


National University of Computer and Emerging Sciences, Lahore Campus

	Course Name:	Information Retrieval	Course Code:	CS317
	Degree Program:	BSCS	Semester:	Fall 2019
	Exam Duration:	180 Minutes	Total Marks:	45
	Paper Date:	26 th Dec 2019	Weight	45
	Section:	ALL	Page(s):	10
	Exam Type:	Final Exam		

Student : Name: _____ **Roll No.** _____

Section: _____

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	4	Total
Marks	/ 2	/ 4	/7	/ 4	/17

Q1) True/False

- a) If I search for term X , and term X has many synonyms, precision is more likely to be a problem than recall.
- b) Smoothing is necessary because otherwise the model would assign a zero probability to queries that contain terms not present in the original document (from which the model was built)
- c) Stemming should be invoked at indexing time but not while doing a query
- d) Words with high Document Frequency (DF) are more discriminative than those with low DF
- e) Stemming increases the size of the lexicon (vocabulary)
- f) Smoothing of a language model is not needed if all the query terms occur in a document.

Q2) a) Which of the following is most likely effective for increasing the PageRank score of a page:

- i. adding an inlink
- ii. adding an outlink
- iii. deleting an inlink
- iv. deleting an outlink

b) Which is most likely going to decrease the PageRank score of the page?

- i. adding an inlink
- ii. adding an outlink
- iii. deleting an inlink
- iv. deleting an outlink

c) What is the most probable type of query for the web query "Microsoft"?

- i. Informational query
- ii. Shopping query
- iii. Downloads and documentation query
- iv. Navigational query

d) Suppose a search returns documents D1, D2, D3, and D4 in this order. The correct results in the system would have been D2, D1, D4, D6, and D5 in this order. What is average precision in this case? [2 Marks]

- i. 0.68
- ii. 0.85
- iii. 0.55

Q3) Apply the SPIMI algorithm to the following collection:

d1: bsbi use term id

d2: sort term id doc id

d3: spimi use term

d4: no term id sort

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 each block just before merging and the result after merging in the following format:

Block 1:

bsbi → 1

...

term → 1, 2

Q4) Assume that postings lists are gap 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

Q5) Consider the following positional index, where docids and positions are delta encoded. Format is as follow:

Docid1: position1, position2,... ..;

Docid2: ...

<information: 11: 7, 18, 33, 12, 46, 31; 2: 3, 149; 54: 17, 11, 291, 30, 44; 5: 433, 67; 77: 54, 12, 3, 22; 19: 4, 12, 333;>	<retrieval: 11: 6, 20, 33, 72, 86, 231; 3: 34, 19; 53: 107, 191, 22, 40, 434; 5: 363, 138;>
--	---

What are all the document Ids and all the absolute positions at which the query phrase “information retrieval” occurs?

Q6) Indexing New York Times newswire from 1991–1995 reveals that it contains about 400 million word tokens, and a lexicon (vocabulary) of size about 1 million (given certain fixed decisions on term normalization, lowercasing, treatment of numbers etc.). What would be a good prediction of how many word tokens and what lexicon size one would get in indexing New York Times newswire from 1991–2000?

Q6) How does Zipf’s law ensure effective inverted index compression?

By storing the gap between consecutive document indices, Zipf’s law ensures that: 1) for frequent words, there will be many duplicated small gaps; 2) for infrequent words, the posting list is short. Therefore, by encoding the gap with variable codes, redundancy can be exploited for effective compression

Q7) Suppose we have a collection that consists of the 4 documents given in the table below.

Document Id	Document Text
Doc1	click go the shears boys click click click
Doc2	click click
Doc3	metal here
Doc4	metal shears click here

Perform retrieval using two models. Fill in these scores in the below table:

a) Tf.IDF

Query	Doc1	Doc2
click		
shears		
click shears		

b) Language model using Jelinek Mercer smoothing ($\lambda = 0.7$)

Query	Doc1	Doc2
click		
shears		
click shears		

Q8) Suppose a word w has occurred 20 times in a document D with 100 words. Assume that the probability of the word according to the collection (background) language model, $p(w|C)$ is 0.01, and the Dirichlet prior smoothing parameter is 200.

- What is the estimated probability of this word in the document language model $p(w|D)$ if we use Dirichlet prior smoothing?
- If we increase the smoothing parameter, would the estimated probability $p(w|D)$ become larger or smaller?
- Why? State the reason for your answer in part b.

Q9) a) Is it possible that one system outperforms the other in terms of MAP by loses to the other in terms of precision at 10 documents? If your answer is yes then justify with an example.

b) Give two advantages of average precision as compared to $P@5$. Justify with examples.
Advantage 1 with example:

Advantage 2 with example:

c) In what situation a system's Mean Average Precision performance will be equal to its Mean Reciprocal Rank performance?

d) What is the advantage of using NDCG as compared to Average Precision? Justify with example.

Q10) In the standard PageRank algorithm, every page starts with the same “inherent” values, and then acquires more value through its inlinks. A. It has been suggested that a link from page P to page Q should be considered as more “significant” if P and Q are from different domains than if they are from the same domain. Give an argument in favor of this.

Q10) In the class, we discussed several ways to evaluate ranking systems, for e.g., precision, recall, NDCG etc. However, for all these metrics, we need the relevance values for the results. Two possible methods to collect relevance values are: a) click feedback from users, b) expert judgements. Write one advantage and one disadvantage of each of these methods. [2 Marks]

a) Click feedback:

Advantage

Disadvantage

b) Expert judgments:

Advantage

Disadvantage

Q11) Suppose we use the number of times a term occurs in all the documents to form a vector to represent each term. For example, if a term T occurred once in document 1, 10 times in document 3, 5 times in document 4, ..., we would have a vector like $V(T)=(1, 0, 10, 4, \dots)$. Suppose we use dot-product or cosine measure to compute the similarity between two vectors representing two terms.

a) What kind of term pairs would have the highest similarity?

b) Suppose we do clustering of terms based on such a similarity function on a collection of product reviews from Amazon. Can we expect to obtain some meaningful clusters of terms? For example, could we expect feature terms describing a particular kind of product (e.g., cell phones) be grouped together? Why?

c) What will be effect (on type of clusters formed) of adding IDF weighting to the weight of each element in a vector?

Q12) a) Word2Vec represents a family of embedding algorithms that are commonly used in a variety of contexts. Suppose in a recommender system for online shopping, we have information about co-purchase records for items x_1, x_1, \dots, x_n (for example, item x_i is commonly bought together with item x_j). Explain how you would use ideas similar to Word2Vec to recommend similar items to users who have shown interest in any one of the items. [3 Marks]

b) What should be training data for WordToVec if you want to use it for following tasks:

Task	Trained on pairs of
Document Ranking	
Query Auto Completion	
Next Query Suggestion	

Q13) Use single link agglomerative clustering to group the data described by the following distance matrix. Show the dendrograms. Show intermediate results by drawing distance matrix.

	A	B	C	D
A	0			
B	1	0		
C	4	2	0	
D	5	6	3	0