



SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
CONTINUOUS ASSESSMENT TEST - I
WINTER SEMESTER 2025-2026

(Handwritten signature)

Programme Name & Branch : M.Tech. (Integrated) Computer Science and Engineering
Course Code and Course Name : CSI 4004 and Text Mining
Faculty Name(s) : Dr.Dheebea J, Dr. Jagalingam P, Dr.Manikandan G, Dr.Yuvaraj N
Class Number(s) : VL2025260501833, 836, 838, 835
Date of Examination : 01.02.2026
Exam Duration : 90 minutes **Maximum Marks: 50**

General instruction(s):

- Answer All Questions
- M - Max mark; CO - Course Outcome; BL - Blooms Taxonomy Level (1 - Remember, 2 - Understand, 3 - Apply, 4 - Analyse, 5 - Evaluate, 6 - Create)
- Course Outcomes
CO1: Recognize key areas and issues in Information Extraction and Text Summarization.
CO2: Discover interesting patterns using Clustering and Classification techniques.

Q. No	Question	Module	Marks	CO	BL
1.	a) Perform the following tasks in the given sentence: "Apple Inc. was founded by Steve Jobs and Steve Wozniak in Cupertino, California, on April 1, 1976." i. Identify and extract all the named entities in the sentence and classify each entity into its appropriate category. ii. Provide a detailed breakdown of the sentence, highlighting the recognized entities and their respective categories.	1	5	1	3
	b) Discuss the pipeline of Information Extraction with a simple example. Include a diagram to illustrate the process, showing the steps from text input to the storage of the extracted information in a knowledge graph.	1	5	1	2
2.	a) Consider the following two product reviews: R1 - "The smartphone has a great battery life, and the camera is excellent." R2 - "Battery life is good, but the camera quality is poor." Perform the following tasks: i. Create the Bag-of-Words (BoW) representation for the given reviews after removing stop words and count the frequency of each term. ii. Calculate the Cosine Similarity between the two reviews using their BoW vectors and provide an inference based on the similarity score.	1	5	1	4
	b) List the different types of text summarization techniques. Discuss the key differences between extractive summarization and abstractive summarization with examples.	1	5	1	2
3.	Based on the following 5 product reviews, apply the K-medoids clustering algorithm to group them into two clusters based on their content. The reviews are: 1. The smartphone has an excellent battery life and is very fast. 2. The camera quality of the phone is outstanding, but the battery life could be better. 3. This phone is slow, and the battery drains quickly.	2	10	2	4



VIT

Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

REG.NO.:

SLOT: F1+TF1

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
CONTINUOUS ASSESSMENT TEST - I
WINTER SEMESTER 2025-2026

	4. Great camera, but the phone gets hot after prolonged use. 5. The performance of the phone is great, and the battery lasts all day. The key terms identified from the Bag-of-Words (BoW) model are: battery, camera, performance, fast, quality, slow, drains, hot, lasts, disappointed. i. Use Review 1 and Review 3 as the initial medoids for clustering. ii. Assume that the medoids do not change after the first assignment, and the algorithm converges with stable clusters. iii. After clustering, summarize the characteristics of each group based on the content of the reviews assigned to each cluster.																						
4.	Find the clusters and draw the dendrogram for the given set of documents by using Euclidean distance and average link method in agglomerative clustering. <table><tr><th>Doc IDs</th><th>X</th><th>Y</th></tr><tr><td>1</td><td>4.2</td><td>3.8</td></tr><tr><td>2</td><td>5.1</td><td>4.0</td></tr><tr><td>3</td><td>3.9</td><td>3.2</td></tr><tr><td>4</td><td>8.5</td><td>9.1</td></tr><tr><td>5</td><td>9.0</td><td>8.7</td></tr></table>	Doc IDs	X	Y	1	4.2	3.8	2	5.1	4.0	3	3.9	3.2	4	8.5	9.1	5	9.0	8.7	2	10	2	3
Doc IDs	X	Y																					
1	4.2	3.8																					
2	5.1	4.0																					
3	3.9	3.2																					
4	8.5	9.1																					
5	9.0	8.7																					
5.	Given the following three mobile phone reviews, apply the Naive Bayes classification algorithm to categorize them into Positive and Negative based on the sentiment expressed: R1- The phone has excellent performance and a great camera. R2- The phone's battery drains quickly, and it lags while using apps. R3- Amazing camera and smooth performance, but the battery could be better. Instructions: i. Convert the reviews into a Bag-of-Words (BoW) model. Use the terms: phone, performance, camera, lags, battery, smooth, quickly, amazing. ii. Calculate the prior probabilities for Positive and Negative based on the given reviews. For each review, compute the probability of it being Positive or Negative using the term frequencies and classify the review using Naive Bayes classification. iii. Summarize the results by identifying which reviews were classified as Positive and which as Negative and explain how the classification was done.	2	10	2	3																		
