SARDAR VALLABHBHAI NATIONAL INSTITUTE OF TECHNOLOGY, SURAT
DEPARTMENT OF COMPUTER SCIENCE & ENGG.
B. Tech-3rd Year (6th Sem)
Data Science (CS321)
End-Sem Exam

Date: 20/04/2024 Time: 09:00 AM To 12:00 PM Marks 50

Instructions: 1. Write your admission number/roll no. and other details clearly on the answer books and questions paper. 2.Support your answer with diagrams & examples wherever required.

Q. 1 -	Answer the following:	**Marks 15**
1.	Find TF-IDF for the following document contents' word. Document 1: rain today maximum. Document 2: maximum today maximum Document 3: rain tomorrow average Document 4: maximum tomorrow average	[5] [CO3]
2.	Describe a Data Stream management system with the help of a diagram and applications.	[5] [CO3]
3.	Which are the three factors that will affect the probability of getting false positives in Bloom filter? Explain the relationship between these factors and the probability of getting false positives; whether it is directly or inversely proportional or no relationship.	[5] [CO4]
Q. 2 -	Answer the following:	**Marks 20**
1.	Calculate the number of distinct elements in the following datastream using Flajolet-Martin algorithm: 2, 3, 1, 1, 2, 5, 2, 1 and given a hash function h(x) = (x + 2) mod 64	[5] [CO4]
2.	Compute the minhash signature for each column and calculate the similarity of signature of each column to other columns using Jaccard similarity; i.e. similarity between C1-C3, C2-C4, C1-C2, C3-C4 You need to consider 3 permutations of rows of a matrix as follows: P1: 3 4 7 6 1 2 5 P2: 4 2 1 3 6 7 5 P3: 1 3 7 6 2 5 4	[5] [CO4]

Rows	C_1	C_2	C_3	C_4
1	1	0	1	0
2	1	0	0	1
3	0	1	0	1
4	0	1	0	1
5	0	1	0	1
6	1	0	1	0
7	1	0	1	0

$C_1-C_2 = 0$
$C_1-C_3 = 3/4$
$C_1-C_4 = 1/4$
$C_2-C_3 = 0$
$C_2-C_4 = 3/4$
$C_3-C_4 = 0$

	C_1-C_3	C_2-C_4	C_1-C_2	C_3-C_4
5. P1	3/4 ?	3/4	0	0
P2	2/3	3/2	0	0

3.	Consider the frequent itemset mining problem based on the following transactions table of the supermarket. Find out items that has minimum support of 2. Also find out the association rules that has minimum confidence of 50 %. Solve the problem using a-priori algorithm that is based on candidate generation.	[10] [CO4]

Transactions	Items
1	Milk, Bread, Butter
2	Bread, Cheese
3	Bread, Jam
4	Milk, Bread, Cheese
5	Milk, Jam
6	Bread, Jam
7	Milk, Jam
8	Milk, Bread, Jam, Butter
9	Milk, Bread, Jam

Q. 3 -	Answer the following:	**Marks 15**
1.	You are developing a software for an online supermarket, and your goal is to recommend new items to users, but you don't want to recommend items that they already bought. Each item has a unique long integer id, but the supermarket has so many items (2^{64} items) that you cannot store in memory. Which of the method/algorithm is best suited for testing if the recommended item is unseen by the user? Also explain why?	[05] [CO5]
2.	Following is a utility matrix, representing the ratings, on a 1–5 star scale, of eight items, a through h, by three persons P_1, P_2, and P_3. Compute the following from the data of this matrix. Some values are unknown which need to be ignored.	[10] [CO5]

	a	b	c	d	e	f	g	h
P_1	4			5	1			
P_2	5	5	4					
P_3				2	4	5		

a) Consider the rating of 3, 4 and 5 as "1" and rating of 1 and 2 as unrated (unkown), compute the Jaccard distance between each pair of persons.

b) Calculate similarity between each pair of person using the cosine distance. Consider normalizing values of an original matrix before calculating similarity using cosine distance.

c) Comment on the similarity of persons based on the answers of (a) and (b); which persons are similar or different?