



**National University of Computer & Emerging Sciences, Karachi**  
**Spring-2021 Department of Computer Science**  
**Midterm Exam-I (Sol)**



**March 15, 2021, 09:00 AM – 10:00 AM**

<b>Course Code:</b> CS317	<b>Course Name:</b> Information Retrieval
<b>Instructor Name / Names:</b> Dr. Muhammad Rafi, Zeshan Khan	
<b>Student Roll No:</b>	<b>Section:</b>

Instructions:

- Return the question paper.
- Read each question completely before answering it. There are **3 questions** on **2 pages**.
- In case of any ambiguity, you may make assumption. But your assumption should not contradict any statement in the question paper.

**Time: 60 minutes**

**Max: 40 Marks**

<b>Question No. 1</b>	<b>[Time: 25 Min] [Marks: 20]</b>
-----------------------	-----------------------------------

Answer the following questions briefly using 4-5 lines of answer book. Be precise, accurate and to the point, only answer genuine query in the question. Each question is of 2 marks.

- a. Illustrate at least three problems while performing "Tokenization" of documents.

Tokenization is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens. It is very challenging aspect of information retrieval, the tokenization process need to decide about a lot of different aspect of a natural language/processing like:

1. Direction of parsing for tokenization.
2. Should it treat space as separator for token e.g. Les Vegas a single or two tokens.
3. Treatment of punctuation characters like hyphen (-) co-ordinated.
4. If there is no space in between word boundary, how it will decide about tokens like in Japanese or German compound nouns.

- b. What is token normalization? What are its benefits?

Token normalization is the process of canonicalizing tokens so that matches occur despite superficial differences in the character sequences of the tokens. It can reduce the dictionary size and enable quick search. There are several type of normalizations for tokens. Like mapping different variations of a single term(lexeme) to a common single standard term.

- c. What are the two important features of Extended Boolean Model?

Extended Boolean Model is an improvement on generalized Boolean Model. The two important features are (i) using frequency of the query- terms to build ranking and (ii) partial ranking of terms to maintain a continuous score using the proximity operators.

- d. What do we mean by tolerant retrieval? Give an example

In the context of information retrieval, tolerant retrieval is a mechanism through which the systems retrieve relevant documents by using an imprecise query from the users. For example, incorrect spelling of a term. E.g corat will be able to give results on closest term "corrat".

- e. Why permuterm index is efficient for a wildcard query with only one "\*" symbol?

Consider the wildcard query with single "\*" like: m\*n. The key is to rotate such a wildcard query so that the \* symbol appears at the end of the string – thus the rotated wildcard query becomes n\$m\*. Next, we look up this string in the permuterm index, where seeking n\$m\* (via a search tree) leads to rotations of (among others) the terms man and moron. Now that the permuterm index enables us to identify the original vocabulary terms matching a wildcard query, we look up these terms in the standard inverted index to retrieve matching documents.

- f. A bigram index is used to retrieve document for wildcard query "te\*ti\*al". Suggest how a Boolean query on a bigram index would look like for this? Give an example of term that may contain in the result-set.

The Boolean query using bi-grams will be "\$t AND te AND ti AND al AND l\$"

Example is: testimonial

- g. How is context sensitive spelling correction being different from isolated word correction? Give an example.

Context-sensitive spelling correction is the task of fixing spelling errors that result in valid words, such as the query "I'd like to eat dessert", where dessert was typed when desert was intended. Isolated word/term correction is easy as it will find a closest match in a dictionary, for correct term suggestions, while context sensitive corrections are more challenging, as it requires understanding of the entire query.

- h. What is an extended bi-gram index?

An extended bi-word indexing is a way to encode an extended bi-word in indexing. For example, the query term "renegotiation of the constitution" can be index as a bi-word "renegotiation constitution" by skipping the two terms in between. Hence it targets such words that appears more common in a language.

- i. What is the relations of vocabulary with Heaps Law and Zipfs Law? Explain

Zipfs law leads to heaps law. The zipfs law concerning the frequencies on individual words(token) in a collection. It state that, if  $t_1$  is the most common term in the collection,  $t_2$  is the next most common, and so on, then the collection frequency  $c_{fi}$  of the  $i$ th most common term is proportional to  $1/i$ . Mathematically,  $c_{fi}$  directly proportion to  $1/i$ . Under mild assumptions, the Heaps law is asymptotically equivalent to Zipf's law concerning the frequencies of individual words within a text. Heaps Law help us in determining how many terms will be there in a collection. It states that the vocabulary size can be estimated as a function of collection size  $M = kT^b$  where  $k$  is a constant and  $b=0.4$  as an exponent. This is a consequence of the fact that the type-token relation (in general) of a homogenous text can be derived from the distribution of its types.

