45

ISLAMIC UNIVERSITY OF TECHNOLOGY (IUT)

ORGANISATION OF ISLAMIC COOPERATION (OIC)

Department of Computer Science and Engineering (CSE)

MID SEMESTER EXAMINATION **SUMMER SEMESTER, 2021-2022**

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 3 (three)** questions. Figures in the right margin indicate full marks of the corresponding question.

---

1. a) Definition of big data by Gartner Inc. highlights a number of benefits for more precise data analytic. Briefly outline them. 5

   b) There is a paradigm shift in terms of focus in big data which transforms the traditional descriptive analytics to predictive and prescriptive analytics. Explain it using suitable example. 5

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

   Objective: To detect "evil doers", we hold the following assumptions:

   - There are 150 million people who might be evil doers.
   - Everyone goes to a hotel one day out of 200 days.
   - A hotel's capacity is 250 persons.
   - Total observation period is 350 days.
   - As a pattern for an evil doer we consider: "for a given hotel, 2 persons visit the hotel on 2 different dates for a common purpose".

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

2. a) What is shingle? Suppose you are working to find the text similarity for (i) official email and (ii) book chapter (that may be upto 50 pages long). How do you select the size of shingle for these cases? Justify it. 5

   b) The maximum limits of Jaccard Similarity and Jaccard Similarity of Bags are identical. Justify your position. 5

   c) Consider the following Boolean matrix for 4 (d1,d2,d3,d4) documents: 15

Table 1: Boolean Matrix for Q 2.c)

d1	d2	d3	d4
1	0	1	1
1	0	0	0
1	0	0	1
0	1	0	0
1	0	1	1
0	1	0	0

Use **any three random permutations** to create the Minhash Signature. Show each step for the signature formation. Verify its correctness.

3. a) Term-frequency vectors are typically very long and sparse. Can we apply Euclidean Distance measure in such case directly? Justify your position using a suitable example. 5

b) Consider the following Boolean Matrix (against 4 Elements/Shingles). Show each step to generate Minhash Signature based on the following two Random Hash Functions: 15

- $h1 = (x + 1)\%5$
- $h2 = (4x + 1)\%5$

Finally verify its correctness against Jaccard Similarity.

Table 2: Boolean Matrix for Q. 3.b)

Elements	d1	d2	d3	d4
a	1	1	0	1
b	1	1	0	0
c	1	1	0	0
d	0	1	1	0

c) Briefly explain the motivation of banding technique for Locality-Sensitive Hashing (LSH) for documents. State a comprehensive discussion on the probabilistic analysis of the technique. 10