SRN | | | | | | | | | | | |

PES UNIVERSITY

PES University, Bangalore

(Established under Karnataka Act No. 16 of 2013)

UE17CS412

December 2021 [illegible] ESA B.TECH. VII SEMESTER

UE17CS412 ALGORITHMS FOR INFORMATION RETRIEVAL Question Paper

1.	a)	You are given the following documents Doc1 = English tutorial and fast track Doc2 = learning latent semantic indexing Doc3 = Book on semantic indexing Doc4 = Advance in structure and semantic indexing Doc5 = analysis of latent structures (i) Build Term Document Matrix ii) Find the document for the query "advance AND structure AND NOT analysis"	6+2
	b)	Correct the following statements. If no change is required, indicate 'True', else 'False' a. In a Boolean retrieval system, stemming never lowers precision. b. In a Boolean retrieval system, stemming never lowers recall. c. Stemming increases the size of the vocabulary. d. Stemming should be invoked at indexing time but not while processing a query.	4
	c)	We have a two-word query. For one term, the postings list consists of the following 16 entries: [4,6,10,12,14,16,18,20,22,32,47,81,120,122,157,180] and for the other it is the one entry postings list: [47] How many comparisons would be done to intersect the two postings lists using postings lists stored with skip pointers, with a skip length of $\sqrt{P}$ where P is the length of the posting list.	4
	d)	i. Write one strength and one weakness of Boolean Query based Information retrieval. ii. What is a possible tradeoff issue in skip pointer implementation in Boolean retrieval?	4
2	a)	Describe the BSBI algorithm in detail	5
	b)	Why is dynamic indexing necessary? Describe the logarithmic merge algorithm	5
	c)	A hypothetical TF-IDF values for a three term vocabulary is shown for the three novels Pride and Prejudice (PaP) , Sense and sensibility (SaS) and Withering heights (Wh) in the table. Show that for the query affection, the relative ordering of the scores of the three documents is the reverse of the ordering of the scores for the query jealous gossip	4
	d)	From the following sequence of gamma -coded gaps i.e. 1110001.11010 101 11111011011 11011 reconstruct the postings sequence.	6
3	a)	Below is a table showing how two human judges rated the relevance of 12 documents to a particular information need (0-non relevant,1 –relevant)	4

docID	Judge 1	Judge 1
1	0	0
2	0	0
3	1	1
4	1	1
5	1	0
6	1	0
7	1	0
8	1	0
9	0	1
10	0	1
11	0	1
12	0	1

Calculate the kappa measure between the two judges

SRN | | | | | | | | | | | | |

	b)	Correct the statements below. If no change is required, indicate 'no change' 1. The probabilistic IR model doesn't consider relevance. 2. LSI is a method of soft clustering 3. LSI can retrieve documents when the query and document don't share common terms. 4. Interpolated precision reduces the 'jaggedness' of the PR Curve	4
	c)	What are these? i) Mean Average Precision (MAP) ii) ROC curve iii) Normalized Discounted Cumulative Gain(NDCG)	6
	d)	Consider an information need for which there are 4 relevant documents in the collection. Contrast two systems run on this collection. Their top 10 results are judged for relevance as follows (the leftmost item is the top ranked search result): System 1 R N R N N N N N R R System 2 N R N N R R R N N N i). What is the MAP of each system? Which has a higher MAP? ii). Does this result intuitively make sense? What does it say about what is important in getting a good MAP score? iii) What is the R-precision of each system? (Does it rank the systems the same as MAP?	6
4	a)	The question below has one correct answer. In your answer script, write your chosen correct answer with reason in 1-3 sentences: I. PageRank can be thought of as concentrating on one half of HITS 1. Hub 2. Authority II. In URL frontier of the crawler 1. Front queue manages prioritization and back queue manages politeness 2. Front queue manages politeness and back queue manages prioritization III. In the case of a dangling node (web page) : 1. Basic PageRank update rule will ensure that the node is visited 2. Scaled PageRank update rule will ensure that the node is visited	6
	b)	Provide necessary explanation with reasons in maximum 4 sentences for each statements : i. In Hyperlink Induced Topic Search (HITS) algorithm, the hub and authority score are normalized after every round unlike PageRank ii. Min-heap is used to ensure politeness in the Mercator scheme.	4
	c)	i) What are the main factors that influence Page Rank? ii) We have a small web comprising of 4 pages A,B,C and D. Assuming the damping factor as 0.5 compute the page ranks of these pages	2+4
	d)	A user of a browser can, in addition to clicking a hyperlink on the page x he is currently browsing, use the back button to go back to the page from which he arrived at x. Can such a user of back buttons be modeled as a Markov chain? How would we model repeated invocations of the back button?	4
5	a)	The question below has one correct answer. In your answer script, write your chosen correct answer with reason in 1-3 sentences: I. TextRank uses PageRank of sentences or phrases and 1. is an abstractive summarization approach 2. is an extractive summarization approach II. Using Non-negative Matrix Factorization (NMF), each document in a corpus is successfully represented as a linear combination of some K basis vectors. A basis vector in this case may possibly represent 1. a topic of the document around which clustering can be done 2. a summary of the document III. For granting bank loan to loan applicants using cost sensitive classification : 1. false Negative cost should be higher 2. false positive cost should be higher	6

SRN

	b)	What is the difference between static and dynamic snippets for the same document being a match for two different queries?	2
	c)	A search user is using a query "Capital of India" and he gets the answer "New Delhi, Mumbai, Islamabad", "Kathmandu, Karachi, New Delhi" and "Karachi, New Delhi, Dhaka" in three successive tries. What is the Mean Reciprocal Rank here? The next query is "Cities of India" and he gets the answer "Dhaka, New Delhi, Islamabad", "Kolkata, Karachi, New Delhi," and "Karachi, Bangkok, Chennai" in three successive tries. What is the Mean Reciprocal Rank here?	6
	d)	An insurance company is evaluating two machine learning based classification models A and B for successful detection of fraudulent claims. The following details are available : • Average insurance claim value = INR 100000 • Average premium by a customer = INR 1000 • On the evaluation dataset, number of False Positive for A and B are 80 and 50 respectively and the number of False Negatives for A and B are 50 and 20 respectively Which classification model should be chosen by the company?.	6