National University of Computer and Emerging Sciences, Lahore Campus

	Course:	Information Retrieval	Course Code:	CS317
	Program:	BS(Computer Science)	Semester:	Fall 2019
	Duration:	180 Min	Total Marks:	75
	Paper Date:	22-05-19	Weight	%
	Section:	CS	Page(s):	8
	Exam:	Final	Reg. No	

READ INSTRUCTIONS CAREFULLY:

Solve long questions on answer sheet

Write answers of short questions on question paper in space provided.

Staple answer sheet with question paper and return both at end of exam

Write neatly and clearly, illegible answers will not be checked.

Cross out rough work after completing paper.

Give reason, show working and explain with example (where applicable) for each question. No marks will be given for merely answering the question.

Question	Total Marks	Obtained
1	18	
2	12	
3	17	
4	11	
5	17	
TOTAL	75	

Solution key:

Red: Answer

Purple: Notes

Green: Marking criterions, where necessary

Question 1 Duplicate Detection and Crawling (5+5+2+2+2+2)

a. **MINHASH**: Given an input matrix N of three documents (D1, D2, D3), showing occurrence of shingle in each document, and three permutations P1, P2 and P3. Signature matrix M is also given for these three documents. M is created using P1, P2 and P3. Now consider a new document D4 as shown. Find the signature of D4, then find the similarity between D4 and other documents, first using signatures and then also using shingle vectors (using Jaccard similarity measure). Fill all the spaces containing question mark (?). Show your working to get any credit.

Note: Shingles are unigrams (one word) and matrix N shows the presence or absence of shingle in document by 1 and 0 respectively. For example, shingle 1 is present in D1 but not in D2. Each column in N can be considered as document vector.

permutations

p1	p2	p3
2	4	3
3	2	4
7	1	7
6	3	2
1	6	6
5	7	1
4	5	5

input matrix N (shingle x document)

Shingles(unigrams)	D1	D2	D3
1	1	0	0
2	1	0	1
3	0	1	1
4	0	1	1
5	0	1	1
6	1	0	0
7	1	0	0

new document

D4
0
1
1
1
1
0
0

signature matrix M

D1	D2	D3	D4
2	1	2	1
2	1	4	1
1	2	1	2

Similarity

	(D1,D4)	(D2,D4)	(D3,D4)
Using signature	0/3	3/3	0/3
Using document's shingle vector	1/7	3/4	1

Note: The signature of D3 is incorrect, so if you have corrected it or used it as it is or used the incorrect signature of D3 as signature of D4 you will get marks as long as rest of your calculations are correct.

You will lose 2 points if either of this process is not correct. I.e. finding signature, similarity using signature and similarity using shingle vectors.

b. **Simhash:** The binary value of each shingle/unigram in question 1 is give in following table. Read the note given in question 1a to figure out which words are present in D1 and D4 and calculate simhash of D1 and D4.

Shingle/word	Binary value
1	01100001
2	00011110
3	00101010
4	00111111
5	11101110
6	10101011
7	00101101

D1: [-2,-2,2,-2,2,0,0,2]

In binary [0,0,1,0,1,0,0,1]

D4: [-2,-2,2,0,4,2,4,-2]

In Binary [0,0,1,0,1,1,1,0]

Note: if you have converted >=0 as 1 that is also fine

c. Find similarity between D1 and D4 from their simhash (using jaccard similarity measure)

Formula for jaccard similarity

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

2/5

(SOLVE Q1 a, b and c on ANSWER SHEET)

d. Based on your answer in question 1a and 2b, which one gives better similarity estimate, simhash or minhash?

Minhash gave better answer as 0 is more close to 1/7 than 2/5

Note: Your answer is judged based on your calculation on previous parts

e. What should be changed in part c to create a 348 bit simhash of D1 and D4?

The binary value of each shingle/word should be changed to 348 bits

f. What will be effect on increasing the size of simhash (for example from 16 bits to 348 bits)?

Less collusions, and more space will be required

g. Define politeness and how it can be achieved while crawling website/s?
Sending request to one server after certain time as given by server or if not given wait for 20 sec before sending next request. Not crawling restricted contents as given in robots.txt file of website.

Question 2: Indexing (5+5+2)

In your assignment you created a positional index in following format

<Term>, <doc 1 >, <TF in Doc1>,<pos 1 in Doc 1>, <pos 2 in doc 1>, ….. , <doc 2>,<TF in Doc2><pos1 in doc 2>, …….

a) Does this positional index supports the queries that demand all the query terms to be in the same sentence of a document? If not, how would you modify the index to support such queries? In either case use the following documents, query and expected results of query to explain your answer, by showing how positional index or your proposed modified index will be used for retrieval.

