Reg. No.										

MANIPAL INSTITUTE OF TECHNOLOGY

MANIPAL

A Constituent Institution of Manipal University

VI SEMESTER B.TECH. (COMPUTER SCIENCE AND ENGINEERING)
END SEMESTER EXAMINATIONS, APRIL/MAY 2017
SUBJECT: ELECTIVE III - INFORMATION RETRIEVAL [CSE 4008]
REVISED CREDIT SYSTEM

Time: 3 Hours 29-04-2017 MAX. MARKS: 50

Instructions to Candidates:

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

1A. If the list lengths are m and n, the intersection takes O (m+n) operations. Can we do better than this? Can we usually process postings list intersection in sublinear time? What solution you will provide for given problem. We have two word query. For one term the postings list consists of the following 16 entries: [4,6,10,12,14,16,18,20,22,32,47,81,120,122,157,180] and for the other it is the one entry postings list: [47].Work out how many comparisons would be done to intersect the two postings lists with the following two strategies. Briefly justify your answers: **5**

i) Using standard postings lists.
ii) Using postings lists should with skip pointers, with the suggested skip length of square root of length.

1B. An IR system returns 3 relevant documents, and 2 irrelevant documents. There are a total of 8 relevant documents in the collection. What is the precision of the system on this search, and what is its recall? **3**

Instead of using recall/precision for evaluating IR systems, we could use accuracy of classification. Consider a classifier (non-ranking IR system) that classifies documents as being either relevant or non-relevant. The accuracy of a classifier that makes c correct decisions and i incorrect decisions is defined as: c/(c+i).

i) Why do recall and precision measures reflect the utility (i.e., quality or usefulness) of an IR system better than accuracy does?
ii) Suppose that we have a collection of 10 documents, and two different boolean retrieval systems A and B. Give an example of two result sets, Aq and Bq, assumed to have been returned by the system in response to a query q, constructed such that Aq has clearly higher utility and a better score for precision than Bq, but such that Aq and Bq have the same scores on accuracy.

1C. What are the several ways to build spelling correction algorithms on the basis of computations of proximity? **2**

2A. With an example, explain distributed indexing. For n=2 and 1<=T<=30, perform a step by step simulation of LMergeAddToken () and LogarithmicMerge () algorithm. Create a table that shows, for each point in time at which T=2 * k tokens have been processed (1<=k<=15), which of the three indexes I0,…,I3 are in use. The first three lines of the table are given below: **5**

	I3	I2	I1	I0
2	0	0	0	0
4	0	0	0	1
6	0	0	1	0

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. How to estimate and distribute the number of terms. **2**

3A. 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, $\alpha = \beta = \gamma = 1$.

3B. 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.

3C. Show the final revised query that would be run. (Please list the vector elements in alphabetical order).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"

4A. Derive a ranking function for query terms. **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. Compute the edit distance between two strings "INTENTION" and "EXECUTION". **2**

5A. Draw a neat diagram, for distributed architecture of a web crawler. **2**

5B. Write the algorithm for scoring documents with SIMNoMerge and K-means clustering algorithm. **3**

5C. Write and explain Navie Bayes' classification algorithm. **5**