

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NOIDA INSTITUTE OF ENGINEERING AND TECHNOLOGY, GREATER NOIDA
(An Autonomous Institute Affiliated to AKTU, Lucknow)

B.Tech

SEM: VI - THEORY EXAMINATION (2025 - 2026)

Subject: Natural Language Processing

Time: 3 Hours

Max. Marks: 100

General Instructions:

IMP: Verify that you have received the question paper with the correct course, code, branch etc.

1. This Question paper comprises of **three Sections -A, B, & C**. It consists of Multiple Choice Questions (MCQ's) & Subjective type questions.

2. Maximum marks for each question are indicated on right -hand side of each question.

3. Illustrate your answers with neat sketches wherever necessary.

4. Assume suitable data if necessary.

5. Preferably, write the answers in sequential order.

6. No sheet should be left blank. Any written material after a blank sheet will not be evaluated/checked.

SECTION-A

1. Attempt all parts:-

- | | | |
|-------|--|----|
| (1.1) | A Named Entity Recognition (NER) system correctly identifies 38 out of 60 named entities. If the system identified a total of 50 entities, calculate: Precision(approximately)(CO1,K2) | 20 |
| (a) | 60% | 1 |
| (b) | 55% | |
| (c) | 75% | |
| (d) | 45% | |
| (1.2) | A parser is used in NLP for:(CO1,K2) | 1 |
| (a) | Removing stop words | |
| (b) | Counting words | |
| (c) | Translating text | |
| (d) | Analyzing grammatical structure | |
| (1.3) | IDF in TF-IDF is calculated as?(CO2,K1) | 1 |
| (a) | TF/IDF | |
| (b) | DF/N | |
| (c) | log(N/DF) | |
| (d) | TF*N | |
| (1.4) | Cosine similarity measures?(CO2,K1) | 1 |
| (a) | Euclidean distance | |
| (b) | Angle between vectors | |
| (c) | Dot product only | |
| (d) | Manhattan distance | |
| (1.5) | Topic modelling helps in discovering .(CO3,K1) | 1 |
| (a) | Hidden themes | |
| (b) | Word spelling errors | |
| (c) | Sentence length | |
| (d) | Data tables | |
| (1.6) | Statistical machine translation depends on ____ data.(CO3,K1) | 1 |
| (a) | Parallel | |
| (b) | Random | |
| (c) | Encrypted | |
| (d) | Numeric only | |
| (1.7) | ELMo produces (CO4,K1) | 1 |
| (a) | Static embeddings | |
| (b) | Random vectors | |
| (c) | Binary vectors | |
| (d) | Contextual embeddings | |
| (1.8) | Hugging Face provides (CO4,K1) | 1 |

- (a) Hardware (b) Pretrained transformer models
(c) Databases (d) Compilers
- (1.9) Statistical machine translation relies on (CO5,K1) 1
(a) Handcrafted grammar rules (b) Large bilingual corpora
(c) Syntax trees only (d) Rule-based parsing
- (1.10) Abstractive summarization is (CO5,K1) 1
(a) Rule-based (b) Generative (c) Static (d) Manual
2. Attempt all parts:-
- (2.1) An NER system identifies 30 entities correctly, misses 10, and incorrectly labels 5 non-entities. Calculate the precision and recall. (CO1,K3) 2
- (2.2) Generate bigrams and trigrams for: 2
“Deep learning improves systems” (CO2,K2)
- (2.3) Explain the architecture of an Information Retrieval system.(CO3,K2) 2
- (2.4) A transformer uses 8 attention heads. If each head outputs a 64-dimensional vector, what is the total concatenated output dimension?(CO4,K2) 2
- (2.5) Compare Rule-based and Statistical methods used in Named Entity Recognition (NER).(CO5,K2) 2

SECTION-B

Answer any one of the following:-

- (3.1) An NLP pipeline processes a document with 400 tokens. After tokenization, stop word removal (removing 35% of tokens), and stemming (reducing unique tokens by 20%), calculate the number of tokens at each stage.(CO1,K3) 6
- (3.2) Given the sentence 'The bank can guarantee deposits will eventually cover future tuition costs', identify all possible parse trees and explain the type of ambiguity. Also compute the number of words and the proportion of nouns assuming 3 nouns. (CO1,K3) 6

Answer any one of the following:-

- (3.3) Bag-of-Words (BoW) Vector Creation 6
Given the following documents:
D1: “Machine learning improves AI systems”
D2: “AI systems use deep learning”
D3: “Machine learning and deep learning are popular”
Construct the Bag-of-Words vector representation for all documents.
(CO2,K3)
- (3.4) Given two word/document vectors: 6
 $A=(1,2,3), B=(2,3,4)$
Compute the cosine similarity between vectors A and B. Show all steps clearly.
(CO2,K3)

Answer any one of the following:-

- (3.5) Explain how email spam filtering works and provide an example of how unwanted emails are detected.(CO3,K2) 6
- (3.6) Consider the following movie reviews:(CO3,K3) 6
R1: “The storyline was amazing and engaging.”
R2: “The acting was terrible and boring.”
R3: “Excellent visuals but weak plot.”
R4: “The movie was not good.”

