Birla Institute of Technology & Science, Pilani
Work-Integrated Learning Programmes Division
First Semester 2018-2019

Mid-Semester Test
(EC-2 Regular)

Course No.	: SS ZG537
Course Title	: INFORMATION RETRIEVAL
Nature of Exam	: Closed Book
Weightage	: 30%
Duration	: 2 Hours
Date of Exam	: 29/09/2018 (FN)

No. of Pages = 2
No. of Questions = 7

Note:
1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

Q.1. Discuss the merits and demerits of hash tables and binary trees as data structures commonly used for Inverted Index Dictionary implementation. [2]

Q.2. Compute the minimum edit distance using Levenshtein algorithm, between the terms HEAP and WEEP. Fill in the table given below by distances between all prefixes as computed by the algorithm. [4]

		W	E	E	P
	0	1	2	3	4
H	1				
E	2				
A	3				
P	4				

Q.3. Consider the following documents:
D1: John gives a book to Mary
D2: John reads book by Mary
D3: John thinks a book is a good gift

(a) Draw a term-document incidence matrix for this document collection (Remove the stop words and construct)
(b) Draw the inverted index representation for this collection. (Remove the stop words and construct) [2 + 2 = 4]

Q.4. Given below are two tables, Table 1 gives the tf values and Table 2 gives the idf values for the 4 terms and 2 documents. Compute the scores and rank the documents for the query "best car insurance". Take the tf-idf weights without normalization. [9]

Term	Doc1	Doc2
Car	27	24
Auto	3	0
Insurance	0	29
Best	4	17

Table1: tf values

Term	df*t*	idf*t*
Car	18,165	1.65
Auto	6723	2.08
Insurance	19,241	1.62
Best	25,235	1.5

Table 2: idf values

Q.5. Given below is the Table 3 which lists the 5 documents in the training set and the appropriate class they belong. Also the test document is given.

(a) Estimate the parameters of Naive Bayes classifier.

(b) Apply the classifier to the test document and classify whether it belongs to a1 or b1. [6]

Table 3: Data for NB estimation

	Docid	Words in document	In class a1	In class b1
Training Set	D1	Good	Yes	No
	D2	very good	Yes	No
	D3	Bad	No	Yes
	D4	very bad	No	Yes
	D5	Very bad very bad	No	Yes
Test Set	D6	good bad very bad	?	?

Q.6. Give the name of the index we need to use if

(a) We want to consider word order in the queries and the documents for a random number of words?

(b) What kind of Index can we use if we assume that word order is only important for two consecutive terms? [2]

Q.7. Given a two-word query. The postings list of one term consists of the following 16 entries: [1,7,13, 22, 24, 25, 28, 30, 32, 33, 50, 84, 117, 139, 161, 178] and for the other it is the one entry postings list: [33].

How many comparisons would be done to intersect the two postings lists using skip pointers, with a skip length as discussed in the class. [3]

****