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) 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

(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) Which metric is used to determine the similarity between two documents in the Vector Space Model?
(a) Jaccard Similarity (b) Cosine Similarity
(c) Euclidean Distance (d) Manhattan Distance.

(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) What is the main goal of relevance feedback?
(a) To adjust the weights of terms based on relevance
(b) To identify duplicate documents
(c) To optimize indexing time
(d) To reduce memory usage in indexing.

(vi) Which of the following is an evaluation measure for ranked retrieval?
(a) Purity (b) Rand Index
(c) Precision at k (d) Skip List Efficiency.

(vii) 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.

(viii) In the Query Likelihood Model, document ranking is based on:
(a) Cosine Similarity
(b) The likelihood that the document generates the query
(c) The Euclidean distance from the query
(d) Jaccard Similarity.

(ix) PageRank in web search algorithms measures a page's importance by
(a) Counting keyword occurrences
(b) Calculating incoming link quality and quantity
(c) Checking page length
(d) Assessing page metadata.

(x) Which metric is typically used to evaluate the quality of clustering by measuring intra-cluster similarity and inter-cluster dissimilarity?
(a) Purity (b) Precision at k
(c) F1 Score (d) Rand Index.

Fill in the blanks with the correct word

(xi) Hierarchical agglomerative clustering can be visualized by _____.

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

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

(xiv) An inverted index stores a list of documents for each term, known as the _____ list.

(xv) In the TF-IDF scheme, _____ frequency assigns a higher weight to terms that are less common across the document collection.

Group - B

2. (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) Why can't we use a term document matrix for efficiency reasons? How can the posting list technique help in this process? How can the Boolean retrieval model work in posting lists? *[(CO3)(Analyse/IOCQ)]*

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

3. (a) Define **edit distance** and calculate the edit distance between the words "intention" and "execution" using the Levenshtein algorithm. Show your calculation steps. Backtrack not required. *[(CO4)(Apply/IOCQ)]*

(b) How does stemming typically affect recall in Boolean document retrieval? Explain with the help of an example. *[(CO1)(Explain/LOCQ)]*

8 + 4 = 12

Group - C

4. (a) What is the full form of BSBI? Outline various steps of this algorithm and comment which step takes most of the time to complete the index creation. *[(CO4)(Remember/LOCQ)]*

(b) What does Heap's law signify? Give the formula and explain various terms used in the formula. *[(CO4)(Remember/LOCQ)]*

(c) What is the idf of a term that occurs in every document? What about when it occurs in only one document? Also state the same when it occurs in half of the documents. *[(CO4)(Apply/IOCQ)]*

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

5. (a) Define the term “Relevance”. How do you evaluate the effectiveness of an IR system? Mention two popular standard test collections you know about. *[(CO4)(Apply/IOCQ)]*

(b) 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. *[(CO4)(Remember/LOCQ)]*

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

Group - D

6. (a) Define **k-Nearest Neighbors (kNN)** classification and explain how it determines similarity between documents. Discuss one limitation of kNN for text classification in high-dimensional data. *[(CO2)(Discuss/LOCQ)]*

(b) Describe the **Rocchio Classification** method and its use in information retrieval. Explain how relevance feedback could improve the Rocchio Classification process in text classification. *[(CO2)(Discuss/LOCQ)]*

(c) Given centroid vectors ***C1=[0.5,0.3,0.4]*** and ***C2=[0.2,0.6,0.3]*** and a document vector ***D=[0.4,0.5,0.4***, calculate the cosine similarity between ***D*** and both centroids. Classify the document based on the closer centroid. *[(CO4)(Apply/IOCQ)]*

4 + 4 + 4 = 12

7. (a) Explain the **Naïve Bayes classifier** in the context of text classification. Discuss how the independence assumption affects the model’s accuracy in real-world applications. *[(CO2)(Discuss/LOCQ)]*

(b) Assume we want to categorize computer-science documents into the following categories: Systems, Theory, AI. Consider performing naive Bayes classification with a simple model in which there is a binary feature for each significant word indicating its presence or absence in the document. The following probabilities given in Table 1 have been estimated by analyzing a corpus of pre-classified training documents.

c	Systems	Theory	AI
$P(c)$	0.35	0.40	0.25
P(theorem| c)	0.05	0.8	0.10
P(search | c)	0.30	0.40	0.60
P(heuristic | c)	0.05	0.01	0.50
P(disk | c)	0.30	0.02	0.01
P(data | c)	0.50	0.01	0.20

Table: 1

Assuming the probability of each evidence word is independent given the category of the text, compute the posterior probability for *each* of the possible categories for each of the following short texts:

- *Data on heuristic search for theorem*
- *Search for data stored on disk*

Assume the categories are disjoint and complete for this application. Ignore any words that are not in the table. *[(CO5)(Solve/HOCQ)]*

4 + (4 + 4) = 12

Group - E

8. (a) What is **Latent Semantic Analysis (LSA)**, and how does it address the issue of synonymy and polysemy in text data?
Discuss the concept of low-rank approximations in SVD and their implications for information retrieval. *[(CO3)(Analyse/IOCQ)]*

(b) Given the following term-document matrix ***M***.

Term	**Doc1**	**Doc2**	**Doc3**
A	1	0	1
B	1	1	0
C	0	1	1

Construct the ***Singular Value Decomposition (SVD)*** of ***M*** and explain how you would use it for dimensionality reduction. *[(CO6)(Construct/HOCQ)]*

4 + 8 = 12

9. (a) Consider a small web graph with pages ***A,B,C*** where A links to B and C, and B links back to A, and C links back to B. If the initial hub and authority scores for all pages are 1, calculate the authority and hub scores after two iterations of the **HITS (Hyperlink-Induced Topic Search)** algorithm. *[(CO6)(Analyze/IOCQ)]*

(b) For the same web graph, calculate the **Page Rank** of all the pages. Assume that a teleportation probability is 10%. *[(CO6)(Solve/HOCQ)]*

5 + 7 = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	37.5	38.54	23.96