Tasks: a) Perform sentiment classification for each review. b) Identify challenges caused by negation and mixed sentiment. c) Explain why Bag-of-Words may fail to correctly classify review R4. d) Suggest one NLP technique that can improve sentiment detection accuracy.

Answer any one of the following:-

(3.7) Analyze the differences between RNN, LSTM, and Transformer models in terms of architecture, ability to capture long-term dependencies, and computational efficiency.(CO4,K2) 6

(3.8) Question a. Define the query Q key K and value V vectors used in the attention mechanism 6

Given, $Q = [1 \ 1]$

Given, $K = [[1,0],[0,1]]$

b. Apply scaled dot product attention to compute attention scores scaling factor = 1 (CO4,K3)

Answer any one of the following:-

(3.9) Explain how parallel corpora are used in Statistical Machine Translation systems (CO5,K2) 6

(3.10) A customer-support dialog system processes 5,000 user queries daily. The system performance statistics are:(CO5,K3) 6

Correct responses = 4,200

Incorrect responses = 500

Unanswered queries = 300

Tasks:

a) Calculate the response accuracy of the dialog system.

b) Compute the error rate.

c) Differentiate between rule-based and neural dialog systems.

d) Explain how context-awareness improves conversational AI systems.

SECTION-C 50

4. Answer any one of the following:-

(4.1) Consider the sentence:(CO1,K3) 10
“Sundar Pichai announced new AI features at Google headquarters in California on Monday.”

a) Identify all named entities and classify them into categories such as Person, Organization, Location, Date, and Technology Term.

b) An NER system correctly identifies 4 entities, incorrectly labels 1 non-entity, and misses 1 actual entity. Calculate:

1. Precision

2. Recall

3. F1-score

c) Explain the importance of contextual understanding in identifying the word “Google” correctly.

d) Discuss one major challenge faced by Named Entity Recognition systems in real-world applications.

(4.2) Design an NLP preprocessing pipeline for a corpus of 2000 documents. If tokenization takes 0.5 sec/doc, stop word removal 0.2 sec/doc, and stemming 0.3 sec/doc, calculate total processing time. Also, if stop words constitute 40% of average 200 words/doc, compute the number of words remaining after preprocessing. Discuss the trade-off between speed and accuracy in each step.(CO1,K3) 10

5. Answer any one of the following:-

- (5.1) Derive the mathematical formulation of CBOW and Skip-gram models in Word2Vec. Explain how training is performed using neural networks and gradient descent.(CO2,K2) 10
- (5.2) Consider the sentence: 10
Consider the sentence: "Machine learning is powerful and widely used in artificial intelligence applications." Using the Skip-gram model of Word2Vec with window size = 2:
a) Generate all possible target-context pairs.
b) Construct the one-hot vector representation for each unique word in the vocabulary.(CO2,K3)

6. Answer any one of the following:-

- (6.1) Compare LSA and LDA in terms of assumptions, mathematical foundation, scalability, and interpretability. Which model performs better for large-scale topic modelling and why?(CO3,K2) 10
- (6.2) Explain the difference between Information Extraction and Information Retrieval, and analyze their roles in modern search engines and NLP systems(CO3,K2) 10

7. Answer any one of the following:-

- (7.1) Distinguish between ELMo, BERT, and FastText in terms of context handling and representation learning.(CO4,K2) 10
- (7.2) Critically analyze the challenges faced by Transformer-based models in streaming and real-time conversational AI systems. Discuss the impact of long-sequence processing, energy consumption, scalability, and multilingual speech understanding on system performance.(CO4,K4) 10

8. Answer any one of the following:-

- (8.1) Analyze the role of text summarization and dialog systems in modern conversational agents and chatbots. Discuss major challenges in multilingual communication and context handling.(CO5,K4) 10
- (8.2) (a) Analyze how different features such as term frequency and sentence position influence the performance of extractive summarization systems. Discuss their advantages and limitations. (b) Consider the following sentences with given feature scores: | Sentence | Term Frequency (TF) | Position Score | 10
| ----- | ----- | ----- |
| S1 | 6 | 1.0 |
| S2 | 4 | 0.7 |
| S3 | 5 | 0.5 | Using the scoring formula: $\text{Score} = (0.6 \times \text{TF}) + (0.4 \times \text{Position})$ (i)
Calculate the final score for each sentence.
(ii) Rank the sentences based on their scores.
(iii) Select the top 2 sentences for the summary and justify your selection.(CO5.K3)