Question Paper

Exam Date & Time: 21-May-2022 (10:00 AM - 01:00 PM)

MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
(A constituent unit of MAHE, Manipal)

Deaprtment of Information & Communication Technology
VI Semester End Semester Examination May-2022

INFORMATION RETRIEVAL [ICT 4035]

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

Descriptive Questions

Answer all the questions. Section Duration: 180 mins

MISSING DATA, IF ANY, MAY BE SUITABLY ASSUMED

1)

A) Compute the ranks for the 5 documents using tf-idf model for the query "run happy" by considering the following 5 documents (Assume stemming and stop-word removal are not required). Show all the steps. (5)

doc1: phone ring person happy person

doc2: dog pet happy run jump

doc3: cat pet person happy

doc4: life smile run happy

doc5: life laugh walk run run

B) Consider the documents given below: (3)

Doc 1 breakthrough drug for schizophrenia

Doc 2 new schizophrenia drug

Doc 3 new approach for treatment of schizophrenia

Doc 4 new hopes for schizophrenia patients

Draw the term-document incidence matrix for this document collection.

Draw the inverted index representation for this collection

Write the results for the following queries:

a. schizophrenia AND drug

b. for AND NOT(drug OR approach)

C) Generate the entries in the permuterm index dictionary for the term "sing". If someone wants to search for s*ng in a permuterm wildcard index, what key(s) would one do the lookup on? (2)

2)

A) Consider the table given below with the information of the relevance of a set of 10 documents to a particular information need (0 = nonrelevant, 1 = relevant) given by two judges Mark and Susan. Assume that, an IR system returns the set of documents {2, 5, 6, 7} for a query. (5)

Document Id	Mark	Susan
1	1	1
2	1	0

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

i. Calculate the kappa measure between the two judges.

ii. Calculate precision, recall, and F1 of the system if a document is considered relevant only if the two judges agree.

iii. Calculate precision, recall, and F1 of your system if a document is considered relevant if either judge thinks it is relevant.

iv. Find the MAP based on Mark's judgement.

v. Find the MAP based on Susan's judgement

B) Consider the following fragment of a positional index with the format: (3)

word: document:⟨⟨position, position,...⟩...⟩; document:⟨⟨position,...⟩...⟩

Gates: 1:⟨⟨3⟩⟩; 2:⟨⟨6⟩⟩; 3:⟨⟨2,17⟩⟩; 4:⟨⟨1⟩⟩;

IBM: 4:⟨⟨3⟩⟩; 7:⟨⟨14⟩⟩;

Microsoft: 1:⟨⟨1⟩⟩; 2:⟨⟨1,21⟩⟩; 3:⟨⟨3⟩⟩; 5:⟨⟨16,22,51⟩⟩;

The /*k* operator, word1 /*k* word2 finds occurrences of word1 within *k* words of word2 (on either side), where *k* is a positive integer argument. Thus *k* = 1 demands that word1 be adjacent to word2. Determine the set of documents that satisfy the query Gates /k Microsoft for k=1, k=2, and k=5.

C) Find the soundex code for the following: (2)

i. Mary

ii. Chebyshev

3) Illustrate SVD dimensions and sparseness with the help of a neat diagram. Find the Singular Value Decomposition of the matrix given below: (5)

A)

$$A = \begin{bmatrix} 4 & 4 \\ -3 & 3 \end{bmatrix}$$

B) Let the static quality scores for Doc1, Doc2 and Doc3 be 0.25, 0.5 and 1. Write the posting list for impact ordering when each postings list is ordered by the sum of the static quality score and the Euclidean normalized tf values for the table given below: (3)

	Doc1	Doc2	Doc3
Car	0.88	0.09	0.58
Auto	0.10	0.71	0
Insurance	0	0.71	0.70

C) From the following sequence of *γ*-coded gaps, reconstruct the gap sequence and then the postings sequence: 1110001110101011111101101111101 (2)

4) Differentiate between cost per click and cost per mil pricing model. The Goto method ranked (5)

A) advertisements matching a query by *bid*: the highest-bidding advertiser got the top position, the second-highest the next, and so on. What can go wrong with this when the highest-bidding advertiser places an advertisement that is irrelevant to the query? Why might an advertiser with an irrelevant advertisement bid high in this manner? Suppose that, in addition to bids, we had for each advertiser their *click-through rate*: the ratio of the historical number of times users click on their advertisement to the number of times the advertisement was shown. Suggest a modification to the Goto scheme that exploits this data to avoid this problem.

B) The following list of R's and N's represents relevant (R) and nonrelevant (N) returned documents in a ranked list of 20 documents retrieved in response to a query from a collection of 10,000 documents. The top of the ranked list (the document the system thinks is most likely to be relevant) is on the left of the list. This list shows 6 relevant documents. Assume that there are 8 relevant documents in total in the collection. (3)

R R N N N N N N R N R N N N R N N N N R

i. What is the precision of the system on the top 20?

ii. What is the F1 on the top 20?

iii. What is the uninterpolated precision of the system at 25% recall?

C) Consider a collection of 806791 documents. Determine the inverse document frequency of the terms van, car, insurance, and vehicle by considering their document frequency as 18165, 6723, 19241, and 25235 (2)

5)

A) Outline the steps for deriving the transition probability matrix. Consider a web graph with three nodes 1, 2 and 3. The links are as follows: 1 → 2, 3 → 2, 2 → 1, 2 → 3. Write the transition probability matrices for the surfer's walk with teleporting, for the teleport probability: α = 1. Compute surfer's probability distribution vector after 2 steps assuming he starts walk at node 1. (5)

B) Describe the difference between relevance feedback and query expansion in terms of user interaction. Given the query "elvis music" and the term frequencies for the three documents as given in the table, Use Rocchio to compute the new query vector by assuming doc3 as relevant through relevance feedback, with α = 2, β =γ =1. Show the detailed steps. (3)

	Elvis	Presley	Mississippi	Pop	Music	life
Doc1	3	4	0	6	0	0
Doc2	4	0	4	0	0	3
Doc3	5	3	0	4	4	0

C) Show that $\lambda = 2$ is an eigenvalue of C and find the corresponding eigenvectors. (2)

$$C = \begin{pmatrix} 6 & -2 \\ 4 & 0 \end{pmatrix}$$

-----End-----