Name: ____________________ Reg #: ____________________ Section: ______________

National University of Computer and Emerging Sciences, Lahore Campus

Course:	Information Retrieval	Course Code:	CS 4051
Program:	BS (Data/Computer Science)	Semester:	Spring 2023
Duration:	60 mins	Total Marks:	27
Paper Date:	25-Feb-23	Weight	%
Section:	BDS-6A, BDS-6B, BCS-8A, BCS-8B	Page(s):	6
Exam:	Midterm 1 Exam Solution		

Instruction/Notes: Attempt the examination on the question paper and write concise answers. Don't fill the table titled Questions/Marks. Extra sheets will not be provided. Last page is for rough work.

Questions	1	2	3	4/5	Total
Marks	/9	/8	/5	/5	/27

Q 1) a) Create inverted index of the following collection:

d1: bsbi use term id
d2: sort term id doc id
d3: spimi use term
d4: no term id sort

Assume you only store word counts and not positions. Assume that main memory can only hold two documents at a time, i.e., the SPIMI algorithm will write to disk each time after two documents, a block, have been processed. Write out the content of a disk block just before merging and the result after merging. [5 Marks]

Solution: Note that the terms are sorted in block 1 and block 2, this is necessary as you cannot merge them in linear time if they are not sorted.

Block 1

Dictionary			Posting List	
Terms	Freq	Pointer		
bsbi	1	→	1	
doc	1	→	2	
id	2	→	1	2
sort	1	→	2	
term	2	→	1	2
use	1	→	1	

Block 2

Dictionary			Posting List	
Terms	Freq	Pointer		
id	1	→	4	
no	1	→	4	
sort	1	→	4	
spimi	1	→	3	
term	2	→	3	4
use	1	→	3	

After Merging

Dictionary			Posting List			
Terms	Freq	Pointer				
bsbi	1	→	1			
doc	1	→	2			
id	3	→	1	2	4	
no	1	→	4			
sort	2	→	2	4		
spimi	1	→	3			
term	4	→	1	2	3	4
use	2	→	1	3		

Q 1) b) Suppose a language has a small vocabulary containing the words:
he, she, driving, the, car, was, road, on, and, fell, ground.
Convert the following document in count vector (use raw counts as weights). Clearly mention the dimensions and value of each dimension of the vector.
he was driving the car on the road and the car [4 Marks]

Solution

and	car	driving	fell	ground	he	on	road	she	the	was
1	2	1	0	0	1	1	1	0	3	1

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

$\mathbf{d_1}$	information retrieval is process of index search retrieval
$\mathbf{d_2}$	retrieval is used for evaluation of search results retrieval retrieval
$\mathbf{d_3}$	evaluation in information in evaluation process search
Q	information retrieval
Stopwords	is , of, in, for, to

Calculate Tf.IDF score of document $\mathbf{d_2}$ for the given query Q. [5 Marks]

- For solution $\log_{10}$ frequency of tf-idf is used $w_{t,d} = \begin{cases} 1 + \log_{10} \mathrm{tf}_{t,d} & \text{if } \mathrm{tf}_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$ $\mathrm{idf}_t = \log_{10} \frac{N}{\mathrm{df}_t}$

terms in Q	Tf-wt	idf	tf-idf
information	0	log 3/2 = 0.18	0
retrieval	1+ log 3 = 1.48	log 3/2 = 0.18	0.266

Q2) b) Suppose we add some documents to an existing collection. Do the weights of terms in other documents change (Tf.IDF weighting)? Yes/No, Justify your answer. [3 Marks]

The Tf.IDF of terms in existing documents will change.
Term frequency will not be effected but IDF score depends on N which is total documents and DF which is document frequency. N will change for all terms and DF will change for terms which are present in new documents. The terms which are present in new documents, their IDF will decrease, the terms which are not present in new documents their IDF will increase.

Q3 a) What proportion of text will be removed if we remove 5 most frequent words (suppose they are stopwords) from a text corpus? [2 Marks]

According to Zipf's law:
Fraction of most frequent word = 0.1/1 = 10%
Fraction of 2nd most frequent word at rank 2 = 0.1/2 = 5%
Fraction of word at rank 3 = 0.1/3 = 3%
Fraction of word at rank 4 = 0.1/4 = 2.5%
Fraction of word at rank 5 = 0.1/5 = 2%
Total = 10+5+3+2.5+2 = 22.5%

Q3 b) With 16,000 documents and 80, 000 unique vocabulary terms, a document by term matrix requires 16,000 * 80,000 = 128×10^7 units of storage. Suppose documents have 4000 terms on average. If we added 2,400 more documents to the collection, roughly how big would the document by term matrix become? Use Heap's law with k = 10 and $\beta = 0.5$. [3 Marks]

Old Docs	*16,000*
New Docs	*2,400*
Total Docs	*18,400*
Avg Tokens/Doc	*4,000*
*Total Tokens T (18,400*4,000)*	*73,600,000*
K	*10*
β	*0.5*
Total Terms $\lvert V \rvert = KT^{\beta}$	*85790.44*
*Term Matrix requires V*TotalDocs Units*	***1,578,544,139***

Q4 Only for Section BDS-6A and BDS-6B

Q 4) Assume that postings lists are gap (delta) encoded using Elias Gamma codes. Using this encoding, suppose that the postings list for the term information is the bit sequence: 1111 1111 1011 1100 1101 0011 1110 0000 0 and the postings list for the term retrieval is the bit sequence: 1111 1111 1100 0000 0011 1011 1101 111

What docids match the following query: information AND NOT retrieval [5 Marks]

Posting of Information:
1111 1111 10 11 1100 110, 1 0 0, 11 1110 0000 0

111 1100 110 = 998, 10 = 2, 100000 = 32

998, 2, 32
After delta decoding
998, 1000, 1032

Posting of retrieval: 1111 1111 110 0 0000 0011 1, 0, 11 110 1 111

1031, 1, 31
After delta decoding

1031, 1032, 1063

Following document ids satisfy the query : information AND NOT retrieval
998, 1000