| Reg. No. | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|

# MANIPAL INSTITUTE OF TECHNOLOGY

MANIPAL

A Constituent Institution of Manipal University

**V SEMESTER B.TECH. (INFORMATION TECHNOLOGY/COMPUTER AND COMMUNICATION ENGINEERING) MAKEUP EXAMINATIONS, DEC 2017/JAN 2018**

## SUBJECT: PROGRAM ELECTIVE I – INFORMATION RETRIEVAL [ICT 4006]

**(01/01/ 2018)**

Time: 3 Hours MAX. MARKS: 50

**Instructions to Candidates:**

- Answer ALL the questions.
- Write the detailed steps for all the problems.
- Missing data, if any, may be suitably assumed.

1A. Write the optimized algorithm for conjunctive queries and recommend a query processing order for the following queries based on the data given in Table. Q.1A.
(i) (black-money OR deadline) AND (rbi OR corruption) AND (loan OR demonetization)
(ii) black-money AND (NOT rbi) AND (NOT corruption)
(iii) (demonetization AND deadline) AND (NOT corruption)

Table. Q. 1A

| Term | Document-frequency |
|---|---|
| demonetization | 102201 |
| loan | 76008 |
| rbi | 106802 |
| corruption | 160547 |
| black-money | 35542 |
| deadline | 21570 |

5

1B. Explain the following theorems with an example.
(i) Matrix Diagonalization Theorem.
(ii) Symmetric Diagonalization Theorem. 3

1C. Write all permuterm index for the term *ping*. Which permuterm key would be used to lookup on the query ng$p*? 2

2A. Write the algorithm for proximity intersection of postings lists p1 and p2. Consider the following positional index in the form given below to write all the documents and all the absolute positions at which the query phrase "information retrieval" occurs?
<word:
docid1: position, offset, offset ... ;
offset_to_docid2: position, offset, offset ... ;
offset_to_docid3: position, offset, offset ... ;
etc. >

| | |
|---|---|
| <information:<br>11:7, 18, 33, 12, 46, 31 ;<br>2: 3, 149 ;<br>54: 17, 11, 291, 30, 44 ;<br>5: 433, 67 ;<br>7: 54, 12, 3, 22;<br>> | <retrieval:<br>11: 6, 20, 33, 72, 86, 231;<br>3: 34, 19;<br>53: 107, 191, 22, 40, 434;<br>5: 363, 138;<br>> |

5

2B. Consider the three documents ($d_1$, $d_2$, $d_3$)

$d_1$= Shipment of gold damaged in a fire

$d_2$= Delivery of silver arrived in a silver truck

$d_3$= Shipment of gold arrived in a truck

and the query q=" gold silver truck" .

Rank the documents using the Probabilistic Model for the given query q by assuming the relevant document set as { $d_2$, $d_3$ } and $d_1$ as the non-relevant document.

2C. Explain "relevance feedback" and "pseudo relevance feedback" methods used in query refinement. 2

3A. Consider a query ($q$) and a document collection consisting of three documents. Rank the documents using vector space model. Assume that tf-idf weighing scheme is used.

$q$: Singapore Russia Ireland

$d_1$: Canada Germany Singapore SriLanka Japan India Ireland

$d_2$: Canada Germany Singapore Malaysia Japan India USA

$d_3$: Australia Germany Russia SriLanka Japan India Russia Ireland

Note: List the vector elements in alphabetical order. 5

3B. Write and explain the FastCosineScore ranking algorithm 3

3C. Consider the following hypothetical data of a medical test where two radiologists R1 and R2 rated 50 images for needing further study. They either said Yes (good for further study) or No (Not good for further study). They both rated 20 images as Yes and 15 images as No by both. R1 said Yes to 25 images and No to 25. R2 said Yes to 30 images and No to 20. Calculate Kappa measure for this data set. 2

4A. Explain any two posting list compression techniques. Compute variable byte code and γ code for the postings list <777, 17743, 294068, 31251336>. Use gaps instead of docIDs wherever possible. 5

4B. Explain hub and authority score with suitable examples. 3

4C. For the following documents compute the similarities between the document using 2-shingles.

$d_1$: Inspired students at Manipal University gave life to their imagination.

$d_2$: Manipal University gave opportunities to students to realize their imagination. 2

5A. Explain URL frontier components of web crawler with a neat diagram. 5

5B. Find SVD for the following matrix.

$$\begin{bmatrix} 2 & 2 \\ -1 & 1 \end{bmatrix}$$

3

5C. Consider the two documents with three zones, namely Author, Title & Body as given in Table. Q.5C. The weights corresponding to Author, Title, Body zones is $g_1$= 0.2, $g_2$=0.2 & $g_3$=0.6. Rank $d_1$ & $d_2$ based on weighted zone scoring.

for the query - Tagore AND Gitanjali

2

Table. Q.5C

| DocID | Zones | | |
|---|---|---|---|
| | Author | Title | Body |
| $d_1$ | Tagore | Gitanjali by Tagore | Tagore wrote the poem Gitanjali in 1912. |
| $d_2$ | William | Creator of Gitanjali | Rabindranath Tagore won Nobel prize for Gitanjali in 1913. |