Birla Institute of Technology & Science, Pilani
Work Integrated Learning Programmes Division
First Semester 2025-2026
M.Tech. in AIML

Mid-Semester Test
(EC-2 Regular Paper)

Course No. : AIMLCZG530
Course Title : Natural Language Processing
Nature of Exam : Closed Book
Weightage : 30%
Duration : 2 Hours
Date of Exam : 20-Dec-2025

No. of Pages = 3
No. of Questions = 7

Note to Students:
1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

Question 1. [4 Marks]

You are preparing data from social media comments. Consider the text:
"I'd love 2 go, but I can't. The concert is 100% sold out :("

Perform the following advanced steps one by one, showing the result after each step.
1. Tokenization (Split text into units, handle contractions).
2. Stop Word Removal (Remove common high-frequency words; assume punctuation is kept).
3. Lemmatization (Reduce words to base form).

Question 2. [4 Marks]

A Consider a toy corpus consisting of only two sentences:

1. "read a book"
2. "read a blog"

You want to calculate the probability of the test bigram "read a map".

a) Calculate the Maximum Likelihood Estimation (MLE) probability P (map|a) based strictly on the corpus counts. What problem do you encounter when calculating the total sentence probability? [2 Marks]

b) Apply Add-1 (Laplace) Smoothing to calculate the smoothed probability P_Laplace (map|a). Assume the Vocabulary size V = 5 (read, a, book, blog, map). [2 Marks]

Question 3. [5 Marks]

Consider a Skip-Gram neural language model trained using the **Negative Sampling** optimization technique. The model processes the following training example:

- **Target Word:** "doctor" (w_t)
- **Positive Context Word:** "hospital" (w_c)
- **Negative Samples (k=2):** "car" (w_1 and "banana" (w_2)

Word	Vector
doctor	[1.0,2.0,−1.0]
hospital	[2.0,−1.0,1.0]
car	[−1.0,1.0,0.5]
banana	[0.5,−0.5,−1.0]

Compute the total SGNS loss for this training example assuming k = **2** negative samples.

Question 4. [4 Marks]

Assume that the following is a portion of a vocabulary extracted from a large text corpus:

Vocabulary = ['river', 'mountain', 'ocean', 'forest', 'desert', 'valley', 'climate', … , …]

Using a word-embedding method such as **Word2Vec**, the corresponding portion of the **word-embedding matrix (M)** for the above vocabulary is shown below:

river	3	5	7
mountain	6	4	2
ocean	5	8	6
forest	2	3	5
desert	7	6	1
valley	4	5	3
climate	6	7	6
…	…	…	…

So that,

$$M = \begin{bmatrix} 3 & 5 & 7 \\ 6 & 4 & 2 \\ 5 & 8 & 6 \\ 2 & 3 & 5 \\ 7 & 6 & 1 \\ 4 & 5 & 3 \\ 6 & 7 & 6 \\ \vdots & & \end{bmatrix}$$

(a) Explain the steps required to extract the embedding vector for the word 'desert' and write the resulting vector. [1 mark]

(b) Using the embedding matrix, write the embedding vectors for 'ocean' and 'forest'. How would you compute the semantic distance between 'ocean' and 'forest'? [2 marks]

(c) Using the word embedding vectors, compute the embedding vector representation of the sentence: 'forest Climate Mountain'. Assume simple vector addition. **[1 mark]**

Question 5. [5 Marks]

A customer-support chatbot is being trained to recognize emotions and encounters the sentence "I am happy with the product." To learn whether "happy" should be more closely related to the positive word "joyful" or kept distant from the negative word "sad," the model uses v(happy) = [0.45, 0.12, −0.18], u(joyful) = [0.18, 0.38, 0.20] with target 1, and u(sad) = [−0.08, 0.55, −0.10] with target 0. Compute the two dot products, compute the two sigmoid values, compute sigma minus target for each, and explain in one comment what these error signs tell the chatbot about adjusting the relationships between "happy," "joyful," and "sad."

Question 6. [4 Marks]

Consider the two sentences:

1. **I will book the table.**
2. **I read the book.**

a) In both sentences, the word *"book"* appears, but with different meanings. Explain how a POS tagging system helps prevent mistranslation or semantic confusion by correctly identifying whether "book" functions as a verb or a noun in each context.[3 marks]

b) Which POS tags are used for "nouns" and "verbs" in the Penn Treebank tagset? [1 mark]

Question 7. [4 Marks]

a) Explain the methodology employed when utilizing pre-trained models, such as BERT or RoBERTa, for Part-of-Speech (POS) tagging. [2 marks]

b) Describe the specific technique utilized by Large Language Models (LLMs), like GPT-4, to accomplish the same task and list any one method. [2 marks]