M.TECH/CSE/3RD SEM/CSEN 6137/2020

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) 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

(iii) For a posting list of size *n*, ideal size of *skip length* would be
(a) n (b) n^2 (c)$\sqrt{n}$ (d) n^3

(iv) Assuming Zipf's law, the fraction of words that appear more than 3 times in any fixed corpus is
(a) 15% (b) 20% (c) 25% (d) 30%

(v) Which of the following is correct about growth rate of vocabulary?
(a) Stemming increases it (b) Spelling error reduces it
(c) Case folding increases it (d) None of the above.

(vi) Time Complexity of finding the Edit Distance between two strings of lengths *m* and *n* is
(a) O(m+n) (b) O(mn) (c) O(1) (d) O(m) + O(n)

(vii) The problem with using MLE (Maximum Likelihood Estimation) estimate in Naive Bayesian classifier shows up when
(a) Training dataset is small
(b) A term occurring rarely is to be classified
(c) The estimate is zero for a term-class combination
(d) All of the above.

(viii) Inversion may occur in
(a) Single Link (b) Complete Link (c) Group Average (d) Centroid

(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) 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 above would be the best **Authority**?
(a) Page A (b) Page B (c) Page C (d) Page D

Group – B

2. (a) Build a Positional Index for the terms: *River, Bank, Interest, Rate.*
Which document(s) match the Query: *Bank /k Interest /k Rate* (/k shows proximity search with a value of 2)?

(b) Draw the inverted index that would be built for the following document collection.
Doc 1: new home sales top forecasts
Doc 2: home sales rise in july
Doc 3: increase in home sales in july
Doc 4: july new home sales rise.

(c) (i) Suppose you have forgotten the first name of the famous playwright “Vijay Tendulkar”. How can you make sure you retrieve background information about him and not that of the famous cricketer “Sachin Tendulkar”? [Give an outline only, no detailed explanation is required].
(ii) Why are Skip Pointers not useful for queries of the form x OR y?

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

3. (a) You are given the Query: *Lemon*e*.
(where * represents a wildcard)
Show two different techniques to match wildcards in the middle of a query. Which one of the two would you prefer and why?

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

(c) Compute the Levenshtine distance between the strings “hill” and “heal” using the Dynamic programming based approach. Clearly mark the output of your steps.

(4 + 1) + 3 + 4 = 12

Group – C

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

(b) Why do we use document frequency in addition to term frequency? Why do we use inverted DF? What is the significance of the TF-IDF score?

(c) The following table lists the term frequencies in three documents.

Term	Doc1	Doc2	Doc3
Affection	115	58	20
Jealousy	10	7	11
Gossip	3	1	7

(i) Tabulate the normalized TF values.

(ii) Which two documents are more likely to be written by the same author? Show your computations in detail.

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

5. (a) Define the terms "precision" and "recall" in the context of information retrieval.

(b) Assume there are 500 terms in a collection. The most frequent term occurs 50 times. Use a power law to find the 4th most frequent term. Write and name the law you used to solve the problem.

(c) Suppose that due to equipment malfunction in a hospital, the results of blood tests on a particular day are unreliable for diabetic patients. The hospital would like to contact all diabetic patients, who had undergone any kind of blood tests on that day, with a request to repeat the tests. The hospital uses an information retrieval (IR) system to identify these patients. Suppose that the collection of patients' medical records contains 10000 documents, 150 of which are relevant to the above query. The system returns 250 documents, 125 of which are relevant to the query.

(i) Calculate the *precision* and *recall* for this system, showing the details of your calculations.

(ii) Based on your results from (a), explain what the two measures mean for this scenario. How well does the hospital's IR system work?

(iii) According to the precision-recall trade-off, what is likely to happen, if an IR system is tuned to aim for 100% recall?

(iv) For the given scenario, which measure (precision or recall) is more important? Justify your answer.

(v) What is the F-score for the hospital's IR System?

2 + 3 + 7 = 12

Group – D

6. (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.

(b) What benefit does the Bernoulli model add to the Naive Bayesian classification process?

(c) You are given 100 similar boxes. One of the boxes contains an iPhone. The rest 99 boxes contain junk. You are given an opportunity to pick one such box, and you pick box number 4. After your choice, 98 of the boxes having junk are removed leaving behind box number 4 and box number 80. What is the probability that the iPhone is in box number 80? Clearly show your assumptions of different events in the problem.

(2 + 2) + 2 + 6 = 12

7. (a) The following table shows the probability of occurrence of two image items, "Bike" and "Jacket" in three documents SA (Sports Assorted), MR (Motor Racing) and BD (Biker's Daily).

Items	**Bike**	**Jacket**
Prob of occurrence in docs		
SA	0.02	0.03
MR	0.03	0.02
BD	0.05	0.01

You have searched for both the items using a query engine capable of supporting image searching. The peculiar property of this search engine is that it shows the most likely matched document first, followed by others in rank of likely match. What would be the order of these documents appearing after your search?

(b) Briefly explain the concept of kNN classifier and how it can be applied in case of classifying documents.
How does kNN classifier compare with the Rocchio classifier?

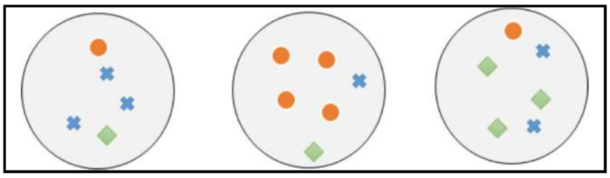
6 + (3 + 3) = 12

Group – E

8. (a) There are six points in a 2-D vector space. These are $d_1(1,2)$, $d_2(2,2)$, $d_3(4,2)$, $d_4(1,1)$, $d_5(2,1)$, $d_6(4,1)$. Use K-means clustering algorithm to find the final clustering outcome for K = 2 when
(i) d_2 and d_5 are chosen as initial seeds
(ii) d_2 and d_3 are chosen as initial seeds
For each case show the steps and final outcome clearly.
What is used as a popular measure for length-normalized vector in case of documents in IR?

(b) From the adjacent figure, calculate the
(i) Rand Index, and

(ii) Purity of the achieved Clustering.



Which one of the two measures would you use to indicate the quality of the achieved clustering? Justify your statement.

(2.5 × 2 + 1) + (3 + 2 + 1) = 12

9. (a) What is Pagerank algorithm? What is the significance of random walk and teleporting in case of pagerank?

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

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

Find a Reduced SVD of the above matrix, such that U and V are square matrices of dimension 2 × 2.

(c) Suppose a librarian wants to arrange a set of 10 documents into two shelves in a library. The shelves are reserved for Science and Arts documents respectively. Each Science document starts with the letter S and are labelled S1 through S6 (meaning there are six of them). The four Arts documents are labelled A1 through A4.

Unfortunately the librarian made some mistakes while arranging these books. He put S1 through S5 in the Science shelf, but put A4 there as well.

A1 through A3 are in the Arts shelf along with S6.

Comment on the librarian's performance in placing documents properly in the two shelves. Is there any other measure on which you want to evaluate him?

(1 + 1) + 6 + (3 + 1) = 12

Department & Section	**Submission Link**
CSE	https://classroom.google.com/c/MTQxNzc1NTM4NDIx/a/Mjg4Mzk3NzQ5NzA2/details