NIRMA UNIVERSITY

Institute of Technology

Semester End Examination (IR), May 2023

B. Tech. in Computer Science and Engineering – Semester VI

2CSDE53 – INFORMATION RETRIEVAL SYSTEMS

Roll/ Exam No. □

Supervisor's initial with date

Time: 3 Hours

Max Marks: 100

Instructions:

1. Attempt all questions.
2. Figures to right indicate full marks.
3. Assume suitable assumptions if required and specify them.
4. Use section-wise separate answer sheet.
5. Draw neat sketches wherever necessary.
6. **Sub-questions of each of the six questions must be written together.**

Section-I

Q.1 Answer the following **[18]**

A. CO1 BL4 Why is Cross lingual information retrieval required in the field of information retrieval? Mention the thrust areas where it is required at extreme level. **06**

B. CO3 BL1 Describe the algorithm to perform intersection operation on two posting lists for a query-based search. Use suitable example to explain the process. **06**

C. CO2 BL3 For the following corpus, apply (Bernoulli or multinomial) naive Bayesian classification for spam mail detection. Assume that the documents are already pre-processed. **06**

Bag-of-words	**Label**
Travel offer booking discount	spam
Offer university graduation	non-spam
Booking offer	spam
Graduation travel	non-spam

Use following text as test sample: discount offer

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

A. CO1 BL6 Draw Crawler architecture and explain working of each component. **08**

B. CO3 BL3 For the utility matrix shown in Table 1, users have rated the items in the scale of 1 to 5. In below matrix, U represents the user and P represents the item. Compute the following: **08**

1. What approach would you follow to fill in the blank entries in below utility matrix? Fill the blank entries using that approach.

2. Find which users are similar to user U4?

Table 1: Utility Matrix for Product ratings

	P1	P2	P3	P4	P5	P6	P7	P8
U1	1	3	4		2		4	
U2		2	3	5	3	3	5	2
U3	2	4	2		4	3		
U4			2	3	5	3	4	5
U5	4	1		3			4	3
U6	4			4	2	3	2	4

OR

B. For the following corpus, do as directed: **08**
CO3 Doc 1: watching Cricket match.
BL3 Doc 2: Our watches are matching.
Doc 3: Watch the time.
Doc 4: Time to start the Cricket match.
1. (1 mark) Apply text-preprocessing on this corpus.
2. (1 mark) Extract and display the list of vocabulary terms.
3. (4 marks) Represent each document using TF-IDF model and show necessary calculation.
4. (2 marks) For a given query " Time to watch the match", determine the ranking of all documents retrieved from the system.

Q.3 Answer the following **[16]**

A. Consider the following documents: **06**
CO3
BL3 **Doc 1** breakthrough drug for schizophrenia
Doc 2 new schizophrenia drug
Doc 3 new approach for treatment of schizophrenia
Doc 4 new hopes for schizophrenia patients

a. Draw the Boolean term-document incidence matrix for this document collection.

b. If the query is "schizophrenia drug", which documents will be retrieved for this query? Consider Euclidean distance as the distance measure.

B. For following documents retrieved in response to a query, calculate precision and recall at each rank position. Assume that there are 5 relevant documents as per the ground truth for this query. **06**
CO2
BL4

Rank position	1	2	3	4	5	6	7	8	9
Relevant?	YES	YES	NO	YES	NO	NO	YES	NO	NO

C. Explain the scenarios in which stemming fail in improving retrieval result in web search engine. **04**
CO1
BL1

OR

C. How do you compare two different IR systems? Explain with a suitable example. **04**
CO1
BL5

2CS29E53

Section II

Q-4 Answer the following. **[18]**

A. What are the issues of average color method in image retrieval? Discuss Histogram based method for image retrieval. **06**
CO3
BL2

B. What are the different possible search types in multimedia information retrieval in addition to conventional text retrieval? Elaborate with applications. **06**
CO1
BL1

C. Find below the small portion of bigrams posting list of corpus. **06**
CO2
BL3

po - point->potato->spoke->depot
ot – potato->depot->carrot->teapot
ta- target->potato->tabus->potato
at- float->bloat->offbeat
pp-apple->applet->trappe

If the misspelled word is "potat", which word is suggested using k gram overlap method of spelling correction from the candidate set "potato" , "depot" and "point". Show all the calculations.

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

A. Elaborate in detail, the scenarios in which user prefer a wild card query for retrieval. **6**
CO1
BL1

B. If user want to search for the wildcard query "s*ng", how system can be built to handle such type of query using the concept of permuterm index. **6**
CO2
BL4

C. Which problem occur with blind relevance feedback? Discuss in detail. **6**
CO3
BL2

OR

C. What is phonetic matching? Write and apply the Soundex algorithm on terms "difficulty" and "difference". **6**
CO3
BL3

Q-6 Answer the following. **[14]**

A. To compute the importance of web page which technique assigns two different score to a web page? Why two different scores are assigned? Justify your answer and describe that technique in detail. **7**
CO3
BL4

B. How web search retrieve the correct documents for the below query: "detail of flights flew **form** Heathrow to Ahmedabad" **7**
CO2
BL2

OR

B. Why compression is required in information retrieval? Discuss the types of compression methods with respect to information retrieval. **7**
CO2
BL2