Reg. No.										

MANIPAL INSTITUTE OF TECHNOLOGY

MANIPAL

A Constituent Institution of Manipal University

VI SEMESTER B.TECH. (COMPUTER SCIENCE AND ENGINEERING)

MAKEUP EXAMINATION, JUNE/JULY 2017

SUBJECT: ELECTIVE III - INFORMATION RETRIEVAL [CSE 4008]

REVISED CREDIT SYSTEM

Time: 3 Hours 22-06-2017 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. Given an inverted index and a query, what solutions to determine whether each query term exists in the vocabulary. **2**

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

2B. Write the algorithm for BSBI and SPIMI. **3**

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: "gold silver truck"
d1: "Shipment of gold arrived in a truck."
d2: "Shipment of gold damaged in a fire."
d3: "Delivery of silver arrived in a silver truck."

3B. The following list of Rs and Ns represents relevant (R) and nonrelevant (N) returned documents in a ranked list of 20 documents retrieved in response to a query from a collection of 10,000 documents. The top of the ranked list (the document the system thinks is most likely to be relevant) is on the left of the list. This list shows 6 relevant documents. Assume that there are 8 relevant documents in total in the collection. **2**
R R N N N N N N R N R N N N R N N N N R N N N N R
i. What is the precision of the system on the top 20?
ii. What is the F1 on the top 20?

3C. Suppose that a user's initial query is **cheap CDs cheap DVDs extremely cheap CDs**. The user examines two documents, $d1$ and $d2$. She judges $d1$, with the content ***CDs cheap software cheap CDs*** relevant and $d2$ with content ***cheap thrills DVDs*** no relevant. Assume that we are using direct term frequency (with no scaling and no document frequency). There is no need to length-normalize vectors. Using Rocchio relevance feedback, what would the revised query vector be after relevance feedback? Assume *alpha*= 1, *beta*= 0.75, *gamma*= 0.25. **3**

4A. Write and explain the Bernoulli model. **5**

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. Give the difference between text centric and data centric in XML retrieval. **2**

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

5B. Write basic feature selection and k – means algorithm. **3**

5C. Explain the process of computing the hub score and authority score for a query. **2**