their utilities?

2+6+3+6+3

7. How do you define a Frequent Item Set?

Explain how random samples can be used to find out Frequent Item Sets from a large number of baskets.

How does SON algorithm reduce False Positives and False Negatives?

Show how SON algorithm can be implemented in M-R framework?

What is the utility of finding Frequent Item Sets?

2+4+3+9+2

8. Answer the following:

(i) Explain Bonferroni's Principle with an example.

(ii) Define Mahalanobis Distance. What are the assumptions for this distance function? Prove that this is indeed a distance function.

(iii) Briefly explain how Citation Analysis can be modeled as a Graph Mining Problem. What are the major issues in Large Graph Analysis?

(iv) What is a Bloom Filter? Find out the optimum number of Hash Functions required to assure a particular rate of False Positives.

3+6+5+6

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