Reg. No.										

MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
A Constituent Institution of Manipal University

VII SEMESTER B.TECH.(INFORMATION TECHNOLOGY/COMPUTER & COMMUNICATION ENGINEERING)

MAKEUP EXAMINATIONS, DECEMBER 2016

SUBJECT: PROGRAM ELECTIVE-III INFORMATION RETRIEVAL [ICT 429]

REVISED CREDIT SYSTEM
(30/12/ 2016)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

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

1A**.** Consider the following THREE documents:

d_1 =”To do is to be. To be is to do.”
d_2=”To be or not to be. I am what I am.”
d_3 =”I think therefore I am. Do be do be do.”

Build a local association cluster for each of following vocabulary term consisting of two cluster elements.

Vocabulary = { to, do, or, I, am }. (5)

1B. Consider the following set of documents.

d_1 = “a quick brown big dog”
d_2 = “dog quick a brown”
d_3 = “un chien quick brown”
d_4 = “un chien big brun rapide”
Q = Brown, Dog, big.

Assume $d_{1,}$ d_2 are relevant documents and d_3, d_4 are non-relevant documents. According to probabilistic model, find the similarity of documents with respect to query Q and rank them. (take log to the base 10) (3)

1C. Calculate the edit distance for the strings – AUTOMATA and AUTOCRAFT. (2)

2A. Consider the following text-

“Brown fox saw lazy sleeping dog under the tree in a hot summer. The fox jumped over the lazy dog”

Assume block size is 4 words. The vocabulary and value of each term in the vocabulary is given in the Table Q. 2A.

Table Q. 2A.

Vocabulary Term	Brown	Fox	Lazy	Dog	Tree	hot	Summer
Value	234	568	123	521	427	628	325

Hash function for generating signature is – ***Value mod 23*** of 5 bits length.

i) Create a signature file for the given text. (5)

ii) Search the word- '*Summer*' in the file according to signature file search method and identify the block to which it belongs. Value for the word '*Summer'* is *325.*

2B. What is the significance of user-oriented measures for retrieval evaluation? Discuss two types of user-oriented measures with an example. (3)

2C. Briefly, describe the appropriateness of Recall and Precision measures. (2)

3A. Consider the text- **ISRO-SPACECRAFT** and pattern-**SPACE**. Write steps for matching the pattern with given text using *Shift-OR* method and mention the string matching position. (5)

3B. Consider the collection of documents D.
$D=\{d_1,d_3,d_5, d_9,d_{12},d_{14}, d_{15}, d_{18}, d_{19}, d_{20}, d_{25}, d_{27}, d_{35}, d_{45},d_{55}, d_{61}, d_{70}, d_{72}, d_{75}, d_{80}, d_{88}, d_{91}\}$. A query Q is executed on D. The set $R = \{d_{14}, d_{20}, d_{25}, d_{61}, d_{70}, d_{88}, d_{91}\}$ is set of documents which are relevant to the query Q. Let $A= \{d_1,d_5, d_{14},d_{18},d_{19},d_{25},d_{35},d_{27}, d_{61},d_{75}\}$ be the set of documents retrieved by an IR system in response to Q. Calculate E-measure E(9) and F-measure F(9) for the parameter b=5. (3)

3C. Write any four disadvantages of Boolean model. (2)

4A. Create a suffix tree and suffix array for the DNA sequence - "GATCGCGGCGTATCCG$". Describe the steps to search a suffix DNA sequence TCCG in suffix array. (5)

4B. Consider the five documents (d_1, d_2 d_3, d_4, d_5), query *q* and corresponding *tf-idf* term weights given in the Table Q. 4B. Using Rocchio method, calculate query *q* weights after two iterations of query reformulation. Among the given document set, Relevant document set is D_r={ d1, d2 d3} and Non-Relevant is D_{nr} ={d4, d5}. Assume α=1 ,β=0.75 & γ=0.15.

Table Q. 4B.

Terms & Vectors	**d_1**	**d_2**	**d_3**	**d_4**	**d_5**	**q**
Bangalore	0.9	0	0	0	0	0
Japan	0	0.8	0	0.71	0.75	1
Tokyo	0	0	0	0.71	0.71	0
Macao	0	0	1	0	0	0.6
NewYork	1	0	0	0.2	0	0
Sidney	0	1	0.2	0	0	0.8

(3)

4C. Describe following features of web crawler.
i) Robustness **ii)** Politeness **iii)** Freshness **iv)** Quality (2)

5A. Generate code for the numbers 36, 43 according to following variable-length coding scheme. (use log to the base 2)
i) Elias γ ii) Elias δ iii) Golumb for b=3 (5)

5B. Consider the symbols and their probabilities as per the Q. 5B table.

Table: Q. 5B.

Symbol	Probability	CDF
a_1	0.7	0.7
a_2	0.1	0.8
a_3	0.2	1.0

String to encode is: “$a_1a_1a_2a_2a_3a_1$”. Generate the tag value for the given string using Arithmetic Encoding technique. (3)

5C. Describe Zipf’s law and derive the expression for frequency of first most occurring English word in terms of number of words in a text collection and harmonic number $H_v(\alpha)$. Also find the frequency of 100th ranking word, if 1st ranking word frequency is 50000. Assume $\alpha = 1$. (2)

6A. Explain basic web-crawler architecture and its functioning along with neat diagram. (5)

6B. Discuss three main types of structural query with an example. (3)

6C. Briefly describe *Rocchio* query expansion, term reweighing method and its advantages for vector model. (2)

****