Question Paper

Exam Date & Time: 25-Jun-2024 (02:30 PM - 05:30 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

VI Semester BTech (IT/CCE) Makeup Examination June 2024
Information & Communication Technology Department

INFORMATION RETRIEVAL [ICT 4035]

Marks: 50 **Duration: 180 mins.**

Descriptive

Answer all the questions. Section Duration: 180 mins

1) Rank the following documents based on a query "machine learning importance" by using the vector space model with term-frequency, inverse document frequency, tf-idf, and the cosine similarity (5)

Document 1: machine learning is the future technology

Document 2: artificial intelligence and machine learning are closely related

Document 3: the importance of data in machine learning

2) Consider a collection of smart phones from 2 different brands X and Y arranged based on price in Indian rupees as follows: (3)

Brand X: 2000, 2250, 3000, 5000, 10000, 12000, 15000, 20000

Brand Y: 1800, 2200, 3500, 8000, 10000, 12000, 14000, 18000

i. Determine the number of comparisons needed with and without skip pointer to find a smart phone priced 12000 from Brand Y. Assume skip pointers in the alternate position of brand Y price list starting from the first position.

ii. How many comparisons would be made to intersect the two posting lists without considering skip pointer of brand Y? Show all comparisons.

iii. How many comparisons would be made to intersect the two posting lists by considering skip pointer of brand Y? Assume skip pointers in the alternate position of brand Y price list starting from the first position. Show all comparisons.

3) Construct a positional index in the format : term:doc1:[position1, position2,...]; doc2:[position1, position2,...]; etc by considering following documents. The number of position in a document is equal to number of terms separated by blank space. Demonstrate the use of positional index by considering query "quick AND brown AND fox" (2)

Doc 1: the quick brown fox jumps over the lazy dog

Doc 2: a brown fox jumps over the lazy dog quick brown

Doc 3: the dog jumps over the lazy fox

4) Use the Robertson-Sparck Jones's probabilistic model to rank the following documents based on the query "artificial intelligence applications". Assume Doc 1 and Doc 3 as relevant to the query. (5)

Doc 1: artificial intelligence algorithms used for medical diagnosis

Doc 2: self-driving cars powered by machine Intelligence applications

Doc 3: artificial intelligence can help educators by automating certain tasks

5) Construct a matrix (rows=terms, columns=documents, cell=1 if term is present in the document, otherwise 0) for the documents given below and show the results for each part of the query " (artificial AND intelligence) OR (machine AND learning)" (3)

Doc 1: artificial intelligence is used in industries including healthcare

Doc 2: machine learning is used to analyse large data sets

Doc 3: deep learning models are used to recognize artificial image

Doc 4: reinforcement learning is type of machine learning

6) Looking at a collection of web pages, you find that there are 3000 different terms in the first 10,000 tokens and 30,000 different terms in the first 1,000,000 tokens. Assume a search engine indexes a total of 20,000,000,000 (2×10^{10}) pages, containing 200 tokens on average. What is the size of the vocabulary of the indexed collection as predicted by Heaps' law? (2)

7) Determine the gap sequence for the posting list: [15, 28, 42, 56, 78, 102, 115, 130, 145, 159] and construct VB encoding. Retrieve the original posting list by decoding the VB encoded posting list with gaps. Consider the same posting list without gaps and apply VB encoding method. Show the detailed steps/computation. (5)

8) Decompose the matrix S using matrix diagonalization theorem. (3)

$$S = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

9) Justify the usage of inexact top k retrieval to produce k documents that are likely to be among the k highest scoring documents for a query. (2)

10) Identify and illustrate any 5 features that a web crawler should provide which are recommended but not necessarily implemented. (5)

11) Apply the front coding compression technique to the 3 list of terms as follows: (3)

(i) apple, application, appoint, approach, appropriate

(ii) case, cash, castle, cast, casual

(iii) stem, stems, stemless, stemware, stemmer

12) Consider the information need for which there are 5 relevant documents in the collection. The top 10 results as follows (the leftmost item is the top ranked search result): R N R N N R N R N R where R indicate "relevant" and N indicates "non-relevant". Determine the 11-point interpolated precision. Draw a table depicting all values of precision, recall, and interpolated precision. (2)

13) Table Q.5A shows how two human judges rated the relevance of a set of 12 documents to a particular information need (0 = nonrelevant, 1 = relevant). Let us assume that you've written an IR system that returns the set of documents {4, 5, 6, 7, 8} for this query. Calculate the kappa measure between the 2 judges and determine the F1-measure assuming a document is relevant only if the two judges agree. (5)

Table Q.5A

Doc ID	Judge - 1	Judge - 2
1	0	1
2	1	1

3	0	0
4	1	1
5	1	0
6	0	0
7	1	1
8	1	1
9	0	0
10	0	0
11	1	1
12	1	0

14) Consider an information need for which there are 4 relevant documents in the collection. Contrast two systems run on this collection. Their top 10 results are judged for relevance as follows (the leftmost item is the top ranked search result): (3)

System 1 : N R N N R R R N N N

System 2 : R N R N N N N N R R

What is the MAP of each system? Which has a higher MAP?

15) Identify any 2 problems pertaining to the block sort - based indexing that helps in making index construction more efficient. (2)

-----End-----