- j. Compare the difference between Block-Sort Based Indexing (BSBI) and Single-Pass In Memory Indexing (SPIMI).

Block-Sort Based Indexing (BSBI)	Single-Pass In Memory Indexing (SPIMI)
<ul style="list-style-type: none"><li>- BSBI uses continuous disk space to collect all terms from document collections by dividing collection into equal parts, iteratively.</li><li>- It uses a data structures to collect termID and docID into memory.</li><li>- The running time is proportional to <math>(T \log T)</math> where <math>T</math> is Number of terms in the collection. Dominated by sorting of terms in a collection.</li></ul>	<ul style="list-style-type: none"><li>- SPIMI add posting directly to posting list and small posting list are stored into the continuous disk blocks.</li><li>- There is no need to map termID and docID pairs and hence no sorting is required. Faster and efficient.</li><li>- The running time is linear in term of <math>T</math>(number of terms in the collection).</li><li>- SPIMI also support compression of posting lists.</li></ul>

**Question No. 2****[Time: 15 Min] [Marks: 10]**

The Boolean retrieval model has several drawbacks, one is the query formulation, which is not straight forward. Below are some of the information needs from user's, given in plain English, you need to transform these into Boolean Model queries. Using the term statistics from the collection. Suggest a best order for your queries processing, using your knowledge on inverted index and related concepts.

**Some of the term statistics from the dictionary / inverted indexes are given:**

Benefit	167	Hot	319	Reduce	980
Cholesterol	6430	Information	23100	Seeds	12323
Cold	767	Local	23900	Success	23451
Eat	37623	Oats	12398	Swim	32414
Effective	231653	Optimize	879	Technique	387912
Exam	21345	Prepare	1213	Water	1998721
Flax	3124	Rate	982345		
Global	21342				

Q1: Eating flax seeds is more effective at reducing cholesterol than eating oats.

Eat AND Flax AND Seeds AND Effective AND Reduce AND Cholesterol AND Oats

Posting sizes: (37623)  $\cap$  (3124)  $\cap$  (12323)  $\cap$  (231653)  $\cap$  (980)  $\cap$  (6430)  $\cap$  (12398)

The most effective query processing order will be:

Reduce AND Flax AND Cholesterol AND Seeds AND Oats AND Eat AND Effective

Q2: Swimming benefits in hot water or cold water

(Swim AND Benefit AND Water) AND (Hot OR Cold)

Posting sizes: (32414)  $\cap$  (167)  $\cap$  (1998721)  $\cap$  (319 U 767)

The most effective query processing order will be:

Benefits AND (Hot OR Cold) AND Swim AND Water

Q3: success rate of exam preparation techniques

Success AND Rate AND Exam AND Prepare AND Technique

Posting sizes: (23451)  $\cap$  (982345)  $\cap$  (21345)  $\cap$  1213  $\cap$  387912

The most effective query processing order will be:

Prepare AND Exam AND Success AND Technique AND Rate

Q4: local or global optimization techniques

(Local OR global) AND Optimize AND Technique

Posting sizes: (23900 U 21342)  $\cap$  879  $\cap$  387912

The most effective query processing order will be:

Optimize AND (Local OR Global) AND Technique

**Question No. 3****[Time: 15 Min] [Marks: 10]**

Consider the following documents in a collection.

Doc 1: new home sales top forecasts

Doc 2: home sales rise in july

Doc 3: increase in home sales in july

Doc 4: july new home sales rise

- a. Draw the inverted index that would be built for this collection. Assuming “in” is the only stop word that you can filter. Give both the dictionary and posting list for each term. [5]

**Inverted Index**

Dictionary	Posting Lists
forecast	forecast->1
home	home->1->2->3->4
increase	increase->3
july	july->2->3
new	new->1->4
rise	rise->2->4
sale	sale->1->2->3->4
top	top->1

- b. Draw a 3-gram index for the dictionary obtained in part(a). [5]

ale	sale
ase	increase
ast	forecast
cas	forecast
eas	increase
for	forecast
hom	home
inc	increase
ise	rise
jul	july
ncr	increase
new	new
ome	home
ore	forecast

rea	increase
rec	forecast
ris	rise
sal	sale
top	top
uly	july

**BEST OF LUCK**