Reg. No. : | | | | | | | | | | | | |

Question Paper Code : 10816

M.C.A. DEGREE EXAMINATIONS, APRIL/MAY 2023.

Elective

MC 4005 — INFORMATION RETRIEVAL TECHNIQUES

(Regulations 2021)

Time : Three hours Maximum : 100 marks

Answer ALL questions.

PART A — (10 × 2 = 20 marks)

1. Consider bitmasks for the documents given in Table 1. Answer the query "Antony AND Brutus AND NOT Calpurnia OR NOT Cleopatra OR Caesar" using boolean retrieval model.

Table 1 : Bit masks for the documents

Terms / Documents	Antony	Brutus	Caesar	Calpurnia	Cleopatra
Antony And Cleopatra	1	1	0	0	0
Julius Caesar	1	1	0	1	0
The Tempest	1	1	0	1	1
Hamlet	0	1	0	0	0
Othello	1	0	0	0	0
Macbeth	1	0	1	1	1

2. Draw the architecture of Information Retrieval (IR) process.

3. Represent the words in the following documents using term frequency technique.

Document 1 : breakthrough drug for schizophrenia

Document 2 : new schizophrenia drug for new disease

Document 3 : new approach for treatment of schizophrenia

Document 4 : new hopes for schizophrenia patients

4. What do you mean by TFIDF? Write down the formula for each component of it.

5. Why do we perform query expansion? List out the methods of doing it.

6. A document collection consists of 200 documents, identified by numbers 1 . . . 200. Suppose that the relevant ones for a given query are those numbered 1 . . . 30. The information retrieval system gives as a result to the query the following answer. Compute precision, recall and F-measure for this retrieval system.

S1={1, 2, 21, 22, 3, 23, 25, 4, 28, 5, 29, 30, 6, 7, 31, 32, 33, 40, 41, 42, 8, 43, 44, 9, 45, 10, 50, 51, 11, 52, 53, 54, 12, 60, 62, 13, 63, 64, 14, 15, 16, 70, 78, 80, 17, 81, 82, 83, 85, 18, 90, 19, 91, 92, 20, 93, 94, 95, 96, 98}.

7. Apply single link agglomerative clustering algorithms on the following distance matrix (A, B, C and D) and show the clusters formed at each step.

$$\begin{bmatrix} 0 & 1 & 4 & 5 \\ 1 & 0 & 2 & 6 \\ 4 & 2 & 0 & 3 \\ 5 & 6 & 3 & 0 \end{bmatrix}$$

8. How will the weight and bias be calculated in Support Vector Machines Algorithm?

9. Draw the working of crawler.

10. What are the various types of indexing techniques available for representing the contents of web documents?

PART B — (5 × 13 = 65 marks)

11. (a) Explain in detail about the components of a search engine.

Or

(b) Write postings merge algorithm and intersection algorithm for arbitrary boolean queries. Explain the steps in it with neat examples.

12. (a) Briefly explain about structured text retrieval models.

Or

(b) Write about various models available for browsing web documents.

13. (a) Draw inverted index and suffix tree for the following document collection.

Document 1 : breakthrough drug for schizophrenia
Document 2 : new schizophrenia drug
Document 3 : new approach for treatment of schizophrenia
Document 4 : new hopes for schizophrenia patients

Or

(b) Consider the following collection of documents,

Document 1 : good movie trailer shown
Document 2 : trailer with good actor
Document 3 : unseen movie

Assuming that an IR system uses the standard term frequency and TFIDF weighting and the user judges the first 2 documents as relevant for the query "movie trailer". What would be the Rocchio-revised queries using term frequency and TFIDF weighting? Use $\alpha = 0.25$, $\beta = 0.75$ and $\gamma = 0.25$.

14. (a) Explain about Latent Semantic Indexing techniques with neat example.

Or

(b) What is meta learning? Write in detail about meta classifier and meta regression.

15. (a) Draw the architecture of parallel and distributed IR and explain in detail each component in it.

Or

(b) How does a search engine assign static and dynamic ranks to web pages? Write in detail.

PART C — (1 × 15 = 15 marks)

16. (a) Assume that there are 1,000,000 documents in total and document frequencies of the terms car, auto, insurance and best are 5,000, 50,000, 10,000 and 1,000 respectively. Compute tf-idf weights for all the terms in Table 2 and form a vector space model with tf-idf weights. Also convert the query "best car insurance" into tf-idf vector. Find cosine similarity between the query and the documents using the computed tf-idf values.

Table 2: Vector space model with term frequencies

Terms/Documents	auto	best	car	insurance
Document 1	27	4	24	14
Document 2	3	33	0	0
Document 3	0	33	29	17
Document 4	13	18	19	31
Document 5	16	23	27	20

Or

(b) For the web graph given in figure 1, write down the teleportation probability matrix for the teleport probability $\alpha = 0.5$. Also calculate PageRank for all the pages (3 iterations) in the graph. Assume that the surfer starts from node 1.

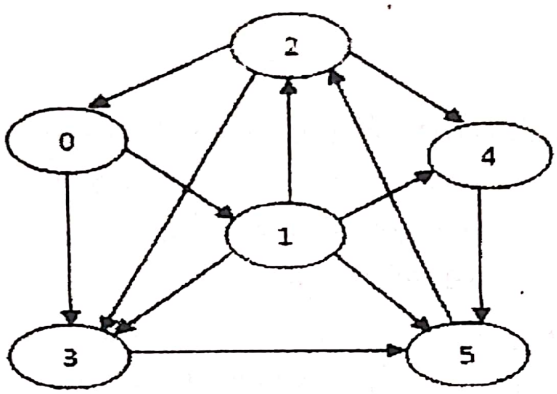

Figure 1: web graph