BIRLA INSTITUTE OF TECHNOLOGY, MESRA, RANCHI
(END SEMESTER EXAMINATION)

CLASS: BTECH **SEMESTER : V**
BRANCH: CSE **SESSION : MO/2023**

SUBJECT: IT447 INFORMATION RETRIEVAL

TIME: 3 Hours **FULL MARKS: 50**

INSTRUCTIONS:
1. The question paper contains 5 questions each of 10 marks and total 50 marks.
2. Attempt all questions.
3. The missing data, if any, may be assumed suitably.
***4*. Before attempting the question paper, be sure that you have got the correct question paper.**
5. Tables/Data hand book/Graph paper etc. to be supplied to the candidates in the examination hall.

			CO	BL
Q.1(a)	Consider the following queries where you need to provide Recommendation for a query processing order Query: (strawberry OR trees) AND (aircraft OR skies) AND (Brutus OR cleopatra) The posting list is given in sizes as follows: Term — Postings size Brutus — 213312 cleopatra — 87009 aircraft — 107913 skies — 271658 strawberry — 46653 trees — 316812 For a conjunctive query, is processing postings lists in order of size guaranteed to be optimal? Justify your answer.	[5]	1	3
Q.1(b)	We have two word query where one term the posting list consists of the 15 entries as given [1,3,5,7,9,11,13,15,17,19,21,25,31,38,45] And for other it is having one entry positing list [19] How many comparisons are required to intersect the two positing list with Using standard positing list. Justify your answer. Consider the Boolean query: x AND NOT y. How this query will be evaluated using naïve evaluation (that is which term will be evaluated first in naïve version). Why is naive evaluation of this query normally very expensive?	[5]	2	2
Q.2(a)	What is front coding? Give one example to explain the benefits of front coding. If you wanted to search for k*ng in a permuterm wildcard index, what key(s) would one do the lookup on?	[5]	2	1
Q.2(b)	Write SPIMI indexing algorithm and discuss its working. How it is better than BSBI algorithm?	[5]	2	1
Q.3(a)	Consider the table of term frequencies for 3 documents denoted Doc1, Doc2, Doc3 as follows. Compute the tf-idf weights for the terms information, retrieval, system, relevant, for each document.	[5]	3	3
Q.3(b)	In the global champion list what is the purpose of considering TFIDF of term rather than only TF whereas for champion list only TF is sufficient. Why do we differentiate between the two? How the common global ordering by g(d) in high and low lists helps make the score computation efficient. Explain your answer.	[5]	4	4

	Doc1	Doc2	Doc3
Information	27	4	24
Relevant	3	33	0
Retrieval	0	33	29
System	14	0	17

PTO

Q.4(a) Consider the list of Rs and Ns represents relevant (R) and nonrelevant (N) returned documents in a ranked list of 25 documents retrieved in response to a query from a collection of 10,000 documents. The top of the ranked list (the document the system thinks is most likely to be relevant) is on the left of the list. This list shows 7 relevant documents. Assume that there are 9 relevant documents in total in the collection. [5] 5 5

R R N NNNNN R N R N NN R N NNN R NNN R N

Compute the precision and F1 score of the system on the top 20? assume that these 20 documents are the complete result set of the system. What is the MAP for the query? Assume, now, instead, that the system returned the entire 10,000 documents in a ranked list, and these are the first 20 results returned. What is the largest and smallest possible MAP that this system could have?

Q.4(b) Consider the following table showing how two human judges rated the relevance of a set of documents to a particular information need (0 = nonrelevant, 1 = relevant). Let us assume that you've written an IR system that for this query returns the set of documents {2, 5, 6, 7, 8}. [5] 5 5

DocID	1	2	3	4	5	6	7	8	9	10	11	12
Judge1	0	0	1	1	1	1	1	1	0	0	0	0
Judge2	0	0	1	1	0	0	0	0	1	1	1	1

(i) Calculate the kappa measure between the two judges using the confusion matrix. (ii) Calculate precision, recall, and F1 of your system if a document is considered relevant if either judge thinks it is relevant.

Q.5(a) Suppose that a user's initial query is: *SHE SELLS SEASHELLS ON THE SEA SHORE*. The user examines two documents, d1 and d2. The user judges d1, with the content *SHELLS SHE SELLS ARE SURELY CHEAP SEASHELLS* relevant and d2 with content *I AM SURE SHE SELLS SEASHORE SHELLS* nonrelevant. Assume that we are using direct term frequency. Draw the table for term-frequency that consists of terms and its frequency of appearance in query as well as documents. Using Rocchio relevance feedback what would the revised query vector be after relevance feedback? Assume $\alpha = 1$, $\beta = 0.75$, $\gamma = 0.25$. What would be the relevance feedback if the negative feedbacks are set to zero? [5] 5 2,3

Q.5(b) Briefly Describe the reasons why relevance feedback is not popular in web search despite of being cost effective. Consider a search for "Find pages like this one" then for Rochhio's algorithm what should be the weight of α, β and γ. [5] 5 1

::::::24/11/2023::::::M