Question Paper

Exam Date & Time: 08-May-2024 (02:30 PM - 05:30 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

VI Semester BTech (IT/CCE) End Semester Examination April-May 2024
Information & Communication Technoogy Department

INFORMATION RETRIEVAL [ICT 4035]

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

Descriptive

Answer all the questions. Section Duration: 180 mins

Missing data if any can be assumed suitably.

1A) Use the tf-idf method to order the set of documents based on the query "pet cat" by considering the following documents: (5)

Document 1: the pet cat sat on the mat

Document 2 i like to play with my cat

Document 3: the pet dog chased the cat

Document 4: the cat and the dog played together

1B) Consider a collection of products from 2 different companies, A & B, sorted by year of their release to the market. (3)

Company A: 1985, 1986, 1990, 1993, 1995, 1997, 2001, 2005, 2006, 2012, 2015, 2020, 2023

Company B: 1983, 1985, 1990, 1994, 1995, 1996, 2000, 2005, 2007, 2009, 2012, 2017, 2020, 2022

i. Find the number of comparisons with and without skip pointers if you want to find a product that was released in the year 2022 by company B. Consider the following skip pointers: 1983 to 1990, 1990 to 1995, 1995 to 2005, 2005 to 2009, 2009 to 2020, 2020 to 2022. Show all comparisons

ii. How many comparisons would be made to intersect the two posting lists using standard posting lists (without skip pointers)? Show all comparisons

iii. How many comparisons would be made to intersect the two posting lists using skip pointers? Show all comparisons

1C) Consider a document collection of 0.5 million documents with articles related to "fruits". Assume posting list for apple has 220000 documents, banana has 90000 documents, grapes has 260000 documents, mango has 110000 documents, orange has 45000 documents, and pomegranate has 325000 documents. Prioritize the query evaluation order based on the posting list sizes for the following query: (orange OR pomegranate) AND (grapes OR mango) AND (apple OR banana) (2)

2A) Construct a term-document matrix for the documents given below. Retrieve relevant documents for the boolean query: "(python OR java) AND (web OR development OR applications)" (5)

Doc 1: python is a popular programming language for data science

Doc 2: java is widely used for enterprise applications development

Doc 3: javascript is a scripting language commonly used for web development

Doc 4: c++ is known for its efficiency and used in system programming

Doc 5: ruby is favoured by developers for its simplicity

2B) Use the probabilistic model to rank the following documents based on the query "climate change effects". Assume Doc 2 and Doc3 as relevant to the query. (3)

Doc 1: rising sea level effects coastal communities

Doc 2: impact of deforestation results in global climate change

Doc 3: government policies to combat climate change

2C) Consider a document containing 3 zones: title, abstract, body with zone weight as 0.3, 0.5, and 0.2 respectively. Calculate the weighted zone score for the document by assuming the relevance score for each zone as 0.8, 0.6, and 0.4 respectively. (2)

3A) Consider a posting list with document IDs: [10, 25, 35, 70, 100, 425, 800] (5)

i. Compress the posting list using variable byte encoding with gaps and without gaps

ii. Compress the posting list using gamma code with gaps and without gaps

iii. Compare the total number of bytes for variable byte encoding with gaps and without gaps

iv. Compare the total number of bytes for gamma code with gaps and without gaps

3B) Decompose the matrix C using symmetric diagonalization theorem. Show detailed steps. (3)

$$C = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$$

3C) How does champion list help in efficient retrieval and ranking of documents? Justify. (2)

4A) Find the Jaccard coefficient/similarity for all possible pairs of documents given below by constructing shingles of size 2 (one shingle = 2 terms). Show all shingles for each document. (5)

Doc 1: football match is a popular sport played worldwide. fans wish their favourite team to kick ball into opposite team goal.

Doc 2: football match attracts large crowd with fans cheering for their favourite team and win against opposite team

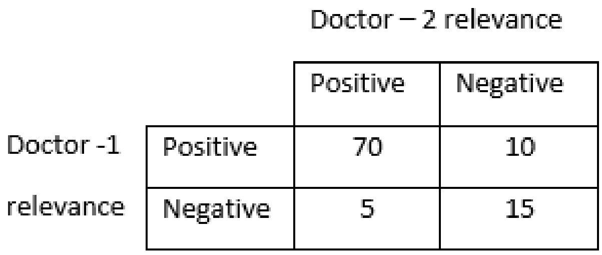
Doc 3: basket ball is another popular sport. Two teams play with the aim to score points by shooting the ball through opposite team hoop

Doc 4: basket ball is a team sport played on a rectangular court. Each team play with the objective to score points by shooting the ball through opposite team hoop.

4B) Illustrate the difference between dictionary search with blocking and without blocking by taking any suitable example. (3)

4C) Consider a information retrieval system that retrieves documents for a given query as: {R, N, N, R, R, N, N, R, R, N}. The total relevant documents for the query are 6. Calculate the precision for each retrieved result and find interpolated precision at recall level 0.3 and 0.7. (2)

5A) A study was conducted to assess the inter-rater agreement between two doctors in diagnosing a certain medical condition. Each doctor independently diagnosed 100 patients as either having the condition (positive diagnosis) or not having it (negative diagnosis). The results are summarized as follows: (5)



		Doctor – 2 relevance	
		Positive	Negative
Doctor -1 relevance	Positive	70	10
	Negative	5	15

5B) Consider documents retrieved by a search engine for 3 queries Q1:N, R, R, N, R; Q2: R, N, R, N, R; Q3: R, R, N, R, N. Find precision for all 3 queries, and compare with their respective MAP (mean average precision). (3)

5C) The key idea of single pass in memory indexing is to generate dictionaries for each block without a need to maintain term-termID mapping across blocks. Identify any 2 advantages of this concept and justify. (2)

-----End-----