Next, predict the Hub and Authority Score if the same structure was used. In this case, you don't need to show any mathematical working, just a concise argument based on the graph structure will be sufficient. *[(CO5)(Predict/HOCQ)]*

(b) We know the Singular Value Decomposition (SVD) of any matrix $A_{(mxn)}$ can be written as $A_{mxn} = U_{mxm}\Sigma_{mxn}V^T_{nxn}$

For the given matrix $A = \begin{bmatrix} 2 & 4 \\ 1 & 3 \\ 0 & 0 \end{bmatrix}$

Compute the Eigen Values corresponding to the above matrix A. *[(CO3)(Compare, Estimate/HOCQ)]*

(5 + 2) + 5 = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	33.33	47.92	18.75

Course Outcome (CO):

After the completion of the course students will be able to

1. Identify basic theories and analysis tools as they apply to information retrieval.
2. Develop understanding of problems and potentials of current IR systems.
3. Learn and appreciate different retrieval algorithms and systems.
4. Apply various indexing, matching, organizing, and evaluating methods to IR problem
5. Be aware of current experimental and theoretical IR research.
6. Analyze and design solutions for some practical problems.

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

INFORMATION RETRIEVAL
(CSEN 6137)

Time Allotted : 2½ hrs **Full Marks : 60**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and any 4 (four) from Group B to E, taking one from each group.

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

Group – A

1. Answer any twelve: **12 × 1 = 12**

Choose the correct alternative for the following

(i) TP/(TP+FN), (where TP means True Positive, FN means False Negative) represents?
(a) F1 Score (b) Precision (c) Recall (d) Accuracy.

(ii) 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.

(iii) 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.

(iv) 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 document with high frequency
(d) none of the above.

(v) The problem with using MLE (Maximum Likelihood Estimation) estimate in Naive Bayesian classifier shows up when
(a) Training dataset is large (b) A term occurring frequently is to be classified
(c) The estimate is zero for a term-class combination (d) None of the above.

(vi) Which of the following is not a problem when using Maximum Likelihood Estimation to obtain the parameters in a language model?
(a) Out-of-vocabulary items (b) Over-fitting
(c) Smoothing (d) Unreliable estimates when there is little training data.

(vii) While plotting a Precision – Recall Curve if precision increases?
(a) Recall Increases (b) Recall Stays Same
(c) Recall Decreases (d) Has no effect on Recall.

(viii) Time Complexity of finding the Edit Distance between two strings of length m and n?
(a) O(m+n) (b) O(mn)
(c) O(1) (d) O(m) + O(n).

(ix) To support Phrase Queries the size of the Inverted Index?
(a) Needs to increase (b) Needs to decrease
(c) Stay Constant (d) Update after every user query.

(x) A metric derived by taking the logarithm of the total number of documents in a collection divided by the document frequency is called
(a) Document frequency (b) tf-idf weight
(c) inverse document frequency (d) None of (a), (b) & (c).

Fill in the blanks with the correct word

(xi) The situation when a statistical model fits exactly against its training data is called _____.

(xii) The steady state probability of a random walk among web pages dictates the _____.

(xiii) A false _____ is saying something is false when it is actually true.

(xiv) The _____ model which generates an indicator for each term of the vocabulary, indicating either the presence or absence of the same.

(xv) Jaccard Coefficient is defined as _____.

Group - B

2. (a) Explain two important differences between Stemming and Lemmatization with a suitable example. *[(CO1,CO2)(Understand/LOCQ)]*

(b) Consider three documents D1, D2, D3 containing the following three sentences respectively:
D1: This fox is brown
D2: Is the fox brown
D3: fox in brown
Mention the techniques you will apply to return all three documents when you query for the phrase "fox brown". *[(CO1,CO2)(Analyze/LOCQ)]*

(c) The following term-document incidence matrix contains occurrences of some terms used by Shakespeare in his plays. Using the Boolean Retrieval model, find out which plays of Shakespeare contain the words Brutus and Caesar, but not Calpurnia.

