# School of Computer Science Engineering and Technology

Course- BTech
Course Code- CSET228

Type- Specialization Core II
Course Name- Data Mining and Predictive Modelling

Year- 2024-25
Date- 22-01-2025

Semester- Even
Batch- IV Semester (All)

**Lab # No. (4.1)**

**CO Mapping**

| Exp No. | Name | CO1 | CO2 | CO3 |
|---|---|---|---|---|
| **1** | Similarity analysis and vectorization | | ✓ | |

1. The Jaccard Similarity algorithm can be used to determine how similar two objects are. The computed similarity might then be used in a recommendation query. For example, the Jaccard Similarity algorithm can be used to display products purchased by similar customers based on previous purchases.
   (a) Assume one customer transaction table consists of three customers where the items purchased by three customers are being shown according to the following table.

| ID | Item 1 | Item 2 | Item 3 | Item 4 | Item 5 | Item 6 | Item 7 | Item 8 | Item 9 |
|---|---|---|---|---|---|---|---|---|---|
| C1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| C2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| C3 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |

Table 1

In Table 1, Customers with id C1, C2, C3 has been shown whereas the items available are Item 1, Item 2, … etc. Out of the nine available items the customers bought the items according to their requirements. In Table 1, 'an item purchased' is represented by '1' and 'an item not purchased' is represented by '0'.
Calculate the similarity index in reference to the customer requirements between each pair (C1 & C2, C2 & C3, C3 & C1) using Jaccard Similarity procedure.

**Sample Output:**

Similarity - Customer C1 and C2 is 0.25
Similarity - Customer C1 and C3 is 0.5
Similarity - Customer C2 and C3 is 0.0

(b) Assume two sets 'S1' and 'S2', consist of some numerical values. Where
S1 = {0, 2, 5, 7, 9}
S2 = {0, 1, 2, 4, 5, 6, 8}
Calculate the similarity between these two sets using Jaccard Similarity procedure.

**Sample Output:**

Similarity between Set S1 and S2 is 0.33

**(40 Minutes)**

2. There exist two text documents which are being examined to calculate their similarity.The files are as follows:
   Document1: The Data Mining and Predictive Modelling courseDocument2:
   Data Mining course is interesting

   (a) Represent the documents by vectors with the help of creation of a word table fromthe documents.

   (b) Calculate the similarity between these two documents using the Cosine similarity procedure.

   **Sample Output:**

   Similarity between document D1 and D2 is: XXXX

   **(20 Minutes)**

3. Create a text summarization using following operations:
   a. Tokenize the sentences using sent_tokenize() function and create the frequency matrix of the words present in a sentence. (use word_tokenize)
   b. Calculate a term frequency and generate a matrix. Also design a table for documents for each word and calculate idf, tf-idf to generate a matrix.
   c. Calculate sentence score to find threshold and generating summary.

   **(30 Minutes)**