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

PES UNIVERSITY

PES University, Bangalore

(Established under Karnataka Act No. 16 of 2013)

UC15MC622

DECEMBER 2017 : END SEMESTER ASSESSMENT (ESA) MCA. V SEMESTER

UC15MC622 – INFORMATION RETRIEVAL

Time: 3 Hrs | Max Marks: 100

General instructions:

- Answer all questions
- In case of ambiguity, feel free to make reasonable assumptions. Clearly state your assumptions.
- Answer neatly and legibly.

Good luck.

1.	a)	I will use a boolean retrieval method to evaluate your answers. Is this acceptable? Justify your answer. If the approach is not acceptable, provide the alternative.	3
	b)	Write the term incidence matrix for the following document corpus. Assume no stemming/stop lists: Doc 1: breakthrough drug for schizophrenia Doc 2: new schizophrenia drug Doc 3: new approach for treatment of schizophrenia Doc 4: new hopes for schizophrenia patients	9
	c)	How can bi-word indexes be used to support phrase queries? What are the other methods of supporting phrare queries.	8
2.	a)	Describe the BSBI algorithm in detail	7
	b)	Describe how Distributed indexing is implemented.	7
	c)	Change the following statements to make them correct. If no change is required, indicate 'no change'. 1. Edit distance provides the actual character transformations to convert from one string to another. 2. Soundex is a method of phonetic correction. 3. Permuterm index allows us to handle wildcard queries with any number of '*' 4. B-tree and reverse b-tree together support query terms with exactly one wildcard '*' 5. SPIMI is slower than BSBI because it uses larger area of memory for the posting list. 6. All scalable index creation algorithms require an external sorting algorithm.	6
3.	a)	What are parametric and zone indices? How does weighted zone scoring work? Note: you do not have to derive the weight scores.	7
	b)	Define/describe the following 1. Cosine similarity. Explain all terms in the formula. 2. Heaps law 3. Variable byte code	6
	c)	State Zipf's law. You have a collection of documents containing exactly 4 terms a, b, c, and d, with	7

SRN

		frequencies a > b > c > d. If the number of tokens in the collection is 5000, what are the frequencies of the individual terms?	
4.	a)	Correct the following statements wherever required. Mark 'no change' if none is required. 1. Rocchio algorithm is similar to decision trees 2. Mutual information is a method of feature selection. 3. k-NN creates Voronoi tessellations 4. Contiguity hypothesis states that documents in the same class form a contiguous region. 5. Bayesian classification models positional dependence. 6. Basic SVM can do 2-class classification only 7. Rocchio is an unsupervised learning method. 8. k-NN training phase is longer than testing phase.	8
	b)	What is the Rocchio classification algorithm? Calculate the centroids for the following document classes. Use raw tf (no IDF) for the vector.	7
	c)	Describe the k-NN classification algorithm.	5
5.	a)	What are the mandatory and the recommended properties of a web crawler?	5
	b)	Explain the working of the URL frontier of a web crawler.	7
	c)	Calculate the document similarity for the following documents using 3-shingles. D1: The sky is clear and the moon is shining D2: The moon is shining as are the stars	7
	d)	You deserve one easy question so here it is: Draw a smiley face.	1

		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

25