BIRLA INSTITUTE OF TECHNOLOGY, MESRA, RANCHI
(MID SEMESTER EXAMINATION MO/2023)

CLASS: BTECH **SEMESTER : VII**
BRANCH: CS/IT **SESSION : MO/2023**

SUBJECT: IT428 INFORMATION RETRIEVAL

TIME: 02 Hours **FULL MARKS: 25**

INSTRUCTIONS:
1. The question paper contains 5 questions each of 5 marks and total 25 marks.
2. Attempt all questions.
3. The missing data, if any, may be assumed suitably.
***4*. Tables/Data handbook/Graph paper etc., if applicable, will be supplied to the candidates**

			CO	BL
Q.1(a)	Which of the following tasks are examples of IR: • Browsing the library to study about the Egyptian sarcophaguses. • Enquiring about the arrival of the Rajdhani Express from the Railway Enquiry • Looking up the dictionary for the meaning of "anachronism" • Scanning through the morning newspaper	[2]	1	K1
Q.1(b)	Why does a search engine require a logging module? Does this module form a part of the indexing phase or the querying phase?	[3]	1	K2
Q.2(a)	What is the utility of a robot.txt and a sitemap in the context of crawling?	[2]	1	K1
Q.2(b)	What do you understand by the age of a web page. Explain in brief the mechanism adopted by a search engine to decide when to recrawl a page.	[3]	1	K2
Q.3(a)	Which of the following metadata or information is part of the dictionary and which are part of the postings: • Average distance between occurrence of the token. • Length of the token. • Number of occurrences in a document • The three nearest tokens in terms of Edit Distance.	[2]	1	K2
Q.3(b)	A query requires the intersection of two tokens with the following postings: T1: 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20 T2: 1,5,10,15,20 If the first token uses skip lists but the second token does not employ them, how many comparisons shall be required to answer the query? The number of skip grams employed is sqrt(p) where "p" is the length of the posting list.	[3]	1	K2
Q.4(a)	Which are the two principal activities that come under fault tolerant querying.	[2]	2	K1
Q.4(b)	Calculate the Levensthein distance between the tokens "follow" and "flow". Draw the table mapping the distance between each prefix of the two terms.	[3]	2	K2
Q.5(a)	What is the main difference between SPIMI and BSBI?	[2]	2	K2
Q.5(b)	A corpus has 400,000 tokens (entries in the vocabulary). The total number of documents present in the corpus is 800,000. If the longest token contains 20 characters, what is the size of the Incidence Matrix if no compression techniques have been applied. Report your answer in GB. You can assume each character requires 1 byte and you have a choice of 8, 16, 32 and 64 bit sized unsigned integers data types.	[3]	2	K2

:::::20/09/2023 M:::::