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

PES UNIVERSITY

PES University, Bengaluru-85
(Established under Karnataka Act No. 16 of 2013)

UE17CS412

December 2020:END SEMESTER ASSESSMENT (ESA), CSE, VII SEMESTER

UE17CS412–ALGORITHMS FOR INFORMATION RETRIEVAL

Time: 3 Hrs. | Answer All Questions | Max Marks: 100

Provide full calculation for the numerical problems. No partial marking will be done.

Provide brief answers in bullets for all theory questions

1	a)	**The question below has one correct answer. In your answer script, write your <u>chosen correct answer with reason in 1-3 sentences</u>:** i. In minimum Edit Distance, for common errors : • cost of such errors is lower • cost of such errors is higher ii. For a large corpus that is static • Hash table is the best choice for the dictionary and array for posting list • B tree is the best choice for the dictionary and linked list for posting list iii. Soundex algorithm as a query term error correction • Uses one index • Uses two indexes	6 (3x2)
	b)	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+2)
	c)	Write a **modified INTERSECT algorithm** for **XOR** operation on two posting lists.	6
	d)	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
2	a)	**The question below has one correct answer. In your answer script, write your <u>chosen correct answer with reason in 1-3 sentences</u>:** I. If pre-processing is done on corpus using Porter Stemmer even before any compression technique is applied, it is an example of 1. Lossy compression 2. Lossless compression	6 (3x2)

		II. Distributed Indexing is for a scenario when 1. The index is very large but can be fitted into a large hard disk attached to the server 2. The index is too big to be fitted into a single server III. In TF IDF, the document frequency is used instead of collection frequency as 1. collection frequency does not capture frequency of the term well 2. collection frequency does not capture rarity of the term well	
	b)	i. What is a common limitation of BSBI and SPIMI algorithm ? ii. Why very large block is not advisable in the blocking method to save space on term pointers in case of dictionary compression ?	4 (2+2)
	c)	i. What is the possible significance of IDF (inverse document frequency) for a single term query ? Give reasons. ii. Why Jaccard coefficient is not a good scoring framework for term ? Give any two reasons.	5 (3+2)
	d)	Using the tf-idf model of ranking, find out which document will be listed in front of the user for the information given below. For the terms t1, t2, t3, t4, t5, and t6, the vector of **document 1** is **(2, 0, 1, 2, 0, 0)**, the vector of **document 2** is **(0,3,2,1, 1, 1)** and the vector of **query** is **(0, 0, 0,3, 0, 4).** *Assume that these vectors already contain the tf-idf values so you do not have to do any further term weighting.*	5
3	a)	**The question below has one correct answer. In your answer script, write your <u>chosen correct answer with reason in 1-3 sentences</u>:** I. In the language model approach of information retrieval, we are finding 1. the most relevant language model of query to match the documents 2. the most relevant language model of documents to match the query II. Language model approach and vector space model approach of Information retrieval : 1. Vector space model rooted in probability theory and language model rooted in linear algebra 2. Vector space model rooted in linear algebra and language model rooted in probability theory III. If the search engine instruments the mechanism of relevance feedback, the main idea behind that is 1. same query should match more relevant documents 2. modified query should match more relevant documents	6 (3x2)

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

	b)	i. What is the idea of Discounted Cumulative Relevance? ii. How does pseudo-relevance feedback implement relevance feedback?	4 (2+2)
	c)	Compare Language Model in IR with Naive Bayes in text classification with respect to number of classes involved and formal model or distribution used inside .	4
	d)	Consider the following search results for two queries Q1 and Q2 (the documents are ranked in thegiven order, the relevant documents are shown in bold). Q1: D1, **D2**, D3, **D4**, **D5**, **D6**, D7, D8, D9, D10. Q2: **D1**, D2, D3, **D4**, D5, **D6**, D7, D8, D9, and D10. For Q1 and Q2 the total number of relevant documents is, respectively, 4 and 5 whereas total corpus is made of 10 documents. Using the TREC "interpolation" rule, the precision value fare calculated for the 11 standard recall levels 0.0, 0.1, 0.2, ... 1.0. (a) Find the average R-precision **over the two queries** Q1 and Q2 (b) Find the MAP (mean average precision) **over these two queries**.	6 (3+3)
4	a)	**The question below has one correct answer. In your answer script, write your <u>chosen correct answer with reason in 1-3 sentences</u>:** 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 (3x2)
	b)	i. In ranked evaluation, typically precision is emphasized. Give an example where RECALL is important for the search user. ii. Between SEO and SEM, how does SEO help SEM ?	4 (2+2)
	c)	i. State the steady state theorem for Ergodic Markov Chain. ii. How is "ad rank" decided for an ad ?	5 (3+2)
	d)	Two web search engines A and B each generate a large number of pages uniformly at random from their indexes. Thirty percent of A's pages are present in B's index, and 50% of B's pages are present in A's index. What is the number of pages in A's index relative to B's?	5

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

5	a)	**The question below has one correct answer. In your answer script, write your <u>chosen correct answer with reason in 1-3 sentences</u>:** I. In personalized Information Retrieval, relevance estimate 1. does not require any query expansion 2. requires query expansion II. In text summarization, for Rouge Score, the precision indicates 1. fraction of number of overlapping words of reference summary 2. fraction of number of overlapping words of system generated summary III. Feature selection and feature transformation : 1. Both can be used in classification as well as in clustering 2. Both can be used in classification but only feature selection is used in clustering	6 (3x2)
	b)	Define **Mean Reciprocal Rank** as a measure for Question Answering System Evaluation.	4
	c)	i. Which shortcomings of hierarchical and K means clustering are addressed in **Scatter Gather** ? ii. What is the difference between static and dynamic snippets for the same document being a match for two different queries?	5 (3+2)
	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 ? Why ?	5