Reg. No.										

MANIPAL INSTITUTE OF TECHNOLOGY

MANIPAL

(A constituent unit of MAHE, Manipal)

VI SEMESTER B.TECH. (INFORMATION TECHNOLOGY/COMPUTER AND COMMUNICATION ENGINEERING) MAKEUP EXAMINATIONS, JUNE 2019

SUBJECT: PROGRAM ELECTIVE III- INFORMATION RETRIEVAL [ICT 4006]

REVISED CREDIT SYSTEM

(18 /6/2019)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

- Answer **ALL** the questions.
- Missing data may be suitable assumed.

1A. Explain various dictionary compression techniques with examples. Also discuss the limitations of each compression technique. **5**

1B. Compute the edit distance between the two strings ALGORITHM and ALTRUISTIC **3**

1C. Consider an information need for which there are 5 relevant documents in the collection. Their top 10 results are judged for relevance as follows (the leftmost item is the top ranked search result).

N R N R R N R N R N

Compute the Mean Average Precision (MAP) of the system. **2**

2A. Consider a query (q) and a document collection consisting of three documents. Rank the documents using vector space model. Assume tf-idf weighing scheme.

q: "six ten eleven"

d_1: "nine eight six two seven one eleven"

d_2: "nine eight six three seven one five"

d_3:"four eight ten two seven one ten eleven"

Note: List the vector elements in alphabetical order. **5**

2B. Table Q.2B shows how two human judges rate the relevance of a set of 12 documents to a particular information need (0 = nonrelevant, 1 = relevant). Let us assume that an IR system has been developed which for a query returns the set of documents {4, 5, 6, 7, 8}.

Table Q.2B

Doc Id	1	2	3	4	5	6	7	8	9	10	11	12
Judge 1	0	0	0	1	1	0	1	1	0	0	0	1
Judge 2	1	1	1	1	0	0	0	0	1	0	0	1

i. Calculate the kappa measure between the two judges.

ii. Calculate precision, recall, and F_1 of the system if a document is considered relevant if either judge thinks it is relevant. **3**

2C From the following sequence of γ -coded gaps, reconstruct first the gap sequence and then the postings sequence: 1101111011111000111010101111111011011 **2**

3A. With a neat diagram, explain the distributed architecture of a web crawler **5**

3B. Consider the three documents (d_1, d_2 d_3)

d_1="pen drive damaged in fire"

d_2="Tom Cruise delivers the pen drive."

d_3=" Tom Cruise at MI bureau"

and the query q="pen drive"

Assume that the search engine uses term-frequency weighting scheme. Using Rocchio method, find reformulated query after two iterations. Assume α =1 β=1 and γ=1. Relevant and Non-Relevant document sets are as below- D_r ={ d_1, d_2 } D_{nr} ={d_3}

Note: Ignore the stop words- the, in, at. (List the vector elements in alphabetical order). **3**

3C. What is Boolean retrieval model? Consider the following document collection.

Doc 1: new home sales top forecasts

Doc 2: home sales rise in july

Doc 3: increase in home sales in july

Doc 4: july new home sales rise

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

Answer the query: home AND sales AND (july OR rise). **2**

4A. What is singular value decomposition (SVD)? Find SVD for the following matrix.

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

5

4B. Consider a query (q) and a document collection consisting of 3 documents. Rank the documents using probabilistic model.

q: "bear cow tiger"

d_1: "cat elephant horse cat"

d_2: "cat lion cow deer"

d_3:"cat deer elephant" **3**

4C. Write an algorithm for Blocked Sort-Based Indexing. **2**

5A. Consider a web graph with three nodes 1, 2, 3 and 4. The links are as follows: 1→2, 2→1, 2→3, 2→4, 3→2, 3→4, 4→2and 1→3. Compute PageRank after six iterations for each of the four pages. Assume that at each step of the PageRank random walk, we teleport to a random page with a probability 0.2. **5**

5B. Which ideas can be exploited to reduce the space for storing URL links in adjacency table of a connectivity server? **3**

5C. What do you understand by the term Shingling? Why is it used in web search? **2**