Reg. No.										

MANIPAL INSTITUTE OF TECHNOLOGY

MANIPAL

(A constituent unit of MAHE, Manipal)

VI SEMESTER B.TECH. (INFORMATION TECHNOLOGY/COMPUTER AND COMMUNICATION ENGINEERING) END SEMESTER EXAMINATIONS, APRIL/MAY 2019

SUBJECT: PROGRAM ELECTIVE III- INFORMATION RETRIEVAL [ICT 4006]

REVISED CREDIT SYSTEM

(3/5/2019)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

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

1A. Write and explain single pass in memory indexing algorithm with suitable example. **5**

1B. What is query optimization? Explain with suitable example. **3**

1C. What is low rank approximation? Explain the construction of low rank approximations of the given term-document matrix. **2**

2A. Consider a web graph with three nodes 1, 2, 3 and 4. The links are as follows: 1→2, 1→3, 1→4, 2→3, 3→4, 3→2, 4→3 and 4→2. Compute 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**

2B. 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:

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). **3**

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

3A. What is singular value decomposition (SVD)? Find SVD for the following matrix.

$$\begin{bmatrix} 4 & 0 \\ 3 & -5 \end{bmatrix}$$

5

3B. Explain with suitable diagram, how priority and politeness is implemented in URL Frontier. **3**

3C. What is the need for robots.txt file in a web server? **2**

4A. Consider a query (q) and a document collection consisting of three documents. Rank the documents using vector space model. Assume tf-idf weighing scheme.

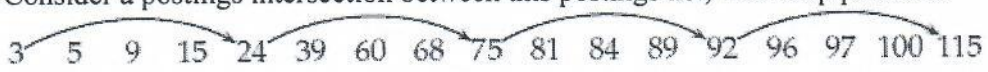
q: "karnataka punjab tripura"
d_1: "nagaland manipur karnataka bihar kerala assam tripura"
d_2: "nagaland manipur karnataka goa kerala assam haryana"
d_3:"gujarath manipur punjab bihar kerala assam punjab tripura"
Note: List the vector elements in alphabetical order. **5**

4B. Consider an information need for which there are 5 relevant documents in the collection. Their top 10 results are judged for relevance as follows (the leftmost item is the top ranked search result).

R N R N N R N N R R.
Draw the 11 point interpolated precision-recall graph. **3**

4C. Explain the different categories into which web search queries can be grouped. **2**

5A. Write an algorithm for posting list intersection with skip pointers.
Consider a postings intersection between this postings list, with skip pointers:



and the following intermediate result postings list (which hence has no skip pointers):

3 5 89 95 97 99 100 101

Trace through the postings intersection algorithm and answer the following queries.

i. How often is a skip pointer followed (i.e., $p1$ is advanced to *skip* ($p1$))?
ii. How many postings comparisons will be made by this algorithm while intersecting the two lists?
iii. How many postings comparisons would be made if the postings lists are intersected without the use of skip pointers? **5**

5B. Consider a query (q) and a document collection consisting of 3 documents. Rank the documents using probabilistic model.

q: "black green yellow"
d_1: "black blue green blue magenta"
d_2: "blue orange purple blue"
d_3:"blue red green magenta" **3**

5C. Explain the following with an example for each.

i. permuterm index
ii. k-gram index **2**