(c) The balanced F measure is defined as the harmonic mean of precision and recall. What is the advantage of using the harmonic mean rather than "averaging" (using the arithmetic mean)? (5)

******

B.Tech Even Semester Examination, May, 2025

Computer Science & Engineering

(6th Semester)

Course No: CSE-603

(Information Retrieval)

Full Marks: 70

Pass Marks: 28

Time: 3 hours

Note:

1. Attempt one question taking one from each unit.
2. Begin each answer in a new page
3. Answer parts of a question at a place
4. Assume reasonable data wherever required
5. The figure in the right margin indicates full marks for the question
6. All the mathematical symbols and abbreviations have their usual meanings.

UNIT-I

1. a) Define Information Retrieval. 1

   b) What is a term-document incidence matrix? 1

   c) How is a term-document incidence matrix constructed from a corpus of documents? 3

   d) Why the term-document incidence matrix is often considered sparse? 3

   e) Define the followings:i) Term frequency ii) Inverse document frequencyiii) Tf-idf weighting 2*3=6

2. Discuss in detail the three classical Information Retrieval models? (14)

UNIT-II

3. a) Explain the concept of an inverted index in Information Retrieval. Discuss its structure, construction process, and importance. 10

   b) Discuss the storage considerations for postings lists in an inverted index. What data structures can be used for postings lists, and what are their advantages in terms of space and time efficiency? 4

4. (a) Discuss Stemming and Lemmatization in Information Retrieval with examples. (6)

   (b) Differentiate between token, type and term with example. (3)

   (c) Discuss any two text-similarity metrics available in Information Retrieval with example. (5)

UNIT-III

5. (a) Differentiate between text classification and clustering with examples. (4)

   (b) Explain the working of Naïve Bayes classifier with example. (5)

   (c) Briefly explain the working of a Decision Tree algorithm. (5)

6. Explain and compare Agglomerative Clustering and K-Means Clustering algorithms. Discuss their working principles, differences, advantages, and disadvantages with suitable examples. (14)

UNIT-IV

7. Explain the working principles of the PageRank and HITS algorithms used in web search. Compare their approaches to ranking web pages, highlighting their similarities, differences, advantages, and limitations with suitable examples. (14)

8. (a) Discuss Hubs and Authorities. (4)

   (b) Explain the Zipf's law and Porter stemmer algorithm. (5+5=10)

UNIT-V

9. (a) Explain precision, recall and F-measure. (5)

   (b) The database collection consists of three documents (D = 3) with the following content:

   D1: "Shipment of gold damaged in a fire"

   D2: "Delivery of silver arrived in a silver truck"

   D3: "Shipment of gold arrived in a truck"

   Sort and rank the documents in descending order according to the similarity values for the query "**gold silver truck**". (5)

   (c) An IR system returns 8 relevant documents, and 10 non-relevant documents. There are a total of 20 relevant documents in the collection. What is the precision of the system on this search, and what is its recall? (4)

10. (a) Briefly discuss relevance feedback and query expansion. (5)

    (b) Discuss Hamming Distance and Euclidean Distance. (4)