Name _____ Roll No _____ _____
Section

National University of Computer and Emerging Sciences, Lahore Campus

Course:	**Information Retrieval**	**Course Code:**	**CS317**
Program:	**BS(Computer Science)**	**Semester:**	**Fall 2018**
Duration:	**60 Minutes**	**Total Marks:**	**17**
Paper Date:	**1-Oct-18**	**Weight**	**13%**
Section:	**ALL**	**Page(s):**	**4**
Exam:	**Midterm-1**	**Roll No:**	

Instruction/Notes: ***Attempt the examination on this question paper.***. *You can use extra sheets for rough work but do not attach extra sheets with this paper. Do not fill the table titled Question/marks*

Question	1	2	3	Total
Marks	/ 7	/ 7	/ 3	**/17**

Q1)a) What is advantage of stemming text before indexing? Give examples [2 Points]

Solution:

Stemming helps to match words with same meaning but in different inflectional or derivational form. For example, if a query contains word sing and document has word singing then stemming will convert singing into sing and query word will match the document words with same meaning but in different form.

Q1)b) What is advantage of using Hashmap or dictionary in creation of inverted index. [2 Points]

Solution:

Index is created in linear time in one pass by using hashmaps. Without hashmaps, sorting of (termed,docid) pairs is required which will take O(n lg n) time.

Name _____ Roll No _____ _____
Section

Q1)c) Briefly explain how Map Reduce algorithm divides data for creation of inverted index in distributed environment. You can draw a diagram for illustration. [3 Points]

Solution:

Map reduce divides data based on terms. For example all words starting with particular alphabets will be written to different location on hard disk and will be sent to same inverter.

Q1)c) What is advantage of using inverted index over forward index ? [2 Points]

Solution:

Search is based on keywords and inverted index gives fast access to all documents that contain a particular keyword. In forward index we will have to search entire index to get all documents that contain a particular query word.

Name _____ Roll No _____ _____
Section

Q2)a) Given the three-document corpus and a stop word list below, answer the following question AFTER removing stopwords.

Document 1	information retrieval is process of index search retrieval
Document 2	retrieval is used for evaluation of search results retrieval retrieval
Document 3	evaluation in information in evaluation process search
Query	information retrieval
Stopwords	is , of, in, for, to

Calculate cosine similarity between document 1 and query. Use tf.idf weight for query vector and normalized tf weight for document vector.
[5 Points]

Solution:

Idf of information = log (3/2) = 0.176
Idf of retrieval = log (3/2) = 0.176

normTf of information in doc 1 = 1
normTf of retrieval in doc 1 = 1.3

query vector = <0.176, 0.176>
doc 1 vector = <1,1.3>

query vector length = 0.248
doc 1 vector length = 2.38
Cosine Similarity = (0.176 + 1.3*0.176) / (0.248*2.38) = 0.4048 / 0.59 = 0.686

Q2)b) What is the advantage of using cosine similarity as compared to Euclidean distance. [2 Points]

Solution:

Euclidean distance penalizes document with different length and gives large distance among documents containing same words but have different lengths. Cosine similarity measures the direction of vectors and does not consider two documents different because of their length difference.

Q3) What is the advantage of using IDF weight in Tf.IDF weighting scheme. Illustrate with an example. [3 Points]

Solution:

All words in query are not equally important in assessing relevance of a document. Idf gives rare words more weight and common words less weigh because rare words are more discriminative.