Documents.

Doc 1: *I am a student, and I am currently taking CS102. I was a student in CS101 last semester.*

Doc 2: *I was a student. I have taken CS102.*

Query: *student CS102*

Result: *Doc 1*

(Because only in Doc 1 ***student*** *and* ***CS102*** *occur in same sentence.)*

No, this index will not be able to handle such quesries as there is not information of sentence stored in it.
Modification: Sentence number can also be stored in posting list as follow
<Term>, <doc 1 >, <TF in Doc1>,<pos 1 in Doc 1>, <sentence #>, <pos 2 in doc 1><sentence #>, ….. , <doc 2>,<TF in Doc2><pos1 in doc 2><sentence #>,

b). Consider queries of following format ***word1 /k word 2,*** where k is some positive integer. The query should only retrieve documents in which ***word1*** occurs within ***k*** words from ***word2***.
Does positional index support these form of queries? If not, how would you modify the index to support such queries? In either case use the following query and its expected results to explain your answer, by showing how positional index or your proposed modified index will be used for retrieval.

Query: *student /4 CS102*

Results: *Doc 2*

(Because in Doc 2 ***student*** *and* ***CS102*** *are 3 words apart, which is less than 4)*

Yes, positional index is sufficient for such queries. While retrieval difference between position of terms can be matched with k to see if the document satisfies the retrieval condition.

(SOLVE Q2 a and b on Answer sheet)

b) Which retrieval model elastic search uses?

Vector space model

Question 3: Index Compression and Preprocessing. (2+2+2+2+2+2+6)

a. Answer following questions, with reason and example (where applicable), your answer should not exceed 5 lines. No marks will be given without reason/example

i. V-byte is bit level encoding or byte level encoding?

It's a byte level encoding because each number is represented as 1 or more complete byte/s

ii. V-byte is more useful with delta/gap encoding, True or False?

If the gap between numbers to be encoded is less than the actual numbers in sequence, V-byte will be more useful is done on gaps.

iii. Stemming reduces the size of index, True or False?

Yes. As stemming will reduce the number of words, by mapping words will same stem to one word, the vocabulary will reduce.

iv. Removing stop words only from query (if required) is better option that removing stop words at indexing time, True or False?

Yes. In some cases it can be useful. For example consider a query "To be or not to be". In this query if stopwords are removed the query will become empty. Also if in some cases, such as the one given earlier, we decide not to remove stop words, the index should have these aswell words.

v. List two methods to compress vocabulary of index.

Dictionary as string

Front encoding

vi. What is the edit distance between "time" and "climate"?

Changing climate to time required following steps

Remove c, **limate**, replace l with t, **timate**, remove a, **timte**, remove t, **time**.

So edit distance is 4

b. Encode the following posting list of term *tropical*, first gap/delta encoding document and positions and then encode the gaps using gamma encoding. The format of posting list given below is same as given in question 2

Tropical, 10, 3, 20, 21, 30, 12, 2, 6000, 6500

Gap encoding:

Tropical, 10, 2, 20, 1, 9, 2, 2, 6000, 500

Gamma encoding each number in posting list

10, 1110 010

3, 10 1

20, 11110 0100

1, 0

9, 1110 001

2, 10 0

6000, 1111111111110 011101110000

500, 111111110 11110100

Complete encoded posting lists of tropical

1110 010 10 1 11110 0100 0 1110 001 10 0 1111111111110 011101110000 111111110

11110100

Spaces are just for readability

If you have not gap endoded document ID you will lose 3 points

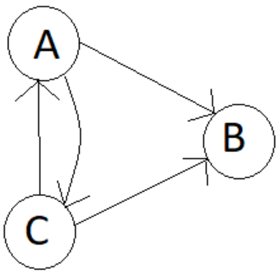
If you have not shown complete posting list in encoded form you will lose 1 point

Question 4: Link analysis (6+2+2)

a) Run two iterations of page rank algorithm on following graph.

