Information Retrieval

Tuesday, April 11, 2023

Course Instructor

Ms. Faryal Saud, Adeel Ashraf Cheema

Serial No:

2nd Sessional Exam

Total Time:1 Hour

Total Marks: 100

_____

Signature of Invigilator

Roll No	Section	Signature

DO NOT OPEN THE QUESTION BOOK OR START UNTIL INSTRUCTED.

Instructions:

1. Verify at the start of the exam that you have a total of six (6) questions printed on four (4) pages including this title page.
2. Attempt all questions on the question-book and in the given order.
3. The exam is closed books, closed notes. Please see that the area in your threshold of cheating.
4. Read the questions carefully for clarity of context and understanding of meaning and make assumptions wherever required, for neither the invigilator will address your queries, nor the teacher/examiner will come to the examination hall for any assistance.
5. Fit in all your answers in the provided space. You may use extra space on the last page if required. If you do so, clearly mark question/part number on that page to avoid confusion.
6. Use only your own stationery and calculator. If you do not have your own calculator, use manual calculations.
7. Use only permanent ink-pens. Only the questions attempted with permanent ink-pens will be considered. Any part of paper done in lead pencil cannot be claimed for checking/rechecking.

	Q-1	Q-2	Q-3	Total
Total Marks	20	20	20	**60**
Marks Obtained				

Vetted By: _____ Vetter Signature: _____

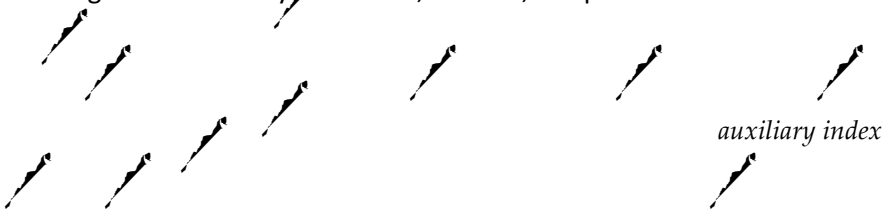
University Answer Sheet Required: No ☐ Yes ☐

Q1: Define following in maximum two lines. 20

Parametric Index	There is one parametric index for each field (say, date of creation); it allows us to select only the documents matching a date specified in the query
Zone	a zone can be thought of as an arbitrary, unbounded amount of text.
Bag of Words	the bag of words model, the exact ordering of the terms in a document is ignored but the number of occurrences of each term is material
Lossy Compression	lossy compression, which discards some information. Case folding, stemming, and stop word elimination are forms of lossy compression.
H Law	estimates vocabulary size as a function of collection size: $M = kT^b$ where T is the number of tokens in the collection.
Logarithmic Merging	up to n postings are accumulated in an in-memory auxiliary index, which we call Z0. When the limit n is reached, the $20 \times n$ postings in Z0 are transferred to a new index I0 that is created on disk. The next time Z0 is full, it is merged with I0 to create an index Z1 of size $21 \times n$. Then Z1 is either with I1 into Z2 (if I1 exists); and so on. We service search requests by querying in-memory Z0 and all currently valid indexes Ii on disk and merging the results.
Auxiliary Index	Maintain a new index while the old one is still available for querying. The auxiliary index is kept in memory. Searches are run across both indexes and results merged
Map Phase	The map phase of MapReduce consists of mapping splits of the input data to key-value pairs.
Reduce Phase	reduce phase, we want all values for a given key to be stored close together, so that they can be read and processed quickly
k-Gram Index	A k-gram is a sequence of k characters. Thus cas, ast and stl are all 3-grams occurring in the term castle. We use a special character $ to denote the beginning or end of a term, so the full set of 3-grams generated for castle is: $ca, cas, ast, stl, tle, le$.

Q2: Write short answers to following 20

a) How do we manage Indexing when collections are modified frequently with documents being added, deleted, or updated. We want new documents to be included into query processing as soon as they are added, deleted, or updated.



b) An enterprise search server for a large corporation must index a multi-terabyte collection with a comparatively large vocabulary, because of the presence of documents in many different languages. Write down their process of decision to come up with a solution to store the dictionary.

Lossy Compression
Removal of stop words, stemming etc
Lossless Compression
Dictionary as a string

c)

document frequency *t*
t

d) What is Query Optimization, and how a query contains AND operator and OR operator affects the optimization process.

Query optimization

t

Brutus Caesar Calpurnia *t*

Q3: Given a document X containing terms t1, t2 and t3 with frequencies (inside brackets) as follows; 20

t1(3), t2(2), t3(1)

Let us assume that the collection contains 10,000 documents and document frequencies of these terms are as follows;

t1(50), t2(1300), t3(250)

Then, find the TF-IDF weight of terms t1, t2, and t3 in document X. Where TF-IDF is defined as

$$TF - IDF = TF_{t,d} * \log \frac{N}{DF_t}$$

D D D D

,

D ,

9

$$TF_{t,d} = \frac{Count\ of\ t\ in\ document\ d}{Total\ number\ of\ words\ in\ d}$$

D / /.

32 D / ; /. - 32 ; . ,/7,

D ,

D /0.

D ; /0. ,

D

,

, . ,

9

$$IDF_t = \log \frac{N}{DF_t}$$

F D

,

D D D D

,

$$TF - IDF = TF_{t,d} * \log \frac{N}{DF_t}$$

D D /9

D/ ; / - ; 1-1

D/ ; , - , / ;
/. . . . - 3. ; 3,1

TF-IDF for t1 = 5.3

D D 09

D0 ; 0- 1

D0 ; /. . . . - /1. . ; 0,

TF-IDF for t2 = 1.3

D D 19

D1 ; /- 1

D1 ; /. . . . - 03. ; 1,5

TF-IDF for t3 = 1.23