# Islamic University of Technology

Organisation of Islamic Cooperation (OIC)

## Department of Computer Science and Engineering (CSE)

**MID SEMESTER EXAMINATION** **SUMMER SEMESTER, 2018-2019**
**DURATION: 1 Hour 30 Minutes** **FULL MARKS: 75**

## 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) The term "Big Data" is a misnomer. Explain. Although there are a number of different ways to define big data, IBM uses three parameters for big data. Briefly explain each of them with relevant reference data and example. [5]

   (b) *Big Data introduces a paradigm shift in terms of analytic focus. We are moving from descriptive analytics to predictive and prescriptive analytics.* Explain this concept. [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) Mining massive dataset is often considered as the discovery of "models" from different aspects such as statistics and machine learning. In particular special care must be taken if machine learning is the correct choice for building models. Briefly strengthen the argument. [5]

   (b) Explain 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 50 million people who might be evil doers.
   - Everyone goes to a hotel one day out of 50 days.
   - A hotel's capacity is 50 persons.
   - Total observation period is 500 days.

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

   (c) What is power law? Prove that it is essentially a transformation from exponential to linear relationship. Mention few applications that exhibit power law. [5]

3. (a) Define Jaccard Similarity. This measure can not capture the preference of the users. Justify with example. Propose an alternative measure to address it and comment on its upper limit. Finally show your previous example for this new measure. [5]

   (b) What is shingle? Comment on its size. Is the traditional shingle capable to identify similar news article on the web? Justify your position. [5]

(c) What is Minhash Signature? Why is it used for large number of document similarity? 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). [5+10]

Table 1: Input Matrix

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

(d) What is Locality-Sensitive Hashing (LSH) for documents? *"It addresses both memory and computational time"*- Justify with a suitable example. [5]