Nirma University

Institute of Technology

Semester End Examination (IR/RPR), May 2022

B. Tech. in Computer Science & Engineering, Semester-VI

2CSDE53 INFORMATION RETRIEVAL SYSTEMS

Roll / Exam No.	19BCE245	Supervisor's initial with date	

Time: 3 Hours Max. Marks : 100

Instructions:

1. Attempt all questions.
2. Figures to right indicate full marks.
3. Use section-wise separate answer book.
4. Draw neat sketches wherever necessary.
5. Simple calculator is permitted.

SECTION – I

Q-1. CLO1 **Answer the following:** **[18]**

[A] For implementing an IR system, would you prefer Boolean model or vector space model? Justify your choice with appropriate claims. [6]

[B] How inverted indexing mechanism helps in solving proximity-based queries for an IR system? Illustrate with an example. [6]

OR

[B] A query is given to a search engine system. The query consists of two terms T1 and T2. Assume that following are the posting lists for these two terms in a corpus. [6]

T1 - D1 --> D3 --> D4 ---> D5 ---> D8 --> D10

T2 – D1 --> D2 --> D4 ---> D6 ---> D9 --> D10 --> D13

Apply merging algorithm to obtain the list of documents in response to this query. Show all the steps of the algorithm and provide a count of total number of comparisons as per the merging algorithm.

[C] Why is cosine similarity approach preferred over Euclidean distance-based approach in IR systems? Explain with appropriate example. [6]

OR

[C] What is the role of relevance feedback in the performance of an IR system? Does it improve the system performance? Discuss. [6]

Q-2. CLO2 **Answer the following:** **[16]**

[A] Given a query vector q = <0,2,1,3> and following document vectors, determine using Cosine similarity the ranking of documents with respect to the query. [8]

D1 = <1,1,1,1>

D2 = <0,2,1,3>

D3 = <2,1,0,2>

D4 = <0,0,1,3>

D5 = <1,2,3,4>

2CSDE53

[B] For a system with four states (A,B,C,D), the state transition probability matrix is given as mentioned below. [8]

$$\begin{array}{c|cccc} & A & B & C & D \\ \hline A & 0.1 & 0.4 & 0.4 & 0.1 \\ B & 0.2 & 0.5 & 0.1 & 0.2 \\ C & 0.4 & 0.2 & 0.2 & 0.2 \\ D & 0.3 & 0.1 & 0.3 & 0.3 \end{array}$$

1. If you start with state A as initial state (S_0) then what is the probability that in iteration S_3 you will come back to state A?
2. Calculate probability of achieving the sequence ACCBDC?

Q-3. Answer the following: [16]

CLO3

[A] Following is the rank given by five search engines to six documents. Some of the documents not ranked by search engines do not appear in their ranking list. Each ranked list contains the highly ranked document as the first in the list and so on. [8]

S1 – A,B,C,D,E
S2 – B,D,F,E,A,C
S3 – C,B,A
S4 – E,D,A,C,B,F
S5 – E,B,A,C,F,D

Determine the combined rank based on Condorcet meta search method.

[B] Does the following matrix have orthonormal set of vectors? If yes then explain, else convert them into orthonormal set of vectors. [8]

$$\begin{bmatrix} 1 & 1 & -1 \\ 0 & 1 & 2 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

SECTION – II

Q-4. Answer the following: [18]

CLO2

[A] Is it possible to have a text corpus in which inverse document frequency (IDF) of all the terms is equal? Represent such corpus using a term document matrix. [4]

OR

[A] Illustrate the application of Information Retrieval in medical domain with an example. [4]

[B] Discuss the pros and cons of following search queries: [6]

1. Keyword based query
2. Proximity query
3. Natural Language Processing (NLP) Query

[C] For the following corpus, apply naive Bayesian classification for predicting the class label of the test sample. Assume that the documents are already preprocessed. Apply Bernoulli model of naïve [8]

2CSDE53

Bayes classification.

Text	Label
North south east west south	Positive
East west	Negative
North north south east	Negative
West east east	Positive

Use following text as test sample: north east

OR

[C] For the corpus above in Que. 4-C, apply multinomial naïve Bayes classifier to classify the same test sample. [8]

Q-5. CLO3 **Answer the following:** **[18]**

[A] Assume that an IR system returns a ranked list of 20 total documents for a given query. Assume that according to a gold-standard labelling there are 8 relevant documents for this query, and that the only relevant documents in the ranked list are in the 1st to 8th positions in the ranked results. List precision and recall at each rank position and show the precision-recall curve. [12]

[B] Compute Eigen values and obtain Eigen vectors from following matrix. [6]

$$\begin{vmatrix} 8 & -5 \\ -5 & 8 \end{vmatrix}$$

Q-6. CLO1 **Answer the following:** **[14]**

[A] Should numbers be removed as a part of stop words removal? Discuss. [4]

[B] Singular Value Decomposition (SVD) transforms semantically similar terms closer to each other in k-concept space representation. Support or refuse this statement with appropriate justification. [5]

[C] List and describe three features of an ideal search engine. [5]