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

[(CO3)(Apply/IOCQ)]
4 + 4 + 4 = 12

3. (a) Show an example of two differently spelled proper nouns whose Soundex Codes are the same. Show an example of two phonetically similar proper nouns whose Soundex Codes are different. *[(CO1)(Show/LOCQ)]*

(b) Assume you need to match Wildcard Queries.
Show what terms you will index for the token "Horse".
Show what terms you will index for the token "House".
Using the indices that you just built, suggest a wildcard query which will return a positive match for Horse but not House. Justify your answer. *[(CO2)(Show/LOCQ)]*

(c) What do you mean by Phrase Index. Explain with the help of examples. *[(CO1)(What/LOCQ)]*
(2 + 2) + (2 + 2 + 2) + 2 = 12

Group - C

4. (a) A term "xyz" appears approximately 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)]*

(b) Assume the following fragments comprise your document collection:
Doc 1: Interest in real estate speculation
Doc 2: Interest rates and rising home costs
Doc 3: Kids do not have an interest in banking
Doc 4: Lower interest rates, hotter real estate market
[Assume the following are stop-words: an, and, do, in, not]
Construct the Term-Document Matrix to show the Term Frequency (tf) and Inverse Document Frequency (idf) of each term for the above documents.
For the Term-Document Index you created, which document has the highest score for the query: (Real estate interest rates)? *[(CO4)(Construct, Analyze/IOCQ)]*
4 + (5 + 3) = 12

5. (a) Define the terms "precision" and "recall" in the context of information retrieval. Explain the quantitative relationship between relevance, precision / recall, retrieval amount, and True/False positives/negatives. *[(CO3)(Remember/LOCQ)]*

(b) (i) Why is Relevance Feedback necessary in query evaluation?
(ii) What is the difference between Explicit Relevance feedback and Pseudo Relevance Feedback? *[(CO3,CO4)(Analyze /IOCQ)]*

(c) A document collection contains 3 relevant documents D_1, D_2 and D_3 for a certain query. A system retrieves D_1 at rank 2 and D_2 at rank 5; it does not retrieve D_3. Compute the precision and recall values at
(i) Rank 2 (i.e., when D_1 is retrieved), and
(ii) Rank 5 (i.e., when D_2 is retrieved). *[(CO3,CO4)(Analyze/IOCQ)]*
(2 + 2) + (2 + 2) + (2 + 2) = 12

Group - D

6. (a) Compute the Query Likelihood (P(Q/D)) for the following documents:
- D1: Xerox reports a profit, but revenue is down.
- D2: Lucene narrows quarter loss, but decreases further.

- Query Q: revenue down

Assume that the collections consists of only these two documents.
[Use Jelinek-Mercer Smoothing with $\lambda = 0.5$] *[(CO3)(Apply,Solve/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 probability 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)]
6 + (1 + 3 + 2) = 12

7. (a) In the Figure 1 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.
Use Rocchio Classification Method to classify the document. Show your working clearly.

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	?

Fig.1

[(CO6)(Apply, Solve/IOCQ)]

(b) For the same data, this time use K-Nearest Neighbours method to show two different sets of classifications depending on different values of K.
In general, how will you choose a good value for k? *[(CO6)(Apply, Solve/IOCQ)]*

(c) Why is KNN called a Lazy Learner? Justify whether Rocchio Classification Technique can be considered to be the same. *[(CO5)(Relate,Illustrate/LOCQ)]*
4 + (3 + 3) + 2 = 12

Group - E

8. (a) Figure 2 shows an actual clustering, how many True Negative pairs would be there in the Ideal Cluster scenario? How many True Positive pairs would be there in the Ideal Cluster scenario? What is the Accuracy of the Actual Clustering?

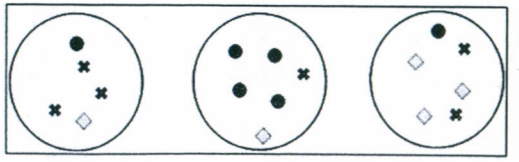

Fig. 2

[(CO4)(Analyze, Compare/IOCQ)]

(b) What are some of the issues with using the K-means algorithm for clustering? In what way K-medoid algorithm is better? *[(CO3,CO6)(Understand/LOCQ)]*

(c) What is a dendrogram? *[(CO6)(What/LOCQ)]*
(3 × 2) + (2 + 2) + 2 = 12

9. (a) Figure 3 shows a toy web graph. Calculate the PageRank score of each page assuming the surfer starts at page A and the probability of typing a URL is 10%. Clearly show all steps to calculate the score after 3 iterations or till score convergence, whichever is earlier.

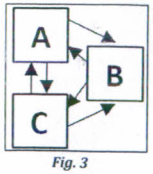

Fig. 3