Nirma University

Institute of Technology

Semester End Examination (IR), December - 2019

B. Tech. in Information Technology, Semester-VII

IT702 Information Retrieval Systems

Roll / Exam No.		Supervisor's Initial with Date	

Time: 3 Hours | Max Marks :100

Instructions: 1. Attempt all questions.
2. Figure to the right indicate full marks.
3. Section wise separate answer book to be used.
4. Draw neat sketches wherever necessary.
5. Assume necessary data wherever required, and indicate clearly.

SECTION – I

Q.1 CO1_BL4 Represent following corpus in matrix form using TF-IDF representation. **[14]**

D1: "Exam ends today, results awaited."

D2: "Admission season will start after exam results."

D3: "Enquiry counter is available for admission."

Show every step in pre-processing clearly. For the query "admission enquiry", report the ranking order of retrieved documents using cosine similarity.

Q.2 **Answer the following:** **[16]**

[A] CO2_BL3 Can a neural network without any hidden layer implement XOR operation on binary variables? If yes, discuss the process in detail. If no, justify your answer and provide the solution discuss it in detail. [10]

[B] CO1_BL4 Critically compare Boolean model with vector space model for document representation using an appropriate example. [6]

OR

[B] CO1_BL2 Describe the concept and strategies used to obtain relevance feedback in context of Information Retrieval systems. [6]

Q.3 **Answer the following:** **[20]**

[A] CO3_BL4 Assume the following user-item rating matrix. [12]

	I1	I2	I3	I4
U1	4	2	5	5
U2	4	2	1	
U3	3		2	4
U4	4	4		
U5	2	1	3	5

Please note that the empty cell in the matrix denotes that the item is yet not rated. Use user-based collaborative filtering to estimate the user U1's rating for the item I2. Use Pearson correlation to calculate similarity between users. Use the following formula to compute the rating

$$P_{a,i} = \bar{r}_a + \frac{\sum_{u \in U} (r_{u,i} - \bar{r}_u) \cdot w_{a,u}}{\sum_{u \in U} |w_{a,u}|},$$

$P_{a,i}$ is the rating of user a for the item i. $\bar{r}_a$ is the average rating of user a. U is the set of 3 users most similar to user a and who have rated the item i. $w_{a,u}$ is the similarity between users a and u.

[B] CO2_BL1 Following documents are represented by Term Frequency (TF) weight vector. Apply k means algorithm on these documents to partition them into two clusters. Run the algorithm for two iterations. [8]

D1 <0,1,1,0,3,1,0,2>
D2 <1,0,0,1,1,1,0,0>
D3 <2,0,0,1,1,0,0,1>
D4 <1,1,0,0,0,0,1,0>
D5 <0,0,3,0,1,2,0,1>

OR

[B] CO2_BL1 Discuss about the long tail phenomenon which is typically observed in Retail & Marketing scenario. [8]

SECTION – II

Q.4 Do as directed. **[20]**

[A] CO3_BL2 Can information retrieval be useful in medical domain? Justify your answer. [6]

[B] CO2_BL4 Consider a meta search system with five underlying search system, which have ranked four candidate documents or pages a, b, c, and d as follows: [14]

System 1: a, b, c, d
System 2: b, a, d, c
System 3: c, b, a, d
System 4: c, b, d
System 5: c, b

Use Borda, Condorcet, and Reciprocal ranking methods to compute the final ranking of the meta search system.

Q.5 **Answer the following:** **[18]**

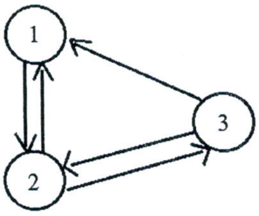
[A] CO3_BL3 For the following corpus, apply naive Bayesian classification for spam email detection. Assume that the documents are already pre-processed. Fit multinomial distribution to the data. [12]

Text	**Class**
Travel offer booking discount	spam
Offer university graduation	non-spam
Booking offer	spam
Graduation travel	non-spam

Use following text as test sample: university travel.

[B] CO2_BL2 Given a choice, which system would you prefer: the one which can handle (a) Boolean queries, or (b) phrase queries, or (c) proximity queries, or (d) natural language queries? Comment on each. [6]

Q.6 CO3_BL4 Consider the following web graph. **[12]**



Calculate PageRank of each of the pages in the web graph using 4 iterations of power iteration method. Assume damping factor d = 0.8.

OR

Q.6 CO3_BL4 Assume a text corpus of 100 documents. Documents in this corpus are belonging to 2 categories namely cricket and movies. Assume that documents are represented as TFIDF vectors of size 200. If we wish to use genetic algorithm to classify these documents in two classes, suggest suitable chromosome encoding, fitness function selection method, crossover operator and mutation operator. **[12]**