PRESIDENCY UNIVERSITY
BENGALURU

SET - B

SCHOOL OF ENGINEERING
END TERM EXAMINATION – MAY/JUNE 2024

Semester : Semester VI - 2021

Course Code : CSA3017

Course Name : Information Retrieval

Program : BCA

Date : June 03, 2024

Time : 09.30am to 12.30pm

Max Marks : 100

Weightage : 50%

_____

Note: 1. **Answer ALL 5 FULL Questions.**
2. **Each Full Question carries 20 Marks**
3. **Scientific and non-programmable calculator are permitted.**
4. **Do not write any information on the question paper other than Roll Number.**

_____

1.a.	List the various evaluation methods of Information Retrieval. **[Knowledge]**	**(CO1)**	**(04 Marks)**
1.b.	Explain the components of IR System. **[Comprehension]**	**(CO1)**	**(06 Marks)**
1.c.	Consider the scenario. Let's say the relevant items are: {2,4,6,8,10,12,14,16} The system returns the following top 5 results: [1,2,4,6,20] Calculate the Precision@k **[Application]**	**(CO1)**	**(10 Marks)**

or

2.a.	List the applications of IR. **[Knowledge]**	**(CO1)**	**(04 Marks)**
2.b.	Explain the importance of web search during information retrieval **[Comprehension]**	**(CO1)**	**(06 Marks)**
2.c.	Calculate the Precision@k for the following scenario. Let's say the relevant items are: {2, 4, 6, 8, 10,12,14,16, 20, 35, 56, 78, 93} The system returns the following top 5 results: [1, 2, 4, 6, 20] **[Application]**	**(CO1)**	**(10 Marks)**

3.a.	List the various types of indices. **[Knowledge]**	**(CO2)**	**(04 Marks)**
3.b.	Explain the query processing. **[Comprehension]**	**(CO2)**	**(06 Marks)**
3.c.	Consider the list x = [3, 9, 2, 1, 4, 5] Illustrate the step-by-step construction of Min-heap **[Application]**	**(CO2)**	**(10 Marks)**

Or

4.a.	List the various techniques used for compressing the dictionaries in information retrieval. **[Knowledge]**	**(CO2)**	**(04 Marks)**
4.b.	Explain the life cycle of inverted indices. **[Comprehension]**	**(CO2)**	**(06 Marks)**
4.c.	Suppose we have a posting list representing document IDs [10, 15, 17, 21, 25]. Illustrate the interpolative coding to compress this posting list. **[Application]**	**(CO2)**	**(10 Marks)**

5.a.	List the applications of Kullback-Leibler Divergence in Information Retrieval. **[Knowledge]**	**(CO3)**	**(04 Marks)**
5.b.	Explain Kullback – Leibler Divergence **[Comprehension]**	**(CO3)**	**(06 Marks)**
5.c.	Suppose we have the following information for a collection of documents:	**(CO3)**	**(10 Marks)**

- Average length of the "title" field: 10
- Average length of the "content" field: 100
- Average length of all fields in the collection: 150

Calculate the field weights for each field. **[Application]**

or

6.a.	List the key components of the Robertson/Sparck Jones Weighting Formula. **[Knowledge]**	**(CO3)**	**(04 Marks)**
6.b.	Describe probabilistic classifier. **[Comprehension]**	**(CO3)**	**(06 Marks)**
6.c.	Suppose we have two longer documents:	**(CO3)**	**(10 Marks)**

Document 1: "Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data."

Document 2: "Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence."

Calculate the similarity between two documents using Jaccard Similarity. **[Application]**

7.a	Describe the benefits of caching in information Retrieval. **[Knowledge]**	**(CO4)**	**(04 Marks)**
7.b.	Explain Minimizing Adjudication Effort **[Comprehension]**	**(CO4)**	**(06 Marks)**
7.c	Consider	**(CO4)**	**(10 Marks)**

- Total Number of Requests: 1000
- Number of Cache Hits: 800
- Cache Access Time: 1 millisecond (ms)
- Primary Data Source Access Time: 20 milliseconds (ms)

Calculate the cache hit rate, cache miss rate and average access time. **[Application]**

8.a   Describe the relevance of Queuing Theory in Information Retrieval.   **(CO4)**   **(04 Marks)**
       [Knowledge]

8.b.   Explain Query scheduling. **[Comprehension]**   **(CO4)**   **(06 Marks)**

8.c   **(CO4)**   **(10 Marks)**

Let's assume we have data for 3 queries:

Query	Relevant Retrieved (A)	Total Retrieved (B)	Total Relevant (C)
1	6	10	8
2	3	10	5
3	4	10	6

Calculate Precision, Recall, and F1 Score for Each Query. **[Application]**

9.a   List the various techniques used to retrieve the information. **[Knowledge]**   **(CO1)**   **(04 Marks)**

9.b   Explain the evaluation process during the information retrieval   **(CO1)**   **(06 Marks)**
       [Comprehension]

9.c   Information retrieval system accepts query as an input and it refers the   **(CO1)**   **(10 Marks)**
       document repository which it has for giving relevant set of documents to
       the query. Consider for query Q, the set of relevant documents listed are
       as follows: Rq = { d3, d5, d9, d8, d39, d44, d56, d79, d81, d511}

       Answer set generated for the query by our algorithm and ranked are as
       follows:

Rank	Document
1	d123
2	d86
3	d56
4	d6
5	d8
6	d9
7	d511
8	d129
9	d187
10	d25
11	d38
12	d44
13	d113
14	d250
15	d3

Illustrate the Performance evaluation for the query in this system.
[Application]

Or

10.a   List the difference between Static inverted indices and dynamic inverted indices. **[Knowledge]**   **(CO2)**   **(04 Marks)**

10.b   Explain the process of constructing the indices. **[Comprehension]**   **(CO2)**   **(06 Marks)**

10.c   Suppose we have the following input data and their frequencies   **(CO2)**   **(10 Marks)**

Character	Frequency
A	5
B	9
C	12
D	13
E	16
F	45

Predict the Encoded message for the input data "**ABACDCEF**" using the assigned Huffman codes. **[Application]**