Program: B. Sc. Engg. in Computer Science and Engineering **Date: 12 September, 2021**
Semester: Winter 2020-2021 **Time: 2:30 PM – 4:00 PM**

ISLAMIC UNIVERSITY OF TECHNOLOGY (IUT)
ORGANISATION OF ISLAMIC COOPERATION (OIC)
Department of Computer Science and Engineering (CSE)

Semester Final Examination **Winter Semester: 2020-2021**
Course Number: CSE 4739 **Full Marks: 75**
Course Title: Data Mining **Duration: 1 Hour 30 Minutes**

There are **3 (THREE)** questions. Answer **ALL** of them. The symbols have their usual meanings. The examination is **Online** and **Closed Book**. Programmable calculators are not allowed. Marks of each question and corresponding CO and PO are written in the brackets.

---

1. a) Consider the following transaction database of Table 1. (15) [CO3, PO2]

Table 1: A transaction database

Transaction ID	List of items in the transaction
T1	B, A, T
T2	A, C
T3	A, S
T4	B, A, C
T5	B, S
T6	A, S
T7	B, S
T8	B, A, S, T
T9	B, A, S

Mine the frequent patterns including conditional pattern base and conditional FP-Tree using Frequent Pattern Growth (PF-Growth) algorithm. Assume that the minimum support count is 2.

b) Briefly explain the general architecture of an Information Retrieval (IR) system. (10) [CO3, CO4, PO2, PO3]

2. a) For the distance matrix (Table 2), perform the iterations of agglomerative clustering (single linkage) and draw the corresponding Dendogram. (10) [CO4, CO5, PO3, PO5]

Table 2: Distance Matrix

	A	B	C	D
A	0	1	4	5
B		0	2	6
C			0	3
D				0

b) Bag of Words (BoW) and TF-IDF are the simplest form of text representation in numbers. Support you have the following exam reviews given by 2 (two) students. (8) [CO3, CO4, PO9, PO5]

Student 1: This question is very hard and lengthy
Student 2: This question is standard and good

Create a TF-IDF representation of the reviewers' comments.

c) Consider the exam review of the two students in the previous question 2 (b), calculate a similarity distance score between those two comments using Jaccard similarity distance. (7) [CO4, CO5, PO5, PO6]

3. a) Consider the following hyperlink graph shown in Figure 1. Assume that the web surfer will click the hyperlinks in a page uniformly random.

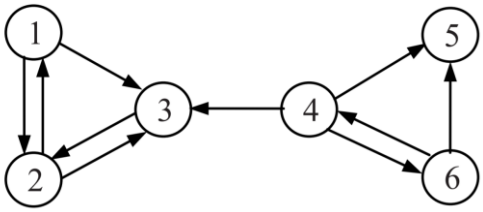

Figure 1: A hyperlink graph for web structure mining

Answer the followings:

i. Write the transition probability matrix from the graph. (3) [CO4, CO5, PO7, PO8]

ii. Is the matrix a stochastic matrix? Explain. (3) [CO4, CO5, PO7, PO8]

b) Find the rank of the web pages in the web graph of Figure 1 using PageRank Algorithm. Show up to 2 iterations. (12) [CO4, CO5, PO9, PO9]

c) With the use of Hyper link Induced Topic Search (HITS) algorithm, explain how hub and authority scores are used for web page ranking. (7) [CO4, CO5, PO9, PO9]