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

PES UNIVERSITY

PES University, Bangalore

(Established under Karnataka Act No. 16 of 2013)

UC14MC622

DECEMBER 2016: END SEMESTER ASSESSMENT (ESA) MCA V SEMESTER

UC14MC622- INFORMATION RETRIEVAL

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

1.	a)	How to process a query using an inverted index and the basic Boolean retrieval model? Give an example.	4
	b)	Describe extended Boolean model with westlaw. And write a query using Westlaw syntax to retrieve the given information. **Information needed:** Information on the legal theories involved in preventing the disclosure of trade secrets by employees formerly employed by a competing company	4+2
	c)	Define Stemming. Explain the rules used in Porter's algorithm for stemming with an example.	1+4
	d)	A school child has been introduced to the alphabets of 3 languages such as Kannada, English and Hindi simultaneously, in total it had (48+26+48). Next day the child was asked to write only Kannada alphabets. The child has written 42 character out of which 29 were from Kannada, 7 from English and 6 from Hindi. Calculate the precision and Recall of the child with respect to kannada.	5
2.	a)	Write a dynamic programming algorithm for computing the edit distance between strings s1 and s2. Trace the algorithm for the term "DOGS" and "GODS" and write the 5 X 5 Matrix for the same.	5+5
	b)	Explain the properties of soundex algorithms for Phonetic correction. Find the soundex codes for the term "phonetically".	6+4
3.	a)	Define Heap's law and Zip's law.	4
	b)	Explain Block Storage dictionary compression with an example.	6
	c)	Write the four possible combinations for boolean match function to compute score using ST(d,q) and SB(d,q).	2
	d)	Consider a collection with 9,87,659 documents. The below table shows the terms present in a document along with their term frequencies(tf) and the document frequencies (df_t). Calculate idf and tf-idf for the terms.	4+4

Terms	tf	df_t
Computer	14	8,29,258
Software	6	67,899
Operating system	12	17,954
Programming	3	45,865

PTO

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

4.	a)	Consider the word "shakespeare" as query term with weights g1 = 0.23, g2= 0.26 and g3= 0.51, what are all the distinct score values a document may get? Title: 5, 13, 45,88 Author : 13, 20, 45, 92 Body : 45, 88	5
	b)	Define Feature selection. Write basic feature selection algorithm for selecting the k best features.	2+3
	c)	Consider the following frequencies for the class coffee for the term "roasted" in 100,000 documents of Reuters-RCV1: use X^2 feature selection method and compute the value for X^2.	10

	$e_{coffee} = 1$	$e_{coffee} = 0$
$e_{roasted} = 1$	$N_{11} = 10$	$N_{10} = 23$
$e_{roasted} = 0$	$N_{01} = 143$	$N_{00} = 99,824$

5.	a)	Define crawling. Explain basic crawlers architecture with a block diagram.	1+4+3
	b)	Describe the features that a crawler must and should provide.	2+6
	c)	Write a short note on web graphs.	4