

SRN

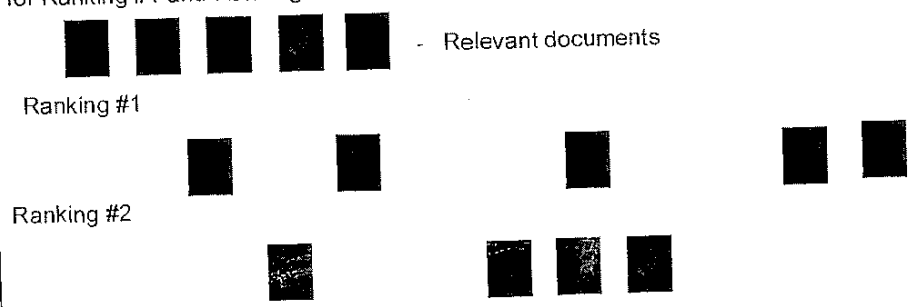
--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

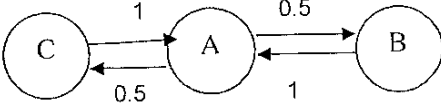


PES University, Bangalore
(Established under Karnataka Act No. 16 of 2013)

UE14CS414
(SN)

December 2017: END SEMESTER ASSESSMENT (ESA) B.TECH. VII SEMESTER
UE14CS414 ALGORITHMS FOR INFORMATION RETRIEVAL QP (Dr SN)

1.	a)	Complex Boolean retrieval systems like Westlaw use many operations that go beyond strictly Boolean operators. Name some of them	4															
	b)	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"	8															
	c)	Distinguish between Stemming and Lemmatization using simple examples Compute the Edit distance between Saturday and Sunday using Levenshtein approach.	2+6															
2.	a)	How would you create the dictionary in blocked sort-based indexing on the fly to avoid an extra pass through the data?	4															
	b)	Apply MapReduce to the problem of counting how often each term occurs in a set of files. Specify map and reduce operations for this task	8															
	c)	Convert decimal number 777 to Variable-byte encoding and Gamma encoding	4															
	d)	Show that, for the query affection, the relative ordering of the scores of the three documents in table below is the reverse of the ordering of the scores for the query jealous gossip. <table border="1" data-bbox="389 1102 836 1228"> <thead> <tr> <th>term</th><th>SaS</th><th>PaP</th><th>WH</th></tr> </thead> <tbody> <tr> <td>affection</td><td>0.996</td><td>0.993</td><td>0.847</td></tr> <tr> <td>jealous</td><td>0.087</td><td>0.120</td><td>0.466</td></tr> <tr> <td>gossip</td><td>0.027</td><td>0</td><td>0.254</td></tr> </tbody> </table>	term	SaS	PaP	WH	affection	0.996	0.993	0.847	jealous	0.087	0.120	0.466	gossip	0.027	0	0.254
term	SaS	PaP	WH															
affection	0.996	0.993	0.847															
jealous	0.087	0.120	0.466															
gossip	0.027	0	0.254															
3.	a)	What are tiered indexes? Give an example. What is query proximity?	4+2															
	b)	You are given a set of relevant documents as well as others. Compute Recall and Precision for Ranking #1 and Ranking #2  - Relevant documents Ranking #1 Ranking #2	8															
	c)	i) What are Probability Ranking Principle (PRP) and Probabilistic Retrieval Strategy? ii) What is Singular value Decomposition (SVD) and how is it utilized in Latent Semantic Indexing?	6															

4	a)	Let us consider Goto method of advertising using bidding. What can go wrong with this when the highest bidding advertiser places an advertisement that is irrelevant to the query? Why might an advertiser with an irrelevant advertisement bid high in this manner?	4																								
	b)	Consider the following five URLs: www.pes.edu representative of typical Universities www.newindianexpress.com representative of online print media www.wikipedia.org representative of monolithic web-scale info. base www.flipkart.com representative of online stores https://shakespeare.mit.edu representative of classical literary work How do these sites exhibit the core information like link characteristics, type of content, frequency of visit etc	5																								
	c)	What as per your opinion are the modules required for building basic crawler architecture? You may explain this by means of a figure briefly touching upon the required modules	7																								
	d)	You are given a simple Markov chain with 3 states A,B and C and the numbers between the links are the transition probabilities. Derive the transition matrix for this. 	4																								
5	a)	Consider the following training set D1: "Indian Bangalore Indian" -> Class=Y D2: "Indian Indian Chennai" -> Class=Y D3: "Indian Delhi" -> Class=Y D4: "Paris French Indian" -> Class=N What is the class of the following test document: D5: "Indian Indian Indian Paris French"	6																								
	b)	How do you define the boundaries between classes using Rocchio classification? Show that Rocchio classification can assign a label to a document that is different from its training set label	4																								
	c)	A search engine returned 5 documents in response to a query. The documents are completely described in terms of the weights and type of keywords they have in the following table <table border="1" data-bbox="446 1155 901 1333"> <thead> <tr> <th></th><th>Computer</th><th>Repair</th><th>Science</th></tr> </thead> <tbody> <tr> <td>D1</td><td>6</td><td>4</td><td>0</td></tr> <tr> <td>D2</td><td>3</td><td>0</td><td>7</td></tr> <tr> <td>D3</td><td>1</td><td>6</td><td>3</td></tr> <tr> <td>D4</td><td>1</td><td>3</td><td>6</td></tr> <tr> <td>D5</td><td>6</td><td>2</td><td>2</td></tr> </tbody> </table> Cluster these into 2 clusters using k-Means algorithm, taking D3 and D4 as initial centers		Computer	Repair	Science	D1	6	4	0	D2	3	0	7	D3	1	6	3	D4	1	3	6	D5	6	2	2	6
	Computer	Repair	Science																								
D1	6	4	0																								
D2	3	0	7																								
D3	1	6	3																								
D4	1	3	6																								
D5	6	2	2																								
	d)	What is single link clustering and complete clustering? Give an example.	4																								