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

PES UNIVERSITY

PES University, Bangalore

(Established under Karnataka Act No. 16 of 2013)

UE16CS412

December 2019: END SEMESTER ASSESSMENT (ESA) B.TECH. VII SEMESTER

UE16CS412 ALGORITHMS FOR INFORMATION RETRIEVAL QP

Time: 3 Hours | Answer all the questions | Max Marks: 100

1.	a)	Let D be a document in a text collection. Suppose we add a copy of D to the collection. How would this affect the IDF values of all the words in the collection? Why?	4
	b)	Given the following index characteristics term / #(postings) sky / 189 000 blue / 230 000 field / 32 000 red / 453 000 high / 345 000 low / 21 000 Propose an order to process the following query: (sky OR field) AND (blue OR red) AND (high OR low)	4
	c)	Consider these documents: • Doc1: solution found for laziness • Doc2: old laziness found • Doc3: old approach for treatment of laziness • Doc4: old hopes for laziness patients i: Draw the term–document incidence matrix for this document collection. ii: Draw the inverted index representation for this collections.	4 +4
	d)	i. How do you perform stopping and stemming to reduce the size of an inverted index? ii. What is the soundex code for the names for Allan, Allen, Alan, and even Allynn given the mapping of the numbers as (B,F,P, V -> 1) , (C,G, K,Q,S,X,Z -> 2), (D,T -> 3), L->4, (M,N ->5), R->6 (A,E, H,I, J,O,U,W,Y -> 0)	2 2
2	a)	i. How would you create the dictionary in blocked sort-based indexing on the fly to avoid an extra pass through the data? ii. Bring out the basic principle of Single Pass In Memory Indexing (SPIMI).	4+2
	b)	i. What is the basic principle of MapReduce for distributed processing? Give an example for illustration. ii. Write a MapReduce program to solve the following problem. Given a large number of credit card transactions, count the total number of transactions for each distinct card number.	6+2
	c)	You are given the following 3 documents. doc1 : two two tea tea doc2: tea tea me you doc3: me me you you Rank these documents for the query **tea me** based on similarity using vector space technique.	6
3	a)	Describe champion list with an example. What is query-term proximity?	4+2

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

	b)	Below is a table showing how two human judges rated the relevance of a set of 12 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 {4, 5, 6, 7, 8}. i. Calculate the kappa measure between the two judges ii;. Calculate precision, recall, and F1 of your system if a document is considered relevant only if the two judges agree. iii. Calculate precision, recall, and F1 of your system if a document is considered relevant if either judge thinks it is relevant	8
	c)	i. What is the intuition behind LSI ii. 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	2+4
4	a)	What do you mean by the term" Google bombing"? Suggest the ways in which the search engine may cope up with this problem.	4
	b)	Given the following two documents and their shingles, calculate the Jaccard Coefficient of $J(d_1,d_2)$ using the concept of comparing the minimum of the hashed values. $h(x) = x \mod 5$ and $g(x) = (2x+1) \mod 5$	6
	c)	i. List any **two** features a crawler should provide. ii. As per your understanding what are the **two important functions** of Mercator Web Crawler as a scalable Web Crawler?	2+2
	d)	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
5	a)	What is attention and briefly explain how is it implemented in text processing?	2+3
	b)	What is a Knowledge Graph and what are some typical examples?	3
	c)	Suppose we have a collection that consists of three documents given below. D1: Malaga Rimini Tibet Tibet Tibet --> class = Y D2: Tibet Tibet sun sun --> class =Y D3: Mexico Malaga Tibet sun --> class= N Assume that we also have the following query: **D4 Tibet** Malaga. Find the class of D4	6
	d)	Cluster to following documents using K-means with K=2 and cosine similarity. D1: "go monster go" D2: "go karting" D3: "karting monster" D4: "monster monster" Assume **D1** and **D3** are chosen as initial seeds. Use tf (no idf). Show the clusters and their centroids for each iteration. The algorithm should converge after **2** iterations.	6

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

Shingle	d_1	d_2
s_1	1	0
s_2	0	1
s_3	1	1
s_4	1	0
s_5	0	1