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	

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)				**new document**
Shingles(unigrams)	D1	D2	D3	D4
1	1	0	0	0
2	1	0	1	1
3	0	1	1	1
4	0	1	1	1
5	0	1	1	1
6	1	0	0	0
7	1	0	0	0

signature matrix M

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

Similarity

	(D1,D4)	(D2,D4)	(D3,D4)
Using signature	?	?	?
Using document's shingle vector	?	?	?

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

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

Formula for jaccard similarity

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

(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?

---

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

---

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

---

g. Define politeness and how it can be achieved while crawling website/s?

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.)

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)

(SOLVE Q2 a and b on Answer sheet)

b) Which retrieval model elastic search uses? ----------

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?

---

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

---

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

---

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

---

v. List two methods to compress vocabulary of index.

---

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

---

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

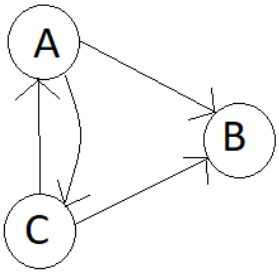
(SOLVE Q3b on ANSWER SHEET)

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



(SOLVE Q4a on answer sheet)

b) What is Hub and Authority in HITS algorithm?

---

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

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

(SOLVE Q5a on ANSWER SHEET)

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

---

c) Give two uses of Clustering in information retrieval.

---

d) What is the use of automatically labeling clusters?

---

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

---

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