VIT

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

Final Assessment Test(FAT) - Apr/May 2025

Programme	B.Tech.	Semester	Winter Semester 2024-25
Course Code	BCSE409L	Faculty Name	Prof. Krithiga R
Course Title	Natural Language Processing	Slot	D1+TD1
		Class Nbr	CH2024250501963
Time	3 hours	Max. Marks	100

Instructions To Candidates

- Write only your registration number in the designated box on the question paper. Writing anything elsewhere on the question paper will be considered a violation.

Course Outcomes

CO1: Understand the fundamental concepts of Natural Language Processing.
CO2: Develop useful systems for language processing and related tasks involving text processing and demonstrate text-based processing of natural language with respect to morphology.
CO3: Check the syntactic and semantic correctness of natural language.
CO4: Select a suitable language modeling & Feature Representation to develop real-world applications.
CO5: Develop computational methods for real-world applications using deep learning.

Section - I

Answer all Questions (7 × 10 Marks)

01. Identify the type of ambiguity present in each of the following sentences in the context of Natural Language Processing (NLP). Justify your answer by explaining why each sentence represents that specific type of ambiguity. [5x2=10 Marks]
a) Are you sad that we do not have a match?
b) The nation needs more vigorous youngsters.
c) It is not clear why the president addressed the mob with anger.
d) My friend gave her notes.
e) The seal on the letter looks thin.

[10] (CO1/K2)

02. a) For the given list of words — 'hoping,' 'hopped,' and 'hopes' — calculate the total number of characters that are removed when each word is processed using the Porter Stemming algorithm. [6 Marks]
b) If the input word is "hopped" (6 letters), how many letters remain after Step 1b is applied, and what is the measure (m) of the resulting stem? [4 Marks]

[10] (CO1/K2)

03. Use Arc-Eager dependency parsing on the sentence **"The talented master handled intriguing young boy with care"** to build the dependency tree, sticking to the various relation operations between head and dependent tokens. Provide step-by-step explanation for the input buffer, stack and relation operations on this parser. Also, mention the total number of left arcs and right arcs used in this exercise.

[10] (CO2/K3)

04. Nithin read through the below set of sentences:

- The students should read many books.
- Learning from books is enjoyed by the students.
- The teacher encourages the students to read many books.
- Books in the library help students with difficult topics.
- The students approach a library or a teacher for clarifying difficult topics.
- The library has many books for the students.

He used a context window size of ± 3 per sentence, after which he omitted determiners, prepositions, conjunctions, modal/auxiliary verbs, infinitive verbs, main verbs, and gerunds while building the word co-occurrence matrix in the order of their appearances in the above paragraph. Help him in evaluating the cosine similarities of the token 'students' with the remaining tokens in the matrix and getting the highest score out of them.

[10] (CO3/K5)

05. Sana was reading the first paragraph from a children's story that contained 4 training sentences:

"The hen sat on the mat. The goat sat on the mat. The hen and the goat played. The goat chased the hen on the ground."

She wanted to try the N-gram model of this paragraph on a new sentence **"the hen and the goat sat on the ground"**. Help her to compute all case-insensitive bigram logarithmic probabilities (up to 4 decimal points) and thereby evaluate the bigram perplexity score on this new sentence.

Note: Include start and end tags for all bigram calculations as well as token count.

[10] (CO4/K4)

06. Nitesh used a Machine Translation (MT) model to translate a native text snippet in Gujarati to English. He compared the following MT output against the reference translation. Help him in computing the BLEU-2 score with brevity penalty.

Reference Translation:
"The quick brown fox jumps over the lazy dog. It swiftly escapes into the dense forest. The dog watches in surprise but does not chase. The fox is too fast for it. This frequently happens in the wild. Animals must be quick to survive."

Machine Translation Output:
"The fast brown fox leaps over the sleeping dog. It quickly runs into the thick forest. The dog looks in shock but does not follow. The fox is too swift for it. This occurs often in nature. Beasts must be fast to live."

[10] (CO4,5/K3)

07. A news summarization model generates a summary for a given reference text. The reference summary and the generated summary are given below:

Reference Summary:
"The government announced a new economic policy to boost employment and trade."

Generated Summary:
"The government introduced a fresh policy to enhance jobs and commerce."
Perform an exhaustive evaluation of the generated summary against the reference using the **Recall-Oriented Understudy for Gisting Evaluation (ROUGE)** metrics, specifically **ROUGE-1**, **ROUGE-2**, and **ROUGE-L**. Calculate precision, recall, and F1-score separately for each ROUGE metric. Interpret the results, highlighting the quality of the generated summary based on the computed ROUGE scores

[10] (CO4,5/K4)

Section - II
Answer all Questions (2 × 15 Marks)

08. Use lower case for all the tokens and construct the required matrices for applying the Viterbi algorithm in a Hidden Markov Model (HMM) using dynamic programming on the given sample sentences:

"Spot will give book.
Can Tom spot the bat?
Kate will book the tickets.
Will took the bat.
Can Spot drop the can?
Will book the spot.
The can will fall.
Tom will bat for the spot."

Now, determine the most probable sequence of tags for the tokens present in the sentence - **"Spot will book the can".**

[15] (CO2,3/K5)

09. Let's assume you are working on an AI-powered virtual assistant for historians that helps analyse historical texts. One major challenge is resolving ambiguous words to improve the accuracy of text interpretation.
Consider that in historical documents, the word "arm" could mean:
A weapon (e.g., "The king ordered his men to take up arms.")
A body part (e.g., "The soldier was wounded in the arm.")
A branch of an organization (e.g., "The military arm of the resistance was growing.")

a) Apply the Lesk algorithm for Word Sense Disambiguation (WSD) to determine the correct meaning of words based on their surrounding context and the word "arm" in the sentence: **[7 Marks]**
"The general supplied weapons to the arm of the rebellion."

b) Explain step-by-step how the Lesk algorithm determines the correct meaning. **[5 Marks]**

c) List the challenges that might arise when using the Lesk algorithm on historical texts, and how would you address them. **[3 Marks]**

[15] (CO3/K2)

BL-Bloom's Taxonomy Levels - (K1-Remembering,K2-Understanding,K3-Applying,K4-Analysing,K5-Evaluating,K6-Creating)

⇦⇨⇦