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

PES University, Bangalore

(Established under Karnataka Act No. 16 of 2013)

UE14CS414 (Prof. Ananthanarayan)

END SEMESTER ASSESSMENT (ESA) B.TECH. VII SEMESTER- Dec. 2017

UE14CS414 – Algorithms for Information Retrieval

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

General instructions:

- In case of ambiguity, feel free to make reasonable assumptions. Clearly state your assumptions.
- Answer the questions in the order
- Answer neatly and legibly. Use the margin only for writing question number and not for numbering any lists you have as part of your answer.

1.	a)	I intend to use boolean matching to evaluate your answer paper. Is this acceptable? Justify your answer.	3
	b)	How do you optimize the execution of boolean queries?	4
	c)	Name the methods used to support phrase queries	3
2.	a)	Explain how dynamic indexing works.	7
	b)	What is a parametric index? What is a zone, how does weighted zone scoring work? Note: you do not have to derive the weight scores.	7
	c)	State Zipf's law. You have a collection of documents containing exactly 4 terms a, b, c, and d, with frequencies a > b > c > d. If the number of tokens in the collection is 5000, what are the frequencies of the individual terms.	6
3.	a)	Your search system retrieved the 5 documents for a query and the relevance scores are captured below. Assuming a total of 10 relevant documents, draw the PR curve.	6
	b)	Explain the following methods used to speed up cosine similarity calculations: 1. Champion lists 2. Cluster pruning	5

Document	Relevant
1	Y
2	Y
3	N
4	N
5	Y

SRN

3	c)	Correct the statements below. If no change is required, indicate 'no change' 1. The probabilistic IR model doesn't consider relevance. 2. An empty document has the same probability of relevance and not relevance. 3. Principal Component Analysis is a subset of Latent Semantic Indexing 4. LSI is a method of soft clustering 5. LSI can retrieve documents when the query and document don't share common terms. 6. Low-rank approximations of SVD maximise the Frobenius norm	6x1 =6
	d)	Calculate the kappa statistic for the following confusion matrix. Use pooled marginals in your calculation. Yes / No Yes: 100 / 20 No: 10 / 70	6
4.	a)	What are the mandatory and the recommended features of a web crawler?	5
	b)	Explain the working of the URL frontier of a web crawler.	6
	c)	Calculate the pairwise document similarity for the following documents using 3-shingles. D1: The sky is clear and the moon is shining D2: The moon is shining so are the stars D3: Where are the stars and the moon today?	6
	d)	How are hub and authority scores defined? Calculate the hub and authority scores using the following adjacency matrix. Stop your calculation after the first iteration.	5

		To			
		1	2	3	4
From	1	0	1	0	0
	2	1	0	1	0
	3	1	0	0	0
	4	1	0	1	0

SRN

5 a) What is the Rocchio classification algorithm? Calculate the centroids for the following document classes. Use raw tf (no IDF) for the vector. [8]

		Class label
D1	hybrid battery efficient	electric
D2	lithium battery recharge	electric
D3	energy recharge mileage	electric
D4	'Fossil fuel' pollution hybrid	fuel
D5	Non-renewable mileage energy	fuel

b) Use k-NN to classify the following document against the document classes in 5a. Use Euclidean distance calculation. Test document: energy efficient recharge [8]

c) Merge the following clusters using complete link clustering. Draw the dendrogram. Use dist(xy, i) = max(dist(y, i), dist(x,i)) where xy is formed by merging clusters x & y and i is any other cluster. [8]

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

d) You deserve one really easy question so here it is: Draw a smiley face. [1]