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) $\frac{TP}{TP+FN}$, (where, TP means True Positive, FN means False Negative) represents
(a) F1 Score (b) Precision (c) Recall (d) Accuracy.

(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) Naive Bayesian Text Classification uses
(a) Supervised Learning (b) Unsupervised Learning
(c) Both (a) and (b) (d) None of the above.

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

(v) 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.

(vi) 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.

(vii) 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.

(viii) 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.

(ix) Latent Semantic Indexing is (choose the correct alternative)
(a) a low-rank approximation to indexing and retrieving documents
(b) has been established as a significant force in scoring and ranking in IR
(c) is not related to clustering of text documents
(d) none of the above.

(x) The steady state probability of a random walk among web pages dictates the _____
(a) Pagerank (b) Teleportation (c) Stochastic Matrix (d) Anchor Text.

Group- B

2. (a) Assume the following fragments comprise your Document Collection:
Doc1: banking on banks to raise the interest rate
Doc2: jogging along the river bank to look at the boats
Doc3: jogging to the bank to look at the interest rate
Doc4: jogging interest along the river Ganga
Assume that you *drop Stop Words*. Assume that you *Stem*.
(i) Construct the **Term-Document Matrix** for the above documents that can be used in **Boolean Retrieval**.
(ii) What documents would be returned in response to the following?
(interest AND jogging) NOT bank [(CO1) (Construct/IOCQ)]

(b) Differentiate between **Bigram Index** and **Phrase Index** with the help of examples. [(CO1) (Differentiate/LOCQ)]

(c) Discuss how you can augment an **Inverted Index with Skip Pointers** for more efficient query processing. [(CO1) (Discuss/LOCQ)]

(3 × 2) + 3 + 3 = 12

3. (a) Calculate the **Levenshtein Distance** between the two terms ***Multiple*** and ***Maple***. Also show the alignment of the two strings. [(CO2) (Analyze/IOCQ)]

(b) Show the **Permuterm Index** for the term ***Choice***. How can you use this index for wildcard matching. [(CO3) (Show /IOCQ)]

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

Group – C

4. (a) Explain the working of the two compression techniques **BSBI** and **SPIMI**. [(CO2) (Explain/LOCQ)]

(b) Suppose the collection of patients' medical records contains **10000 documents**, **150** of which are **relevant for a user query**. The system **returns 250 documents**, **125 of which are relevant** to the query.
(i) Report the ***Precision*** and ***Recall*** for this system, showing the details of your calculations.
(ii) Based on your results from (i), explain what the two measures mean for this scenario. How well would you say that the hospital's information IR system works? [(CO3) (Report/LOCQ)]

(4 + 4) + (2 × 2) = 12

5. (a) Assume the following fragments comprise your document collection:
Doc 1: Argha takes a book to Ankita.
Doc 2. Argha who reads a book helps Ankita.
Doc 3. Whom does Argha think Ankita helps?
Doc 4. Argha thinks cricket is a good sport.
Assume that you drop stop words. Assume that you stem
(i) Prepare the **Vector Space Term-Document Matrix** for the above documents using ***tf-idf*** term weighting. Clearly indicate your final answer.
(ii) Use the above to show the retrieval of the best matching document in response to the **Query:** *help ankita*.
Show the similarity score between the query and the document. (You may take raw ***tf-idf*** scores for ease of calculation). [(CO3) (Prepare/Use/Show/IOCQ)]

(b) Assume there are **500 terms** in a collection. The **most frequent term** occurs **50 times**. Express a power law to **find the 4th most frequent** term. *Write and name the law* you used to solve the problem. [(CO4) (Express/Name/LOCQ)]

(4 × 2) + 4 = 12

Group – D

6. You are given the graph in Fig. 1, which shows two types of data sets, one marked as circles, other as squares.

(i) Add at least one more data point for each data type (circle and square). Then add a point on the graph and mark it as **X.** (Read the next questions before placing **X** on the graph. *Is it possible to place **X** in such a way that it results in different classifications in part (b) and (c) below? If so, then show it in your example*)

(ii) Classify the point **X** using Rocchio Classification Technique. Show your working / justification clearly. (An illustrative example, without exact mathematical calculations is acceptable.)

(iii) Classify the point **X** using **k-Nearest Neighbour (kNN)** Classification Technique.Show your working / justification clearly. (An illustrative example, without exact mathematical calculations is acceptable.) Judiciously choose a value of **k**. [(CO4)(Analyze/Differentiate/IOCQ)]

