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

II SEMESTER M.TECH. (COMPUTER NETWORKING AND ENGINEERING / SOFTWARE ENGINEERING) END SEMESTER EXAMINATIONS, APRIL/MAY 2019
SUBJECT: PROGRAM ELECTIVE I - INFORMATION RETRIEVAL [ICT 5237]
REVISED CREDIT SYSTEM
(29/04/2019)

Time: 3 Hours   MAX. MARKS: 50

Instructions to Candidates:
❖ Answer **All** questions.
❖ Missing data, if any, may be suitably assumed.

1A. Consider the following 5 documents (Assume stemming and stop-word removal are not required) :
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
    (i) Write the term-document incidence matrix.
    (ii) Construct an inverted index.
    (iii) Outline how the query "happy AND run AND pet" can be processed in an optimal way.   **5**

1B. Shown below is a portion of a positional index in the following format:
term: doc1: position1, position2,. . . ; doc2: position1, position2, . . . ; etc.

angels: 2: 36, 174, 252, 651; 4: 12, 22, 102, 432; 7: 17;
fools: 2: 1, 17, 74, 222; 4: 8, 78, 108, 458; 7: 3, 13, 23, 193;
fear: 2: 87, 704, 722, 901; 4: 13, 43, 113, 433; 7: 18, 328, 528;
in: 2: 3, 37, 76, 444, 851; 4: 10, 20, 110, 470, 500; 7: 5, 15, 25, 195;
rush: 2: 2, 66, 194, 321, 702; 4: 9, 69, 149, 429, 569; 7: 4, 14, 404; to: 2: 47, 86, 234, 999; 4: 14, 24, 774, 944; 7: 199, 319, 599, 709;
tread: 2: 57, 94, 333; 4: 15, 35, 155; 7: 20, 320;
Process the following queries by using the positional index method.
    (i) " fools rush in"
    (ii) "fools rush in" AND "angels fear to tread"   **3**

1C. Is the following statement true or false? Justify.
"In a Boolean retrieval system, stemming never lowers the recall".   **2**

2A. Compute the Levenstein edit distance between the strings "intention" and "execution".   **5**
2B. Compute variable byte and γ codes for the postings list <777, 17743, 294068, 31251336>. Use gaps instead of docIDs where possible. Write binary codes in 8☐bit blocks.   **3**
2C. Write any two differences between Single-Pass-In-Memory (SPIM) Indexing and Blocked-Sort-Based-Indexing (BSBI). What is the largest number that can be stored in 4 bytes using unary encoding?   **2**

3A. Compute the ranks for the 5 documents given in Q.1A using tf-idf model for the query "run happy". Show all the steps.   5

3B. 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 in the Table Q.3B, 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.

Table Q.3B

	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

3

3C. Describe the inexact top k document retrieval.   2

4A. Consider Table. Q.4A, 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.

Table.Q.4A

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   5

4B. Find the SVD for the matrix $C = \begin{bmatrix} 5 & 5 \\ -1 & 7 \end{bmatrix}$

3

4C. What is a zone index? Give one example.   2

5A. What is low rank approximation? Write the three-step procedure to find low rank-k approximation to the matrix C. Consider the vocabulary for the inverted index consisting of the 6 terms: elite elope ellipse eloquent eligible elongate; Assume that the dictionary data structure used for this index stores the actual terms using dictionary-as-a-string storage with front coding and a block size of 3. Show the resulting storage of the above vocabulary of 6 terms. Use the special symbols * and ◇   5

5B. What is web crawler? Explain the architecture of web crawler with neat diagram.   3

5C. Explain different pricing model adopted in advertisement on web pages.   2