Roll No											

PRESIDENCY UNIVERSITY
BENGALURU

SCHOOL OF ENGINEERING
MID TERM EXAMINATION - APR 2023

Semester : Semester IV - 2021
Course Code : CSE2051
Course Name : Sem IV - CSE2051 - Information Retrieval
Program : ISE

Date : 17-APR-2023
Time : 2PM - 3:30PM
Max Marks : 50
Weightage : 25%

Instructions:
(i) Read all questions carefully and answer accordingly.
(ii) Question paper consists of 3 parts.
(iii) Scientific and non-programmable calculator are permitted.
(iv) Do not write any information on the question paper other than Roll Number.

PART A

ANSWER ALL THE QUESTIONS **(5 X 2 = 10M)**

1. Explain the architecture of open source engine framework.

(CO1) [Knowledge]

2. Draw the inverted index that would be built for the following document collection.
Doc 1 one fish, two fish
Doc 2 red fish, blue fish
Doc 3 one red bird

(CO1) [Knowledge]

3. Discuss Merging and Insertion Concept of Posting List

(CO1) [Knowledge]

4. What are User's Tasks?

(CO1) [Knowledge]

5. Find the Score of a term if tf=32.

(CO2) [Knowledge]

PART B

ANSWER ALL THE QUESTIONS **(5 X 4 = 20M)**

6. How does the base of the logarithm affect the score calculation? How does the base of the logarithm affect the relative scores of two documents on a given query?

(CO1) [Comprehension]

7. Explain the concept of Inverted index file. How it can be used Information Retrieval.

(CO1) [Comprehension]

8. In a situation, the term frequency for a term is 25, and the document frequency for that term is 37. Find the TF-IDF score of that term if the total number of Documents in the collection is 500

(CO2) [Comprehension]

9. What are the advantages and disadvantages of the Boolean Model?

(CO2) [Comprehension]

10. Write the advantages and disadvantages of the Vector model with examples.

(CO2) [Comprehension]

PART C

ANSWER ALL THE QUESTIONS **(2 X 10 = 20M)**

11. Given a document X containing terms t1, t2 and t3 with frequencies (inside brackets) as follows;
t1(3), t2(2), t3(1)
Let us assume that the collection contains 10,000 documents and document frequencies of these terms are as follows;
t1(50), t2(1300), t3(250)
Then, find document X's TF-IDF weight of terms t1, t2, and t3.

(CO2) [Application]

12. Compute the Cosine Similarity between the query "digital cameras" and the document " digital cameras and videos cameras". Assume N=1000. Use Df as follows:

Terms	Df
Digital	100
video	200
cameras	500

(CO2) [Application]