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, JAN. 2017

SUBJECT: ELECTIVE I - INFORMATION RETRIEVAL [ICT 4006]

REVISED CREDIT SYSTEM
(07/01/2017)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

- Answer **ALL** questions.
- Missing data if any, may be suitably assumed.

1A. Describe briefly the various steps in determining the vocabulary of terms. **5**

1B. Given a query q, where the relevant documents are d5, d15, d21, d22, d32, d40, d45, and d60. An IR system retrieves the following ranking: d5, d3, d21, d36, d30, d45, d80, d28, d23, d12, d15. Calculate the precision and recall values at each retrieved document for this ranking. Plot a precision versus recall curve after interpolating the precision values at the standard recall levels **3**

1C. Explain the following with an example for each. **2**

i. permuterm index
ii. k-gram index

2A. Explain the posting file compression techniques with example. **5**

2B. Consider the postings list < 4,10,11,12,15,62,63,265,268,270,400 > with a corresponding list of gaps < 6,1,1,3,47,1,202,3,2,130 >. Assume that the length of the posting list is stored separately, so the system knows when a postings list is complete. Using a variable byte encoding: **3**

i. What is the largest gap you can encode in 1 byte?
ii. What is the largest gap you can encode in 2 bytes?
iii. How many bytes will the above postings list require under this encoding?

2C. We have defined unary codes as being "10": sequences of 1s terminated by a 0. Interchanging the roles of 0s and 1s yields an equivalent "01" unary code. When this 01 unary code is used, the construction of a γ code can be stated as follows: (1) Write G down in binary using $b = \lfloor \log_2 j \rfloor + 1$ bits. (2) Prepend $(b - 1)$ 0s. (i) Encode the numbers 307 and 819 in this alternative γ code. **2**

3A. Consider a query (q) and a document collection consisting of three documents. Rank the documents using vector space model. Assume tf-idf weighing scheme. **5**

q: "pink gray purple"
d_1: "magenta white pink green black red purple"
d_2: "magenta white pink blue black red yellow"
d_3: "orange white gray green black red gray purple"

3B. Omar has implemented a relevance feedback web search system, where he is going to do relevance feedback based only on words in the title text returned for a page (for efficiency). The user is going to rank 3 results. The first user, Jinxing, queries for: **3**

banana slug

and the top three titles returned are:

banana slug Ariolimax columbianus
Santa Cruz mountains banana slug
Santa Cruz Campus Mascot

Jinxing judges the first two documents Relevant, and the third Not Relevant. Assume that Omar's search engine uses term frequency but no length normalization nor IDF. Assume that he is using the Rocchio relevance feedback mechanism, with $\alpha = \beta = \gamma = 1$. Show the final revised query that would be run. (Please list the vector elements in alphabetical order).

3C. Consider an information need for which there are 4 relevant documents in the collection. Their top 10 results are judged for relevance as follows (the leftmost item is the top ranked search result). **2**

N R N N R R R N R N

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

4A. What is singular value decomposition (SVD) ? Decompose the following matrix into its SVD components. **5**

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

4B. Below is a table showing how two human judges rated the relevance of a set of 12 documents to a particular information need (0 = nonrelevant, 1 = relevant). Let us assume that you've written an IR system that for this query returns the set of documents {4, 5, 6, 7, 8}. **3**

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

i. Calculate the kappa measure between the two judges.
ii. Calculate precision, recall, and F_1 of your system if a document is considered relevant if either judge thinks it is relevant.

4C. Compute the edit distance between two strings "trials" and "zeil". **2**

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

5B. Explain the process of computing the hub score and authority score for a query. **3**

5C. Consider a web graph with three nodes 1, 2 and 3 with α = 0.5. The links are as follows: 1 → 2, 3 → 2, 2 → 1, 2 → 3. Construct the transition probability matrix. **2**