INFORMATION RETRIEVAL
(CSEN 6137)

Time Allotted : 3 hrs **Full Marks : 70**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and any 5 (five) from Group B to E, taking at least one from each group.

Candidates are required to give answer in their own words as far as practicable.

Group – A
(Multiple Choice Type Questions)

1. Choose the correct alternative for the following: **10 × 1 = 10**

(i) Which of the following is not a correct entry corresponding to the permuterm index of the term "hello"?
(a) hello$ (b) ello$h (c) hel$lo (d) o$hell.

(ii) When we have indexed all the text documents in a collection, would we need to do smoothing?
(a) Yes, there will still be unseen words in individual documents
(b) No point in smoothing if we have indexed all terms
(c) No smoothing required if we have already done stemming
(d) Smoothing would be required only in clustering.

(iii) To compress index, we can use:
(a) lossy compression
(b) lossless compression
(c) lossy compression for posting lists
(d) lossy or lossless depending on storage limitations.

(iv) What would happen if the dimension of latent space in Latent Semantic Analysis is decreased?
(a) It would decrease recall
(b) It has no effect on recall
(c) It would increase recall
(d) Lowers recall and interpolated precision.

(v) Stopping and Stemming
(a) Would reduce the size of an index being constructed
(b) Would have no impact of the index being constructed

(c) Stopping would decrease, and Stemming would increase the size of the index being constructed
(d) Stopping would increase, and Stemming would decrease the size of the index being constructed.

(vi) To be able to answer phrase queries, the following should be in place for Inverted indices
(a) the tf-idf score
(b) the position of the term in every document
(c) skip pointers
(d) list of other terms appearing after this term.

(vii) Can we run through the inverted index intersection in time O(m+n), where m and n are the length of the postings lists for the respective terms?
Information OR NOT retrieval
(a) Yes, this is the union of the posting lists of information and retrieval
(b) No, this essentially must go over all the documents in the index
(c) No, it is bounded by O(mn)
(d) Yes, this is the difference between the posting lists of information and retrieval.

(viii) A document is said to be relevant to a query if:
(a) the terms in the query are present in the document
(b) the terms in the query are present in consecutive positions in the document
(c) the terms in the query are present in the doc with high frequency
(d) none of the above.

(ix) Which of the following is not correct while using Relevance Feedback (RF) in case of web search engines?
(a) RF is hard to explain to common users
(b) It improves the recall but ordinary users hardly see any benefit of improving recall
(c) Most users want to finish web interaction with minimal repeat interaction
(d) RF expands the query and makes the search process slower.

(x) In a corpus of n documents, one document is randomly picked. The document contains a total of T terms and the term “data” appears K times. What is the correct value for the product of term-frequency and inverse-document-frequency, if the term “data” appears in approimately one-third of the total documents?
(a) KT*Log(3) (b) K*Log(3)/T
(c) T*Log(3)/K (d) Log(3)/KT.

Group – B

2. Assume the following fragments comprise your document collection:
Doc1: whale, sea, sea, whale, boat, boat, boat, boat, boat

A user submitted the query "revenue down". Compute the rank of D1 and D2 using Maximum likelihood estimation unigram model and a linear interpolation smoothing. (Use $\lambda = 0.5$)

(b) Look at the following Table, where Documents are represented as Vectors, and their tf-idf scores are listed. For the first four documents their classification is also given. You need to Classify the Query Document D5.

Document	China	Japan	Tokyo	Macao	Beijing	Shanghai	Classification
D_1	0	0	0	0	1	0	C
D_2	0	0	0	0	0	1	C
D_3	0	0	0	1	0	0	C
D_4	0	0.71	0.71	0	0	0	C'
D_5	*0*	*0.71*	*0.71*	*0*	*0*	*0*	*?*

(i) Use Rocchio Classification Method to classify the document D_5. Show your working clearly.

(ii) For the same problem, this time use K-Nearest Neighbours method to classify the document D_5. Again, clearly show your working. (Hint: Choose your K judiciously)

(iii) "KNN is called a Lazy Learner". Would you consider Rocchio to be lazy learner too? Justify your answer.

6 + (3 + 3) = 12

Group – E

8. (a) Compare top down vs bottom up hierarchical clustering. What type is HAC (Hierarchical Agglomorative Clustering)?

