Birla Institute of Technology & Science, Pilani
Work-Integrated Learning Programmes Division
Second 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	: 09/03/2019 (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 (a) Why single pass in memory indexing is better than the block sort based indexing?

Q.1 (b) Discuss in brief the limitations of the Boolean retrieval model.

Q.1 (c) Discuss briefly the index construction algorithm used in Distributed Indexing with a suitable diagram. [2 + 2 + 4 = 8]

Q.2. Given a two-word query. The postings list of one term consists of the following 16 entries: [1, 9, 15, 23, 24, 26, 29, 31, 32, 34, 56, 89, 119, 138, 177, 179] and for the other it is the one entry postings list: [34].
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]

Q.3. Give the name of the index we need to use if [1 + 1 + 2 = 4]

(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?

(c) What is the soundex code for the following two names, Robert and Rupert? Assume that the alphabets are mapped to numbers as follows: *(B, F, P, V → 1), (C, G, J, K, Q, S, X, Z → 2), (D,T → 3), (L → 4), (M, N → 5)* and *(R → 6)* .

Q.4. A fragment from an inverted index (augmented with positional information) is given below. Given the phrase query ***"The rumour city",*** find all the occurrences in the given documents and also give the position of the phrase in the corresponding document. [3]

The:	**rumour:**	**city:**
1:34,38,55;	**1:**12,15,19;	**1:**22,26;
2:12,16,25,44;	**2:**3,5,17,41,45,96;	**2:**18,46,52,65;
3:67,87,90,101;	**6:**21,25,55,62;	**3:**5,69,91,105;
4:33,39,45,62;	**3:**4,68,70,85,110;	**8:**32,42,65,93;
	4:15,34,40,65,81;	**4:**32,44,75,83;

Q.5. Compute the minimum edit distance using Levenshtein algorithm, between the terms WAIT and WEKA. Fill in the table given below by distances between all prefixes as computed by the algorithm. [2]

		W	A	I	T
	0	1	2	3	4
W	1				
E	2				
K	3				
A	4				

Q.6. Given below is the table 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 class A or B. [3 + 4 = 7]

	Docid	**Words in document**	**Class A**	**Class B**
Training Set	D1	White	Yes	No
	D2	Dark White	Yes	No
	D3	Green	No	Yes
	D4	Dark Green	No	Yes
	D5	Dark Green Dark Green	No	Yes
Test Set	**D6**	**White Green Dark Green**		

Q.7. Assume that Simple term frequency weights are used (with no IDF factor), and the stop words “is”, “am” and “are” are removed. Compute the cosine similarity of the following two documents: [Show the term frequency matrix] [3]

Doc1: “temperature is bit bit bit low”

Doc2: “low temperature is bit bit harmful”

****