Birla Institute of Technology & Science, Pilani
Work-Integrated Learning Programmes Division
First Semester 2018-2019
Comprehensive Examination (EC-3 Regular)

Course No.	: SS ZG537
Course Title	: INFORMATION RETRIEVAL
Nature of Exam	: Open Book
Weightage	: 50%
Duration	: 3 Hours
Date of Exam	: 24/11/2018 (FN)

No. of Pages = 3
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. Answer the following questions precisely not exceeding two sentences**:** [1 * 5 = 5]

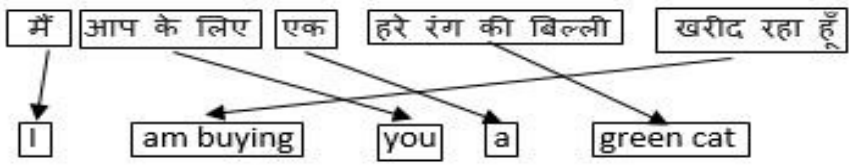
(a) Define the term PRECISION, in Information Retrieval.
(b) What is the difference between Adhoc type of retrieval and Filtering type of retrieval?
(c) What is the difference between BSBI and SPIMI?
(d) Name two methods of handling Phrase queries.
(e) What is the bag of words model?

Q.2. Cluster to following documents using K-means with K=2. [5]
D1: "data collection data"
D2: "data mining"
D3: "mining collection"
D4: "collection collection"
Assume D1 and D3 are chosen as initial seeds. Use tf without idf and Euclidean distance. Show the clusters and their centroids, after each iteration until convergence. Show the intermediate steps clearly.

Q.3 (a) The goal of a retrieval model is to score and rank documents for a query. Different retrieval models make different assumptions about what makes a document more (or less) relevant than another. Suppose you issue the query "lemur" to a search engine. And, suppose that documents D_1 and D_2 contain the term "lemur" once and twice respectively. Answer the following questions.

i. Would the Boolean retrieval model necessarily give both documents the same score? If not, what information would determine which document is scored higher?
ii. Would the Jaccard coefficient necessarily give both documents the same score? If not, what information would determine which document is scored higher?
iii. Would the cosine similarity necessarily give both documents the same score? If not, what would determine which document is scored higher?

Q.3 (b) Using the following phrase aligned sentences (f, e) below, answer questions:



i. Construct the phrase alignment matrix with English words as rows and Hindi words as columns.

ii. Assuming that the alignment matrix from question (i) is the intersection of P(f|e) and P(e|f), identify whether the following phrase pairs are consistent with the alignment:

1. (you, आप के लिए)
2. **(green, रंगकी)**

)

iii. Compute the reordering distance between the following phrase pairs:

1. (you, आप के लिए)
2. (am buying, खरीद रहा हूँ)

iv. In the exponentially decaying cost function $d = \alpha^{|start_i - end_{i-1}|}$, what should be the value of α if the movement of the phrases have to be penalized? [3 + 7 = 10]

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 3 documents. Compute the tf-idf matrix and using that, compute the Euclidean normalized document vectors for each of the documents, where each vector has four components, one for each of the four terms**.** [3]

Table 1: tf values

Term	Doc1	Doc2	Doc3
Cream	15	5	20
Cake	2	22	0
Milk	0	22	15
Butter	3	0	14

Table 2: idf values

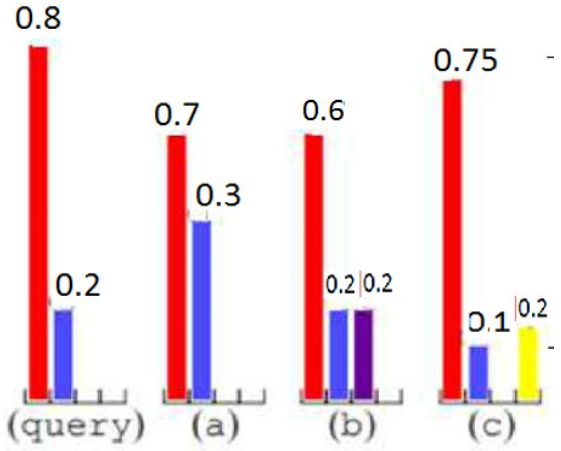
Term	dft	idft
Cream	2312	0.64
Cake	345	1.46
Milk	3030	0.52
Butter	178	1.75

Q.5 (a) Given a grey scale image of size 5x5 pixels in Table 3 with the intensity range K=0,1,2,.., 255. Sketch the histogram to represent the image. [Note in the Y axis you may just show the value of intensities present in the image.]

Table 3

4	100	250	4	200
3	6	35	6	5
5	4	30	35	20
6	3	5	30	10
200	3	4	2	100

Q.5 (b) Given the color histograms for the query and the three images named a, b and c with each histogram having four colors: red, blue, purple, and yellow where the first bin shows number of red pixels, second bin shows blue, third bin shows purple and fourth bin shows yellow. Compute the Canberra distance and Squared chord and rank the images based on both the distances. [2 + 7 = 9]



Canberra Distance

$$d_{Can}(v,w) = \sum_{i=1}^{n} \frac{|v_i - w_i|}{|v_i| + |w_i|}$$

Squared Chord

$$d_{sc}(v,w) = \sum_{i=1}^{n} \left(\sqrt{v_i} - \sqrt{w_i}\right)^2$$

Q.6. Considering the User-Movie ratings matrix shown in Table 4, answer questions below:

Table 4

	M1	M2	M3	M4	M5	M6	M7	M8	M9
U1	?	4	4	2	1	2	?	?	?
U2	3	?	?	?	5	1	?	?	5
U3	3	?	?	3	2	2	?	3	?
U4	4	1	?	2	1	1	2	4	?
U5	1	1	?	?	?	?	?	1	?
U6	?	?	?	?	1	1	?	1	?
Ua	?	?	4	3	?	1	?	5	?

(a) Find the 3 neighbors of Ua using Cosine Similarity.

(b) Predict the rating of Ua for M5 using User-Based collaborative filtering using the similarity calculated in (a).

(c) What is the problem if you have to predict rating for User Ua, Movie M9 using item based collaborative filtering and what is it called? How is this overcome in Content-based recommendation system? [3 + 2 + 2 = 7]

Q.7. Consider the following web pages and the set of web pages they link to:

Page A points to pages C and D.
Page B points to A and C.
Page C points to B.
Page D points to E.
Page E points to A.

Use the below formula for page rank calculations:

$PR(A) = (1-d) + d(PR(T_1)/L(T_1) + \ldots + PR(T_n)/L(T_n))$

(a) For the above given web graph, calculate the page ranks with a damping factor of 0.5 and initial page ranks as [0 0 1 0 0] for the pages [A B C D E] respectively for 2 iterations.

(b) Run the Hubs and Authorities algorithm on the above link of web pages. Show the authority and hub scores for each page for the first two iterations. Assume initial values as a=1 and h=1. [8 + 3 = 11]

****