Islamic University of Technology

Organisation of Islamic Cooperation (OIC)

Department of Computer Science and Engineering (CSE)

SEMESTER FINAL EXAMINATION SUMMER SEMESTER, 2018-2019

DURATION: 3 Hours FULL MARKS: 150

CSE 6279: Big Data Analysis and Management

Programmable calculators are not allowed. Do not write anything on the question paper. **Answer all questions.** Figures in the right margin indicate marks.

1. (a) What is big data analytics? According to Bill Franks big data can be seen as different from traditional data sources in a number of ways. Briefly explain them. [10]

   (b) Churn prediction can be enhanced using big data analytics. Explain it. [5]

   (c) Explain the matrix-vector multiplication by MapReduce. Also describe the technique to cope with the situation when the vector v does not fit in main memory. [10]

2. (a) Explain the hazards related to Total Information Awareness (TIA) in the context of big data. [5]

   (b) Explain base line of the Bonferroni's Principle (BP) to avoid "bogus" false positive. Consider the following scenario: [5+10]

   Objective: To detect "evil doers" We assume that such people periodically gather at a hotel to plot.

   Assumptions:

   - There are 30 million people who might be evil doers.
   - Everyone goes to a hotel one day out of 40 days.
   - A hotel's capacity is 30 persons.
   - Total observation period is 300 days.

   Your task is to apply the BP to test if this approach to detecting evil doers is feasible.

   (c) What is the basic application area of TF.IDF? Explain its formal measure with suitable example. [5]

3. (a) What is shingle? Comment on its size. Is the traditional shingle capable to identify similar news article on the web? Justify your position. [10]

   (b) What is Minhash Signature? Consider the following tables for input matrix and random permutation matrix. Your task is to construct the Minhash Signature for each column and show its correctness (approximate). [15]

Table 1: Input Matrix

C-1	C-2	C-3	C-4
1	0	1	0
1	0	0	1
0	1	0	0
0	0	0	1
0	1	0	1
1	1	1	0
1	0	1	0

Table 2: Random Permutation Matrix

P1	P2	P3	P4
3	7	2	1
4	5	1	6
7	4	6	2
6	2	3	4
1	3	5	3
2	6	4	7
5	1	7	5

4. (a) Briefly discuss how early search engines worked. Also explain how people fooled this naive technique. Finally mention two innovations by Google to combat "spam". [10]

(b) What is PageRank? Briefly explain three basic principles to assess the importance of a page used in PageRank. [10]

(c) Suppose a network has 4 nodes with sufficient connectivity. Apply the PageRank algorithm to write its 4 equations with 4 unknowns. How will you solve it manually? Is this method applicable for a real network with millions of nodes? Justify. [5]

5. (a) Explain the concept of Power Iteration using a suitable example. [10]

(b) There are two problems in Power Iteration for PageRank such as spider-trap and dead-end. Briefly discuss them and outline how PageRank eliminates them. [10]

(c) Discuss the three essential properties of a social network. Explain the term "betweeness" with a suitable example. [5]

6. (a) Briefly describe the long tail phenomenon in the context of recommendation system. [10]

(b) What is content-based recommendation? Suggest some properties for item profiling (i.e. audio CD or mp3). [5]

(c) Explain the concept of UV-Decomposition with suitable example. [10]