Roll No.												

PRESIDENCY UNIVERSITY

BENGALURU

End Term Examinations – December 2025

Date: 05-12-2025 **Time:** 01:00pm – 04:00pm

School: SOCSE/SOE	**Program:**B. Tech CSE,CAI,CCS,CSG,CSTCDV,CBC,CSD,CBD,ISR,IST,ECE	
Course Code : CAI3427	**Course Name:** Language Models for Text Mining	
Semester: VII	**Max Marks**: 100	**Weightage**: 50%

CO - Levels	**CO1**	**CO2**	**CO3**	**CO4**	**CO5**
Marks	**34**	**16**	**36**	**14**	

Instructions:

(i) Read all questions carefully and answer accordingly.

(ii) Do not write anything on the question paper other than roll number.

Part A

Answer ALL the Questions. Each question carries 2marks. **10Q x 2M=20M**

1.	Define Text Mining.	**2 Marks**	**L1**	**CO1**
2.	Differentiate between Text Mining and NLP	**2 Marks**	**L2**	**CO1**
3.	What is Tokenization? Give an example	**2 Marks**	**L1**	**CO2**
4.	List any four text preprocessing techniques.	**2 Marks**	**L1**	**CO2**
5.	Explain Stop Words with an example.	**2 Marks**	**L2**	**CO2**
6.	What is a Corpus in NLP?	**2 Marks**	**L1**	**CO3**
7.	Define N-gram Language Model.	**2 Marks**	**L1**	**CO3**
8.	What is TF-IDF?	**2 Marks**	**L2**	**CO3**
9.	State the purpose of the Viterbi Algorithm.	**2 Marks**	**L1**	**CO4**
10.	Define Topic Modeling.	**2 Marks**	**L1**	**CO4**

Part B

Answer the Questions. **Total Marks 80M**

11.	**a.**	Explain the difference between Text Mining and Natural Language Processing (NLP) with suitable examples.	**3 Marks**	**L2**	**CO1**
	b.	Describe tokenization and the tokenization process using a simple sentence.	**3 Marks**	**L2**	**CO1**
	c.	Summarize the complete Text Mining pipeline, highlighting each step briefly.	**4 Marks**	**L2**	**CO1**
		Or			
12.	**a.**	Describe the steps involved in text preprocessing (cleaning, normalization, tokenization).	**3 Marks**	**L2**	**CO1**
	b.	Explain stop-word removal and state why it is necessary in text mining.	**3 Marks**	**L2**	**CO2**
	c.	Illustrate the different types of research paradigms in NLP (rule-based, statistical, neural).	**4 Marks**	**L3**	**CO1**

13.	**a.**	Differentiate between stemming and lemmatization with examples.	**3 Marks**	**L2**	**CO2**
	b.	Determine the corpus provide the two examples and explain its role in NLP.	**3 Marks**	**L3**	**CO2**
	c.	Interpret the N-Gram language model with examples for unigram, bigram, and trigram.	**4 Marks**	**L3**	**CO2**
		Or			
14.	**a.**	Write any three PoS tags with examples.	**3 Marks**	**L6**	**CO2**
	b.	Explain how lemmatization identifies the base form of a word using any two examples.	**3 Marks**	**L2**	**CO2**
	c.	Examine the role of PoS tagging in understanding sentence structure. Provide examples showing how PoS tags help in downstream NLP tasks.	**4 Marks**	**L2**	**CO2**

15.	**a.**	Describe its purpose in NLP with an example.	**3 Marks**	**L2**	**CO3**
	b.	Sketch an N-Gram model. Give examples of unigram, bigram, and trigram.	**3 Marks**	**L3**	**CO3**

	c.	Explain the Bag-of-Words (BoW) model in detail. Discuss its structure and two major drawbacks.	**4 Marks**	**L2**	**CO3**
		Or			
16.	**a.**	Apply the concept of TDM to compare the similarity between two documents.	**3 Marks**	**L3**	**CO3**
	b.	Calculate Term Frequency (TF) and Inverse Document Frequency (IDF) in detail. Provide formulas and examples.	**3 Marks**	**L4**	**CO3**
	c.	Explain cosine similarity and measures document similarity using vector representations.	**4 Marks**	**L2**	**CO3**

17.	**a.**	Apply one-hot encoding to the vocabulary {apple, banana, grapes}.	**3 Marks**	**L3**	**CO4**
	b.	Evaluate the advantages of using CBOW for large text corpora.	**3 Marks**	**L5**	**CO4**
	c.	Describe the architecture and working of the CBOW (Continuous Bag of Words) model. Include the role of context window.	**4 Marks**	**L2**	**CO4**
		Or			
18.	**a.**	Given the sentence “AI transforms the world,” apply CBOW and list one context-target pair.	**3 Marks**	**L3**	**CO4**
	b.	Explain how the Skip-gram model captures semantic relationships between words with an example.	**3 Marks**	**L2**	**CO4**
	c.	Describe the role of deep learning models (CNN, RNN, LSTM) in document classification using Keras.	**4 Marks**	**L2**	**CO4**

19.	**a.**	Explain the process of Text Mining in detail. How is it applied to extract meaningful information from unstructured text data?	**10 Marks**	**L3**	**CO1**
	b.	Describe the complete workflow of Text Mining, including Extraction, Preprocessing, Analysis, and Evaluation. Provide a neat diagram.	**5 Marks**	**L2**	**CO1**
	c.	Illustrate different methods of Lexical Resource Creation and evaluate their importance in NLP applications.	**5 Marks**	**L3**	**CO1**
		Or			
20.	**a.**	Design a text preprocessing workflow for a sentiment analysis system.	**10 Marks**	**L3**	**CO2**
	b.	Explain how Tokenization combined with Stop Words Removal can enhance text analysis.	**5 Marks**	**L3**	**CO2**

	c.	Evaluate which encoding is more suitable for deep learning models and justify your answer.	**5 Marks**	**L5**	**CO2**

21.	**a.**	Compare N-gram models with neural language models in terms of accuracy, scalability, and handling of unseen words.	**10 Marks**	**L2**	**CO3**
	b.	Explain the working of a Bag-of-Words model and list any two major limitations.	**5 Marks**	**L2**	**CO3**
	c.	Given simple term counts for two documents, compute TF-IDF for one selected term and interpret the result.	**5 Marks**	**L3**	**CO3**
Or					
22.	**a.**	Compare one-hot encoding and embeddings in terms of semantic understanding.	**10 Marks**	**L4**	**CO4**
	b.	Describe how the CBOW model predicts a target word from its context with a simple example.	**5 Marks**	**L3**	**CO4**
	c.	Explain how Keras uses the Embedding layer in document classification tasks.	**5 Marks**	**L2**	**CO4**