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		

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]

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

Name: _____ Reg #: _____ Section: _____

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]

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

[1 2 1 0 0 1 1 1 0 3 1 1]

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

d_1	information retrieval is process of index search retrieval
d_2	retrieval is used for evaluation of search results retrieval retrieval
d_3	evaluation in information in evaluation process search
Q	information retrieval
Stopwords	is , of, in, for, to

Calculate Tf.IDF score of document **d_2** for the given query Q. [5 Marks]

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

	d2				
terms in Q	tf	idf	1+log(tf)	log(idf)	tf-idf
information	0	1.5	0	0.1760913	0
retrieval	3	1.5	1.477121	0.1760913	0.26

Tf.IDF score of document **d_2** for the given query Q = **0+0.26 =0.26**

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]

Yes, the Tf.IDF weights of terms in other documents will change as the inverse document frequency idf of all terms will change due to two factors.

1. The existing terms may occur in new documents hence the total number of documents in which they occur will change. So, there can be a change in denominator of IDF.
2. As IDF includes total number of documents as well so the numerator of all IDFs will by all means change.

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]

Zipf's Law states that collection frequency cf of a term t can be described as $cf_i \propto \frac{1}{i}$ *or* $cf_i \propto \frac{c}{i}$ *where i is the rank as per frequency associated with t and c is a constant (in English c = 0.10).*

$$cf_1 = \frac{0.1}{1}, cf_2 = \frac{0.1}{2}, cf_3 = \frac{0.1}{3}, cf_4 = \frac{0.1}{4}, cf_5 = \frac{0.1}{5}$$

$$\frac{0.1}{1} + \frac{0.1}{2} + \frac{0.1}{3} + \frac{0.1}{4} + \frac{0.1}{5} = 0.228$$

*if we remove 5 most frequent words (stop words) from a text corpus then 0.228*100=22.8% of the corpus will be removed.*

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 tokens 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 β = 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 $\|V\| = KT^{\beta}$	*85790.44*
*Term Matrix requires V*TotalDocs Units*	***1,578,544,139***

Name: _____ Reg #: _____ Section: _____

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]

Name: _____ Reg #: _____ Section: _____

Q4 Only for Section BCS-8A and BCS-8B

Q 4) a) Recommend a query processing order for
(Information OR knowledge) OR (web OR net) AND (retrieve OR salvage) AND (images OR graphics)
given the following postings list sizes: [2 Marks]

Term	Frequency
information	2345
knowledge	2450
Web	412
Net	2463
Retrieve	383
Salvage	567
Images	653
Graphics	4578

	Frequency		Sum of Freq	
(Information OR knowledge)	2345	2450	4795	3
(web OR net)	412	2463	2875	2
(retrieve OR salvage)	383	567	950	1
(images OR graphics)	653	4578	5231	4

[[(retrieve OR salvage) AND (web OR net)] AND (images OR graphics)] OR (Information OR knowledge)

Q 1) b) Using the bi-gram spelling correction technique, list the candidate words in the most appropriate one at the top. Show complete calculations [3 Marks]

X= Standford
W1 = Stamfrd
W2 = Standard
W3 = Stanmore
W4 = Stanford

	1	2	3	4	5	6	7	8	9	10	Similar Bi-gram	Total Bi-grams	Similirity
X= Standford	$S	St	ta	an	nd	df	fo	or	rd	d$			
W1 = Stamfrd	$S	St	ta	am	mf	fr	rd	d$			5	13	0.38
W2 = Standard	$S	St	ta	an	nd	da	ar	rd	d$		7	12	0.58
W3 = Stanmore	$S	St	ta	an	nm	mo	or	re	e$		5	14	0.36
W4 = Stanford	$S	St	ta	an	nf	fo	or	rd	d$		8	11	0.73

Ranking

W4 = Stanford
W2 = Standard
W3 = Stanmore
W1 = Stamfrd

Name: _____ Reg #: _____ Section: _____