M.TECH/CSE/3RD SEM/CSEN 6137/2021

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) What is the **Soundex Code** for *Roheet*?
(a) R30 (b) R30T (c) 30T (d) R300.

(ii) 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

(iii) In case of **Ranked Retrieval**, if you keep looking at more results down the line
(a) Precision can only increase (b) Precision can only decrease
(c) Recall can only increase (d) Recall can only decrease.

(iv) Suppose edit distance between two strings s and t is denoted by d(s,t). Then which of the following is correct?
(a) d(s,t) <= |s| + |t| (b) d(s,t) <= max(|s| , |t|)
(c) d(s,t) >= min(|s| , |t|) (d) None of the above.

(v) Page A points to Page B. Page B points to Page C.
Page C points to page B. Page C points to page D.
Which page would be the best **Authority**?
(a) A (b) B (c) C (d) D

(vi) Which of the following is false in case of Phrase Queries?
(a) Biword indexes can give rise to false positives
(b) Positional indexing can be quite efficient
(c) Longer phrase indexes can expand the vocabulary enormously
(d) Longer phrase indexes can be very efficient in false positive handling and hence is the most adopted approach

(vii) 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.

(viii) In kNN classification, the decision boundary is generated
(a) globally (b) locally
(c) both globally as well as locally (d) neither globally nor locally.

(ix) Which of the following is not correct while using Relevance Feedback (RF) in case of web search engines?
(a) It improves the recall by allowing users to take time to review their initial search results
(b) RF expands the query and makes the search process slower
(c) Image search systems generally allow only positive feedback
(d) Even though it increases the number of true positives, it deteriorates precision by increasing total number of positive predictions to a larger amount.

(x) $A = U\ \Sigma\ V^T$ signifies that Σ is a
(a) Identity Matrix (b) Eigen Vector
(c) Singular Matrix (d) Eigen Value Matrix.

Group – B

2. Consider you have two Posting Lists as shown:
BRUTUS: 2 → 4→6→8→10→12→14→16→18→20→22→24→26→28→30→32
CAESAR: 28
(i) How many comparisons would be needed for a query *(BRUTUS* **AND** *CAESAR) u*sing **Lists Augmented with Skip Pointers**. Show your working, and the sequence of comparisons made to arrive at the result. (Assuming Standard Skip Lengths show the lists augmented with Skip Pointers) [(CO3) (Learn/IOCQ)]
(ii) Consider a **Permuterm Index** for **Wildcard Queries**. Give an example of a term that falsely matches the wildcard query br**s*. (Assume the actual indexed term is *brutus*) [(CO3) (Learn/IOCQ]
(iii) Find the **Levenshtein Edit Distance** between the terms *BRUTUS* and *CAESAR*. Show the **Backtrack** and corresponding **Alignment**. [Insertion/Deletion Cost = 1, Replacement Cost = 2] [(CO4) (Apply/IOCQ)]
(iv) For the technique you used, what is the **Time Complexity** of finding Minimum Edit Distance for two terms, one of size *m* and one of size *n*? [(CO1) (Identify/LOCQ)]

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

3. (a) Explain with an example the intersection of two posting lists, pointing out in particular the advantage of using the skip pointer. Will there be any problem if one posting list with skip pointer undergoes intersection with another posting list without skip pointers? [(CO3) (Understand/LOCQ)]
(b) Compute the Levenshtine distance between the strings “god” and “dog” using the Dynamic Programming based approach. Clearly mark the output of your steps. [(CO4) (Apply/IOCQ)]

(c) Define the term "Relevance". How do you evaluate the effectiveness of an IR system? Mention two popular standard test collections you know about. [(CO2)(Remember/LOCQ)]

(3 + 1) + 4 + (1 + 1 + 2) = 12

Group – C

4. (a) In a given corpus, what would be the fraction of words that are expected to appear more than 3 times and why? [(CO2) Understand/LOCQ]

(b) Why cosine similarity is preferred over Euclidian distance in Vector Space Models? Explain with the help of an example. [(CO2) Understand/LOCQ]

(c) How does stemming typically affect recall in Boolean document retrieval? Justify your reasoning with a simple illustration. [(CO2) Understand/LOCQ]

Two tea testers have been given 4 Tea samples to taste. Below is a snapshot of their observations, where L=Like, N=Do not Like.

	Tea Sample 1	Tea Sample 2	Tea Sample 3	Tea Sample 4
Tester 1	L	L	L	N
Tester 2	L	N	L	L

(d) Do the two testers agree between themselves? Give a measure of this agreement. [(CO4) (Apply/IOCQ)]

2 + 2 + 2 + 6 = 12

5. (a) What are the full forms and significance of the terms BSBI and SPIMI? What are the advantages of keeping a dynamically growing posting list in the SPIMI scheme? [(CO4) (Remember/LOCQ)]

(b) 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? [(CO1) (Remember/LOCQ)]

(c) A term "xyz" appears approximately only in 1/p-th of a set of N documents. A document is chosen at random from this set. The term "xyz" appears K times in this document consisting of T terms in aggregate. What is the tf-idf score for "xyz"? Show the variation of this score for values of p ranging from 5% to 20% in a diagram. You may assume any suitable values for K, T and N. [(CO4)(Analyze/IOCQ)]

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

Group – D

6. (a) Briefly explain the concept of kNN classifier and how it can be applied in case of classifying documents. [(CO5) (Remember/LOCQ)]

(b)

Names	Veterans	Women's Lib	Immigrant Friendly	Support
Anil	9	2	3	REP
Portea	3	10	7	DEM
Kapur	7	3	4	REP
Malik	6	5	5	REP
Yogesh	2	7	10	DEM

Debashis	5	4	8	DEM
Arindam	8	6	9	??

