Reg. No.										

MANIPAL INSTITUTE OF TECHNOLOGY, MANIPAL 576104
(Constituent College of Manipal University)

SECOND SEMESTER M.TECH(SOFTWARE ENGG.) DEGREE END SEMESTER EXAMINATIONS MAY-2016
SUBJECT: ELECTIVE II-INFORMATION RETRIEVAL(ICT-568)

TIME: 3 HOURS **10/05/2016** **MAX. MARKS: 50**

Instructions to candidates

- Answer **ANY FIVE FULL** questions.
- All questions carry equal marks.
- Assume any missing data suitably.

1A. Consider the documents and query Q as given below(consider *in* and *for* as stopwords):

Q :new quantum cryptography approach
D1: breakthrough discovery in quantum mechanics
D2: new quantum cryptography algorithm
D3: new approach for encoding data using quantum cryptography
D4: new hopes for sensitive data users

i. Compute term weights and document weights using the tf-idf weighting scheme.
ii. Rank the documents using the vector space model.

D3, D2, D1, D4

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.

1C. Given a query vector $\vec{q}$ and a document vector $\vec{d}$ in the vector space model. Suppose the similarity between $\vec{q}$ and $\vec{d}$ is 0.08. Suppose we interchange the full contents of the document with the query, that is, all words from $\vec{q}$ go to $\vec{d}$ and all words from $\vec{d}$ go to $\vec{q}$. What will now be the similarity between $\vec{q}$ and $\vec{d}$? Explain your answer.

[5+3+2]

2A. Compute the edit distance between the strings *s1= played* and *s2 = licked.* Write down the array of distances between all prefixes as computed by the dynamic programming algorithm for computing edit distance between string s1 and s2.

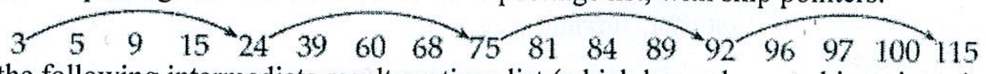
Ans: 4

2B. Describe following techniques for processing wildcard queries:
i. Permuterm index
ii. K-gram index

2C. If the query is: *friends AND romans AND (NOT countrymen)*, how could we use the frequency of *countrymen* in evaluating the best query evaluation order? In particular, propose a way of handling negation in determining the order of query processing.

[5+3+2]

3A. Write an algorithm for postings 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., *p*1 is advanced to *skip*(*p*1))? only once
ii. How many postings comparisons will be made by this algorithm while intersecting the two lists? 19
iii. How many postings comparisons would be made if the postings lists are intersected without the use of skip pointers? 19

(2+1+1)

3B. 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:

i. What is the largest gap you can encode in 1 byte? $2^7 - 1 = 127$

ii. What is the largest gap you can encode in 2 bytes? $2^{14} - 1 = 16383$

iii. How many bytes will the above postings list require under this encoding? 12

3C. Shown below is a portion of a positional index in the format: term: doc1:<position1,position2, . . .>; doc2:<position1, position2, . . . >; etc.

angels: 2: <36,174,252,651>; 4: <12,22,102,432>; 7: <17>;
fools: 2: <1,17,74,222>; 4: <8,78,108,458>; 7: <3,13,23,193>;
fear: 2: <87,704,722,901>; 4: <13,43,113,433>; 7: <18,328,528>;
in: 2: <3,37,76,444,851>; 4: <10,20,110,470,500>; 7: <5,15,25,195>;
rush: 2: <2,66,194,321,702>; 4: <9,69,149,429,569>; 7: <4,14,404>;
to: 2: <47,86,234,999>; 4: <14,24,774,944>; 7: <199,319,599,709>;
tread: 2: <57,94,333>; 4: <15,35,155>; 7: <20,320>;
where: 2: <67,124,393,1001>; 4: <11,41,101,421,431>; 7: <16,36,736>;

Which document(s) if any match each of the following queries, where each expression within quotes is a phrase query?

i. "fools rush in" 2, 4, 7

ii. "fools rush in" AND "angels fear to tread" 4

[5+3+2]

4A. Consider a web graph with three nodes 1, 2 and 3. The links are as follows: 1→2, 1→3, 2→3, 3→2. Compute PageRank after three iterations, hub and authority scores for each of the three pages. Also give the relative ordering of the three nodes for each of these scores indicating any ties. Assume that at each step of the PageRank random walk, we teleport to a random page with a probability 0.1, with a uniform distribution over which particular page we teleport to.

4B. 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. 000000001, 11111111

i. Encode G = 511 and G = 1025 in this alternative γ code. 0000000001, 0000000001

ii. Show that this method produces a well-defined alternative γ code in the sense that it has the same length and can be uniquely decoded.

4C. Discuss the method of Block Sort Based Indexing(BSBI) to reduce the number of disk seeks during sorting.

[5+3+2]

5A. Describe various techniques for reducing the document search space (that is value of $|A|$ such that $K < |A| \ll N$) for score computation for the ranking purpose.

5B. 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: $\left(\frac{1}{2}, 2, 0, \frac{1}{2}, 0, 0, \frac{1}{2}, 0, 2\right)$

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.)

5C. The following pairs of words are stemmed to the same form by the Porter stemmer. Which pairs would you argue shouldn't be conflated? Give your reasoning.

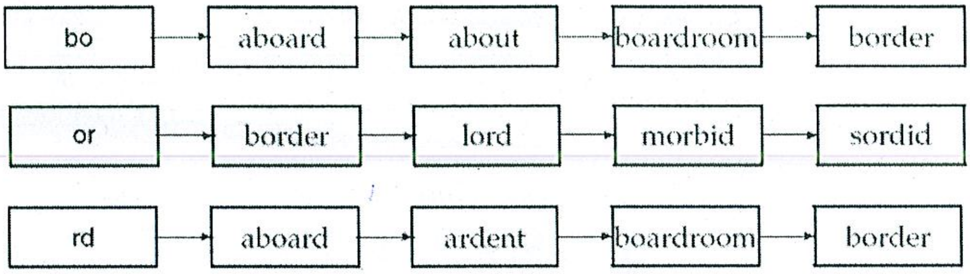
i. abandon/abandonment
ii. absorbency/absorbent
iii. marketing/markets
iv. university/universe

[5+3+2]

6A. With a suitable diagram, explain basic crawler architecture.

6B. Vector space representation is unable to cope with two classic problems arising in natural language processing: *synonymy* and *polysemy.* Explain how Latent Semantic Indexing (LSI) helps to address these problems.

6C. Compute the Jaccard coefficients between the query *bord* and each of the terms in Figure Q.6C that contain the bigram *or.*



2, D1, D4

Fig. Q.6C

[5+3+2]

$\frac{3}{5}$ $\frac{2}{4}$ $\frac{1}{7}$ $\frac{2}{6}$