Reg. No.										

MANIPAL INSTITUTE OF TECHNOLOGY

MANIPAL

A Constituent Institution of Manipal University

V SEMESTER B.TECH. (INFORMATION TECHNOLOGY/COMPUTER AND COMMUNICATION ENGINEERING) END SEMESTER EXAMINATIONS, NOVEMBER 2017

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

REVISED CREDIT SYSTEM
(27/11/2017)

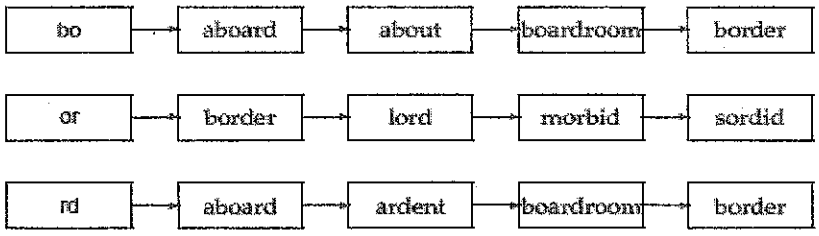
Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

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

1A. Explain the following dictionary compression techniques by giving an example for each.
(i) Array of fixed-width entries
(ii) Dictionary as a string
(iii) Dictionary as a blocked storage 5

1B. Compute the Jaccard coefficients between the query *bord* and each of the terms given in Fig. Q.1B that contain the bigram *or*.



3

1C. Consider a two-word query. For one term the postings list consists of the following 17 entries: [4,6,10,12,14,16,18,20,22,32,47,81,120,122,157,180,200], and for the other it is the one entry postings list: [47]. How many comparisons would be done to intersect the two postings lists with the following two strategies. Justify the answer.
(i) Using standard postings lists
(ii) Using postings lists stored with skip pointers from 4 to 14, 14 to 22, 22 to 120, 120 to 200 for the first posting list

2

2A. Consider the following four documents:
d_1: new car loans top predictions. d_2: car loans hike in November.
d_3: increase in car loans in November. d_4: November new car loans hike.
(i) Draw the term-document incidence matrix.
(ii) Write the inverted index representation.
(iii) Write the algorithm for intersecting two posting lists.
(iv) Find the results for the query "loans AND NOT new" using the posting lists. 5

2B. Write and explain the algorithm for inversion of a block in single-pass-in-memory (SPIM) indexing. 3

2C. Write the algorithm for computing the weighted zone score from two posting lists. 2

3A. Explain various components of a complete search system with a neat diagram. 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. Find the reformulated query using Rocchio method. Assume α =1, β=1 and γ=1. Relevant document set = D_r ={ d_1, d_2 } and Non-Relevant document set= D_{nr} ={d_3}

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

3C. Explain the different pricing models adopted in advertisement on web pages. 2

4A. Consider a web graph with four nodes 1, 2, 3 and 4. The links are as follows: 1→2, 2→1, 2→3, 3→2, 4→3, 4→2, 4→1 and 1→3. Compute the 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 0.4. 5

4B. Consider the following set of documents d_1, d_2, d_3 & query q.

d_1: Manipal University is synonymous with excellence in higher education.

d_2: Every institute has world class facilities and pedagogy. MIT is a constituent institute of MU. Chemnitz University of Technology, Germany is one of Partner University for student exchange programs.

d_3: MIT, Manipal has MOUs with Jack F Welch Technology Center (GE) Bangalore for faculty consultation, research, student projects.

q: Manipal Institute of Technology.

Assume d_2, d_3 are relevant documents and d_1 is non-relevant document. Rank the documents using the Probabilistic Model With respect to query q.

Note: Ignore stop words-of, is, for, to, in, has, with and. (Use log to the base 10). 3

4C. What is low rank approximation? Explain how low rank approximations to the given term-document matrix is computed. 2

5A. Consider a query (q) and a document collection consisting of three documents. Rank documents using vector space model. Assume tf-idf weighing scheme.

q: Bhadra Ghataprabha Malaprabha

d_1: Krishna Godavari Bhadra Yamuna Narmada Ganga

d_2: Krishna Godavari Bhadra Caveri Narmada Ganga Malaprabha

d_3: Godavari Ghataprabha Caveri Narmada Ganga Ghataprabha Malaprabha

Note: List the vector elements in alphabetical order. 5

5B. What is singular value decomposition (SVD)? Find SVD for the following matrix.

$$\begin{bmatrix} 2 & 4 \\ 4 & 4 \\ 4 & 2 \end{bmatrix}$$

3

5C. What is web crawler? Explain the architecture of a web crawler with a neat diagram. 2