The probability of being at a node at t=0 are P(A)= 0, P(B)=0.75 P(C)=0.25

The teleportation probability is =0.1



As B is a dead end so assuming hypothetical links from B to A,B and C

Adj matrix = $\begin{matrix} 0 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \end{matrix}$

P =0.9 * $\begin{matrix} 0 & 1/2 & 1/2 \\ 1/3 & 1/3 & 1/3 \\ 1/2 & 1/2 & 0 \end{matrix}$ + 0.1*[1/3]3x3

P= $\begin{matrix} 0.033 & 0.48 & 0.48 \\ 0.33 & 0.33 & 0.33 \\ 0.48 & 0.48 & 0.033 \end{matrix}$

Xt=0= [0 0.75 0.25]

Xt=1= Xt=0 * P = [0.37 0.37 0.25]

Xt=2= Xt=1 * P = [0.26 0.42 0.31]

Alternative solution (only giving equations)

P(A)= 0.9(P(C)/2 +P(B)/3) +0.1/3

P(B)= 0.9(P(A)/2 + P(C)/2+ P(B)/3) +0.1/3

P(C)= 0.9(P(A)/2 +P(B)/3) +0.1/3

If you have not handled B as a dead end, you will get only 1 points

(SOLVE Q4a on answer sheet)

b) What is Hub and Authority in HITS algorithm?

A hub it a webpage that has links to other pages. For example MOOC directory. An authority is a webpage that is source of information about certain topic. For example Home page of Fast nu website.

c) What is the importance of anchor text in search engines?

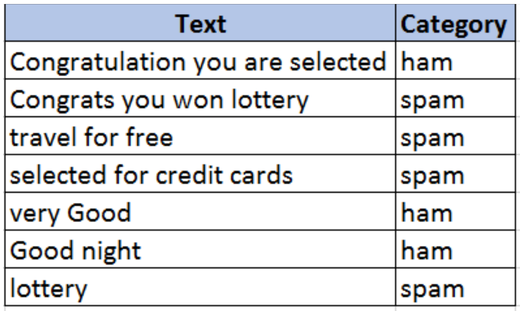
Anchor text gives important information about the content of the page it is referring to. This text can be used in indexing the reference page as well. For example a lot of pages will link IBM's page with anchor text "Big Blue", and although big blue does not appear anywhere on the website of IBM, we can get useful information that Big Blue term is used to refer to IBM

Question 5: Classification and clustering (7+2+2+2+2+2)

a) Consider the following training date set to detect spam or non-spam(ham) emails.

Given a new email ***"congratulations you are selected for lottery"*** classify it either as ham or spam using multinomial Naïve Bayes. Use Laplace smoothing while calculating probabilities.

Show all your calculation.



Text	Category
Congratulation you are selected	ham
Congrats you won lottery	spam
travel for free	spam
selected for credit cards	spam
very Good	ham
Good night	ham
lottery	spam

P(Ham)= 3/7

P(Spam)= 4/7

Probabilities:

Vocabulary	Spam	Ham
congratulation	0+1/12+15	1+1/8+15
you	1+1/12+15	1+1/8+15
are	0+1/12+15	1+1/8+15
selected	1+1/12+15	1+1/8+15
congrats	1+1/12+15	0+1/8+15
won	1+1/12+15	0+1/8+15
lottery	2+1/12+15	0+1/8+15
travel	1+1/12+15	0+1/8+15
for	2+1/12+15	0+1/8+15
free	1+1/12+15	0+1/8+15
credit	1+1/12+15	0+1/8+15
cards	1+1/12+15	0+1/8+15
very	0+1/12+15	1+1/8+15
good	0+1/12+15	2+1/8+15
night	0+1/12+15	1+1/8+15

P(Spam|Email)=4/7 *(0+1/12+15) *(1+1/12+15)* (0+1/12+15) *(1+1/12+15) * (2+1/12+15) *(2+1/12+15)

=5.3X10^{-8}

P(Ham|Email)= 3/7* (1+1/8+15) * (1+1/8+15) * (1+1/8+15) * (1+1/8+15) * (0+1/8+15) * (0+1/8+15) =4.6X10^{-8}

(SOLVE Q5a on ANSWER SHEET)

b) What are the objectives of SVM while finding the decision boundary?

Maximize the margin between two classes whilst minimizing the error/loss

c) Give two uses of Clustering in information retrieval.

Image retrieval

Getting “More like this” result

d) What is the use of automatically labeling clusters?

Clusters are labeled so that user can easily see what the cluster is about. Automatic labeling can make the process quick as compared to manual labeling.

If you have given only 2nd reason you will get only 1 point

e) How will you to automatically label clusters of text documents, label should be in text form?

By using mutual information of terms and cluster.

You could have given another valid way

f) How will you automatically label clusters of images, label should also be an image?

Image that is closest to centroid can be used as label of cluster.