(b) Explain single link and complete link Clustering concepts as used by the HAC algorithm with suitable diagrams. How is the dendogram concept used in HAC?

(c) Perform a 2-means clustering to convergence for the points below.
A: (1, 1), B: (1, 2), C: (4, 5), D: (6, 7)
Start with the two seeds A and B. For each iteration give (i) the coordinates of the centroids and (ii) the assignments of points to centroids.

(2 + 2) + (2 + 2) + 4 = 12

9. (a) We know, Singular Value Decomposition SVD of a matrix *A* can be written as $A = U \Sigma V^T$.

Given $A = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}$

Find the Eigen Values for the above matrix A.
Hence, show the matrix Σ.
Now, find U and V.

(b) Can SVD techniques be used in Hierarchical Clustering? If so, then how can the Eigen Values help?

(4 + 1 + 5) + 2 = 12

Doc2: whales, sea, sea, water
Doc3: whale, water, water, whale, whale
Doc4: whales, whales, whales

Assume that you do not Stem. Assume articles and prepositions are stop words.

(i) Construct the *Term-Document Incidence Matrix* for the above documents which can be used in *Boolean Retrieval*.

(ii) Show the vector that will be returned from the above *Term-Document Incidence Matrix* for the query:
(whale AND *sea)* AND NOT *(whales* AND *sea)*
And hence state which documents match the query.

(iii) Now, modify the Matrix to show the Term Frequency (tf) and Inverse Document Frequency (idf) of each term for the above documents.

(iv) For the Terms and Documents above, return the documents according to their Rank if you were to use raw tf and idf values (without log), for the query phrase:
Whale Boat in the Sea

(v) What are the disadvantages of using a term-document incidence matrix? How can it be overcome?

(vi) Draw the inverted index representation for the collection given.

(6 × 2) = 12

3. (a) (i) What is the advantage of using skip pointers with posting lists?

(ii) Explain with an example the intersection of two posting lists, pointing out in particular the advantage of using the skip pointer.

(iii) Will there be any problem if one posting list with skip pointer undergoes intersection with another posting list without skip pointers?

(b) How can permuterm index be used for wildcard queries?

(c) Describe the dynamic programming based approach to calculate the edit distance between two strings.

(1 + 4 + 1) + 3 + 3 = 12

Group – C

4. (a) Mention in what way BSBI and SPIMI indexing schemes are different. Also mention their relative merits and demerits.

(b) Write the equations describing Heap's law and Zipf's law. Explain various symbols used in these equations and mention some typical values used for

the constants. Do the values output by these equations converge in the long run?

(c) Why is term frequency not used alone but we have the document frequency as well? Why do we use inverted df? What is the significance of the tf-idf score?

(3 + 1) + (3 + 1) + (1 + 1 + 2) = 12

5. (a) Define the terms "precision" and "recall" in the context of information retrieval. Explain the quantitative relationship between Relevance, precision / recall and True/False positives/negatives.

(b) Explain the meaning of Relevance Feedback (RF). In what way it is related to Pseudo RF?

(c)

	Doc1	Doc2	Doc3
car	27	4	24
auto	3	33	0
insurance	0	33	29
best	14	0	17

Table A

term	df_t	idf_t
car	18,165	1.65
auto	6723	2.08
insurance	19,241	1.62
best	25,235	1.5

Table B

Table A contains the term frequency values for some terms in documents Doc1, Doc2 and Doc3. *Table B* shows a set of precalculated values of df and idf for the very same set of terms. Using these two tables find out the tf-idf weight and Term Vectors.

(2 + 3) + (1 + 1) + 5 = 12

Group – D

6. (a)

	docID	words in document	in $c = China$?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

Using the Naïve Bayesian classifier verify that the fifth document i.e. the test set belongs to c = China.

(b) What problem does the Laplace smoothing technique solve in case of Naive Bayesian classification of documents? Mention the technique used and the rationale behind it.

6 + (3 + 3) = 12

7. (a) Use Query Likelihood Model to predict how the two documents will be ranked for the given query:
D1: xyzzy reports a profit but revenue is down
D2: Quorus narrows quarter loss but revenue decreases further