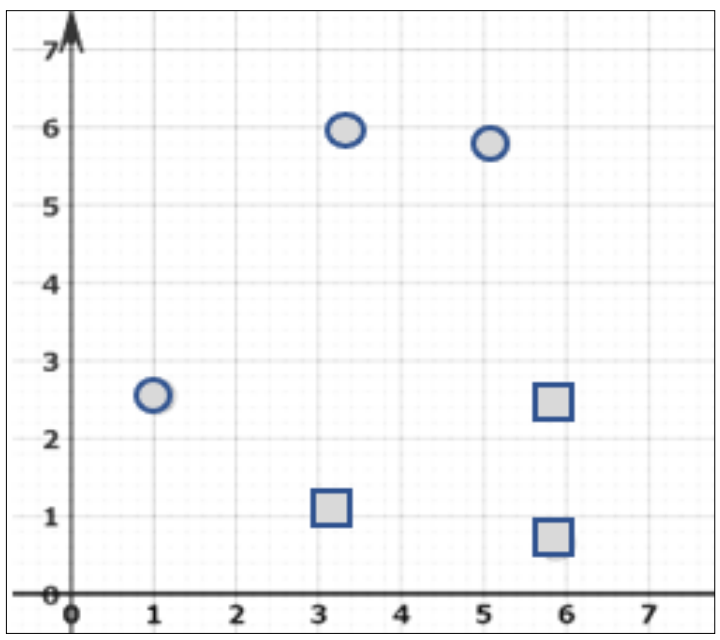

Fig. 1

(3 × 4) = 12

7. (a)

Table 1

Colour	Type	Origin	Stolen?
Red	Sports	Domestic	Yes
Red	Sports	Domestic	No
Red	Sports	Domestic	Yes
Yellow	Sports	Domestic	No
Yellow	Sports	Imported	Yes
Yellow	SUV	Imported	No
Yellow	SUV	Imported	Yes
Yellow	SUV	Domestic	No
Red	SUV	Imported	No
Red	Sports	Imported	Yes

In Table 1, various types of car details are shown. Attributes of a car are colour, type, origin. Subject is: stolen or not.

(i) Find the apriori probability of a car being stolen.

(ii) You want to buy a red domestic SUV. You want to know whether it is a stolen one or not. Find out the conditional probabilities of various attributes needed to classify this car.

(iii) Use NB classifier to find out whether the car you chose is indeed a stolen one or not.

[(CO5) (Design/HOCQ)]

(b) Use **Query Likelihood** model to predict the rank of the following documents:

D1: Humpty Dumpty sat on a wall.

D2: Humpty Sharma ki dulhania.

Query Q: Humpty Dumpty

(i) Use **Jelinek-Mercer Smoothing** with $\lambda = 0.5$

(ii) Now, consider you build **Generative Language Model(s)** $\mathbf{M_1}$ and $\mathbf{M_2}$ respectively for $\mathbf{D_1}$ and $\mathbf{D_2}$ using a **Unigram-Word Model**. Consider a STOP probability of 0.1 after each token is ingested. Justify which model is more likely to generate the above **Query Q**. Clearly show your working.

[(CO5) (Calculate/Solve/IOCQ)]

(1 + 3 + 2) + (3 + 3) = 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,CO6) (Analyze/IOCQ)]

(b) The matrix in figure 4 shows distances between pairs of points. Here, distance is used as the metric to define similarity such that lower values signify more similarity, higher values indicate less similarity.

(i) Predict the clustering of all five points using **Hierarchical Agglomerative Clustering (HAC)**.

(ii) Estimate the result of Clustering using **Complete Link Clustering**. Also, show the corresponding **Dendrogram.**
[(CO6)(Show/Calculate/HOCQ)

	A	B	C	D	E
A					
B	4				
C	6	7			
D	8	9	10		
E	10	11	12	13	

(2 + 2 + 2) + (3 × 2) = 12

9. (a) We know the **Singular Value Decomposition (SVD)** of any matrix $\boldsymbol{A_{mxn}}$ can be written as
$A_{mxn} = U_{mxm}\, \Sigma_{mxn}\, V^T_{nxn}$
For the given matrix $A = \begin{bmatrix} 2 & 4 \\ 1 & 3 \\ 0 & 0 \end{bmatrix}$

(i) Compute the Eigen Values for the above matrix $\boldsymbol{A}$.

(ii) Hence, show the matrix $\boldsymbol{\Sigma}$.

(iii) Now construct the matrices $\boldsymbol{U}$ and $\boldsymbol{V}$. [(CO5) (Compose/Construct/HOCQ)

(b) The following toy **Web Graph** in Fig. 2 consists of 3 webpages **A**, **B**, and **C** showing the in links and out links.

(i) Construct the **Adjacency Matrices** corresponding to **Hubs** and **Authorities** from the given graph.

(ii) Assume the **Initial Hub Weight Vector** is **[1 1 1]**. Predict the **Weight Vectors** for **Hubs** and **Authorities** after two iterations each. [(CO6)(Construct/Predict/HOCQ)]

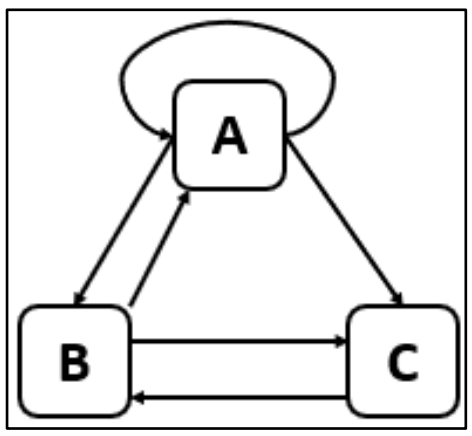

Fig. 2

(3 + 1 + 3) + (1.5 + 3.5) = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	22.92	52.08	25

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