In the above table, data for a few authors are shown. These authors support one of the two ideologies: DEM or REP. They write on similar topics, that include issues related to Veterans, Women's Lib and Immigration Policies. Their coverage levels in these issues in their writings are shown in the various cells in the table in a scale of 1 to 10. Note that for the last author, Arindam, the ideology supported is not yet determined. Use a kNN classifier to determine this missing information, assuming a suitable value of 'k'. [(CO6) (Design/HOCQ)]

(c) In the kNN classification, it is a common practice to compute nearness values using Euclidean distance. How is this related to cosine similarity?
[**Hint**: It might be useful to think with unit normalized vectors; you may assume 2D attributes]. [(CO1)(Analyze/IOCQ)]

(2 + 2) + 5 + 3 = 12

7. (a) 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. What benefit does the Bernoulli model add to the Naive Bayesian classification process? [(CO5) (Analyze/IOCQ)]

(b) Two cricket teams IND and ENG play in the international arena. There are lots of names that are common in reports from both these countries. So it is difficult to identify a report referring to which country, IND or ENG.
However analysts have created the following table for giving us hints as to what country the report could possibly refer to based on some keywords. This is shown in the table below.

DocId	Keywords	Country?
1	Deepak, Ish, Akshar, Afsar	IND
2	Sourav, Rahul, Akshar	IND
3	Ish, Rahul, Kartika	IND
4	Ish,Deepak,Rahul	ENG

Now you have retrieved a report Doc5 where the following keywords are present: Ish,Deepak, Sourav,Akshar
You need to use MLE based estimation with NB classifier to find out whether the document belongs to IND or ENG.

(i) Find out the apriori probalility of a document to belong to IND.
(ii) Find out the conditional probabilities of the terms needed to classify Doc 5.
(iii) Use NB classifier to find out whether the report refers to IND or not. [(CO5) (Design/HOCQ)]

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

Group – E

8. (a) What are some of the issues with using the K-means algorithm for clustering? In what way K-medoid algorithm is better? Comment on the suitability of K-means in case of IR systems. [(CO3) (Analyze/IOCQ)][(CO6)(Analyze/IOCQ)]

(b) In a certain library there are two shelves to keep books on Physics (PHY) and Chemistry (CHE). There are six books to be sorted in these two shelves. The

proportion of PHY, CHE and BIO (Biology) content for each book is given in the table below.

	PHY	CHE	BIO
A	0.12	0.52	0.36
B	0.09	0.33	0.58
C	0.52	0.33	0.15
D	0.17	0.23	0.60
E	0.23	0.06	0.71
F	0.70	0.14	0.16

Use the K-means algorithm to find out which books will go to CHE shelf and which ones to PHY shelf. As the starting seeds, use the ones with maximum content in that subject.

Ex: F has highest content of PHY among all books, so this should be starting seed for PHY cluster.

The stopping criteria is there is no change in the cluster assignment in two successive iterations. [(CO4) (Apply/IOCQ)]

(c) Outline an algorithm to perform the same as in (b) but this time the number of shelves to sort these books into are not known in advance. You do not have to show the computation, but you have to show that your algorithm is indeed doing a "good" sorting. You may assume suitable optimization criteria. [(CO4)(Experiment/HOCQ)]

(2 + 1 + 1) + 4 + 4 = 12

9. (a) What is Pagerank algorithm? What is the significance of random walk and teleporting in case of pagerank? Describe the steps to compute pagerank of a web page. [(CO6) (Remember/LOCQ)]

(b) Prove the Matrix Diagonalization theorem. How can you extend the above theorem to cover the case for a symmetric square matrix?

In general, a term incidence matrix is not symmetric where the number of terms are many orders higher than the number of documents present. How can the idea of diagonalization be applied in such a case? Give an outline of your approach, no proof required. [(CO1) (Design/HOCQ)]

(c) The following table shows the term incidence matrix of six terms in three documents.

Term¤	Doc1¤	Doc2¤	Doc3¤
T1¤	1¤	1¤	1¤
T2¤	0¤	1¤	1¤
T3¤	1¤	0¤	1¤
T4¤	1¤	1¤	0¤
T5¤	0¤	0¤	1¤
T6¤	1¤	0¤	0¤

An SVD decomposition of the matrix C (from the above table) is given below:

C= $U\sum V$

It is known that one of the eigenvalues of C^TC is 2.0.

Find $\sum$. Which one of U or V is easier to compute and why? [(CO3)(Experiment/HOCQ)]

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

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	34%	42%	24%

Course Outcome (CO):

After the completion of the course students will be able to

CO1. Identify basic theories and analysis tools as they apply to information retrieval.
CO2. Develop understanding of problems and potentials of current IR systems.
CO3. Learn and appreciate different retrieval algorithms and systems.
CO4. Apply various indexing, matching, organizing, and evaluating methods to IR problem
CO5. Be aware of current experimental and theoretical IR research.
CO6. Analyze and design solutions for some practical problems

*LOCQ: Lower Order Cognitive Question; IOCQ: Intermediate Order Cognitive Question; HOCQ: Higher Order Cognitive Question

Department & Section	**Submission link:**
CSE	https://classroom.google.com/c/MTQxNzc1NTM4NDIx/a/NDYzODM0NTYzMTY4/details