B/G/TY

Reg. No:

VIT
Vellore Institute of Technology

Summer Term Final Assessment Test – July 2025

Course: **CSI4004 - Text Mining**

Class NBR(s): **1092 / 0202**

Slot: **C1+TC1+C2+TC2**

Time: **Three Hours**

Max. Marks: **100**

- **KEEPING MOBILE PHONE/ANY ELECTRONIC GADGETS, EVEN IN 'OFF' POSITION IS TREATED AS EXAM MALPRACTICE**
- **DON'T WRITE ANYTHING ON THE QUESTION PAPER**

General Instructions: Non Programmable calculator is permitted.

Answer ALL Questions

(10 X 10 = 100 Marks)

1. a) What is topic representation and different types of topic representation approaches?

   b) Consider the following set of sequences and find the probabilities for that:

   - The cat purred softly.
   - Dogs bark loudly.
   - Birds fly high.
   - The small bird sang.

   Find if the following is a correct tag or not, based on typical English POS tagging:

   - The cat purred softly (D, N, V, Adv)
   - Dogs bark loudly (N, V, Adv)

2. Consider the following two documents:

   Document 1: "machine learning improves prediction accuracy"

   Document 2: "prediction models use machine learning"

   Compute the cosine similarity between the two documents. Based on the similarity score, discuss whether these documents are likely to fall into the same cluster in a distance-based clustering approach. Explain the different distance based clustering algorithm?

3. Consider the following table presents a dataset of 10 objects, with attributes Color, Type, Origin, and a Class label (Yes/No). Use the Naive Bayes classifier to predict the class of an object based on the given attributes.

Sr. No.	Color	Type	Origin	Satisfied?
1	Red	Casual	Domestic	Yes
2	Red	Casual	Domestic	No
3	Red	Casual	Domestic	Yes
4	Yellow	Casual	Domestic	No
5	Yellow	Casual	Imported	Yes
~~5~~	~~Yellow~~	~~Casual~~	~~Imported~~	~~Yes~~
6	Yellow	Formal	Imported	No
7	Yellow	Formal	Imported	Yes
8	Yellow	Formal	Domestic	No
9	Red	Formal	Imported	No
10	Red	Casual	Imported	Yes

Compute the prior probabilities for each class (Yes/No). Calculate the conditional probabilities for each attribute given the class (Yes and No). Classify a new object with the following attributes: Color = Red, Type = Formal, Origin =Domestic, Use the Naive Bayes classifier formula to predict whether the class is Yes or No.

4. In the context of AI-driven societies and hyper-connected environments, data exploration must evolve from traditional descriptive analysis to predictive and prescriptive insights. Discuss how data exploration can be redefined using future lens.

5. Consider a Global Crisis Monitoring System (GCMS) which uses continuous text streams from multiple real-time sources including news feeds, social media, satellite communication logs, and IoT sensors to detect emerging global issues such as pandemics, climate-related disasters, and geopolitical conflicts. As a data scientist for GCMS you need to improve topic detection in high-velocity text streams. Based on the data.

   (i) Describe the unique challenges involved in performing topic detection in real-time text streams compared to static corpora.

   (ii) Explain the concept of topic evolution and how it can be handled in streaming text environments. Provide one example of an algorithm that supports dynamic topic modeling.

6. Describe the process of detecting and describing events and trends from textual data. Include techniques such as burst detection, temporal topic modeling, and named entity recognition. Explain with real-world examples and discuss how accuracy and relevance are maintained in dynamic text environments.

7. Discuss the various approaches to sentiment analysis, highlighting the challenges in handling sarcasm, domain-specific language, and multilingual data with examples.

8. Evaluate the methodologies used in cross-text analysis, including similarity measures, textual entailment, and clustering. Discuss its applications in areas such as fake news detection, plagiarism identification, and opinion aggregation.

9.a) Develop a content-based spam email classifier for an email system. Explain the classification process using a suitable machine learning algorithm of your choice. Justify your selection based on its effectiveness in handling spam classification. Compare this classifier with any one other classifier in terms

   i) Preprocessing
   ii) Cost-sensitivity
   iii) ability to handle high-dimensional feature sets.

OR

9.b) (i) Explain in detail the various evaluation metrics used in email classification, including the confusion matrix and cost sensitivity. Support your explanation with examples.

ii) Compare the performance of any two classifiers with respect to the PU1 and ZH1 experiments. Highlight how dataset characteristics and language differences affect their performance.

10.a) Develop an multilingual document clustering system for organizing medical research papers written in English, Spanish, and Chinese. Explain how Latent Multilingual Semantic Analysis (LMSA) combined with term alignments can be effectively applied in this scenario. How does LMSA conceptually and functionally differ from traditional LSA and Multilingual LSA (MLSA)?

OR

10.b) a) Discuss the different feature selection methods used for clustering?

b) You are clustering a collection of medical research papers across topics like cancer, infections, and nutrition. Many topic-specific terms appear in certain documents, while generic terms (e.g., study, effect) occur in nearly all. Which feature selection method would you choose to improve clustering quality? Justify your choice by comparing it with at least one other method?

⇔⇔⇔ B/G/TY ⇔⇔⇔