Reg. No. | | | | | | | | | | |

MANIPAL INSTITUTE OF TECHNOLOGY

MANIPAL

A Constituent Institution of Manipal University

II SEMESTER M.TECH.(SOFTWARE ENGINEERING) END SEMESTER EXAMINATIONS, APRIL 2017

SUBJECT: PROGRAM ELECTIVE I - INFORMATION RETRIEVAL [ICT 5237]

REVISED CREDIT SYSTEM
(25/04/2017)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

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

1A. Consider the following documents: 5

doc1 : phone ring person happy person
doc2 : dog pet happy run jump
doc3 : cat purr pet person happy
doc4: life smile run happy
doc5 : life laugh walk run run

Given the query *happy person smile*, rank the documents outlined above using Vector space model. Use term frequency for weighting the document and query terms.

1B. Construct the inverted index required for ranked retrieval for the five documents given in Q.1A. Assume that no stemming or stop-word removal is required. 3

Relating to the sample documents above, outline how the processing of the following Boolean query can be optimised: *happy AND run AND pet* 2

1C. What do you understand by the term *Shingling*? Why is it used in web search?

2A. Table Q.2A shows the output of two information retrieval systems on the same two queries in a competitive evaluation. The top 15 ranks are shown. Crosses correspond to a document which has been judged relevant by a human judge; dashes correspond to irrelevant documents. There are no relevant documents in lower ranks. 5

(i) Explain the following evaluation metrics and give results for query Q1 for both systems.
Precision at rank 10. Recall at precision 0.5.

(ii) Give the formula for mean average precision (MAP), and illustrate the metric by calculating Systems MAP.

(iii)For each system, draw a precision-recall curve. How could one create more informative curves?

2B. Consider the following documents. 3

Doc 1: whale, sea, sea, whale, boat, boat, boat, boat, boat
Doc 2: whales, sea, sea, water
Doc 3: whale, water, water, whale, whale
Doc 4: whales, whales, whales

(i) Construct the term document matrix under the assumption that the terms are not stemmed.
(ii) Construct the corresponding document document matrix.

2C. What role does stemming play in automatic indexing? Briefly describe the principles behind the Porter Stemmer. 2

Table Q.2A

System 1

Rank	Q1	Q2
1	-	X
2	X	-
3	X	-
4	X	-
5	-	-
6	-	-
7	-	-
8	X	-
9	X	-
10	X	-
11	X	-
12	-	-
13	-	X
14	-	X
15	X	-

System 2

Rank	Q1	Q2
1	X	X
2	X	-
3	X	-
4	-	X
5	X	X
6	X	-
7	-	-
8	-	-
9	-	-
10	-	-
11	X	-
12	X	-
13	-	-
14	-	-
15	X	-

3A. Discuss the various dictionary compression and Postings Compression techniques with suitable examples. 5

3B. Define Edit distance. Show how dynamic programming can be used to calculate the edit distance between *able* and *belt*. 3

3C. From the following sequence of γ -coded gaps, reconstruct first the gap sequence and then the postings sequence: 1101111100011101010111111101101111011 2

4A. Find Singular Value Decomposition (SVD) for the following matrix. How is this useful in Latent Semantic Indexing? 5

$$\begin{bmatrix} 5 & 5 \\ -1 & 7 \end{bmatrix}$$

4B. Suppose that a user's initial query is *cheap coats cheap jackets extremely cheap coats*. The user examines two documents, *d*1 and *d*2. She judges *d*1, with the content *cheap coats jackets cheap* relevant and *d*2 with content *cheap blazers* nonrelevant. Assume that we are using term frequency. Using Rocchio relevance feedback equation, what would the revised query vector be after relevance feedback? Assume $\alpha = 1$, $= 0.75$, $\gamma = 0.25$. 3

4C. Table Q.4C shows how two human judges rated the relevance of a set of 12 documents to a particular information need (0 = nonrelevant, 1 = relevant). Assume an IR system which for this query returns the set of documents {3, 4, 5, 6, 7}. 2

Table Q.4C

Doc Id	1	2	3	4	5	6	7	8	9	10	11	12
Judge 1	1	0	1	1	1	1	1	1	1	0	0	1
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 the system if a document is considered relevant if both judges agree.

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 $\alpha = 0.5$. The links are as follows: $1 \rightarrow 2$, $3 \rightarrow 2$, $2 \rightarrow 1$, $2 \rightarrow 3$. Construct the transition probability matrix. 2