Department of Computer Science and Engineering

Central Institute of Technology Kokrajhar

End Semester Examination

B. Tech

Course Title: **Speech and Natural Language Processing** Course Code: **UCSE613**
Session: **Jan-Jun, 2025** Full Marks: **100** Time: **3:00** hrs

---

Figure in the margin indicates full marks.
Question1 is compulsory, Answer any three from the rest!

1 [A] Fill in the blanks 2 x 10 = 20

i. The task of sentiment analysis involves classifying text into categories like positive, negative, or neutral. When labeled training data is sparse, pre-built word lists called _____ annotated for intensity can be used.
ii. In evaluating binary text classification systems, a _____ matrix is used to visualize performance by showing the counts of true positives, false positives, true negatives, and false negatives.
iii. _____ is used for automatic evaluations of language modelling.
iv. _____ is the process of breaking down text into individual words or phrases
v. Using _____ analysis helps understand the sentiment expressed in text.
vi. Unlike generative classifiers such as Naive Bayes that model the likelihood P(d|c) and prior P(c), _____ classifiers like Logistic Regression attempt to directly compute the posterior probability P(c|d).
vii. _____ is the process of organizing unstructured text into predefined categories
viii. A key task in Information Extraction is _____, which involves labeling certain kinds of proper nouns like personal names, organizations, and locations.
ix. Language models commonly perform calculations in _____ space to avoid underflow when computing the joint probability of a sentence
x. The level of linguistic representation that deals with structures and effects in related sequences of sentences, such as texts or dialogues, is called _____.

[B] Multiple Choice Questions 2 x 10 = 20

i. Which of the following is a core component of Natural Language Processing (NLP)?
   a) Image Recognition
   b) Language Analysis
   c) Mathematical Proof Generation
   d) Robot Kinematics
ii. Which of the following is listed as a potential career field for someone working in NLP?
   a) Archaeology
   b) Astronomy
   c) Humanitarian organizations

d) Culinary Arts

iii. Analyzing a word into its meaningful components, such as breaking "cats" into "cat" and "s", is an example of which level of linguistic analysis?

a) Syntax
b) Semantics
c) Morphology
d) Phonetics

iv. What is the smallest meaning-bearing unit of a language?

a) Phoneme
b) Syllable
c) Word
d) Morpheme

v. Which of the following is a key challenge for computers in processing natural language?

a) Lack of processing power
b) Inability to store large amounts of text
c) Ambiguity
d) Limited vocabulary

vi. What is the main goal of a probabilistic language model?

a) To generate grammatical sentences
b) To parse the syntactic structure of a sentence
c) To assign a probability to a sentence
d) To identify the topic of a document

vii. In evaluating a binary text classification system, what does a "True Positive" represent?

a) An item correctly labeled as belonging to the negative class
b) An item incorrectly labeled as belonging to the negative class
c) An item incorrectly labeled as belonging to the positive class
d) An item correctly labeled as belonging to the positive class

viii. What is the term for the percentage of items actually present in the input that were correctly identified by the system?

a) Precision
b) Accuracy
c) Recall
d) F-measure

ix. What problem does "smoothing" primarily address in language modeling?

a) Reducing the size of the vocabulary
b) Speeding up probability calculation
c) Dealing with zero probability N-grams
d) Handling misspellings in the text

x. What is a key difference between a Generative Classifier (like Naive Bayes) and a Discriminative Classifier (like Logistic Regression)?

a) Generative classifiers are typically faster to train.
b) Discriminative classifiers model the distribution of the features given the class, $P(d|c)$.
c) Generative classifiers are only used for binary classification tasks.
d) Discriminative classifiers attempt to directly compute $P(c|d)$.

2 a. Explain the concept of *ambiguity* in natural language and why it poses a significant challenge for NLP systems. Provide an example. 5 + 10 + 5 = 20

b. Describe the process of *tokenization* in basic text processing. Explain the difference between a *token* and a *type* and how these relate to

vocabulary size in a corpus.
c. Compare and contrast *Lemmatization* and *Stemming* as methods for word normalization, explaining their goals and how they differ.

3 a. Explain the purpose of a *Probabilistic Language Model*. How does a language model assign a probability to a sentence, particularly using the Chain Rule? 10 + 10 = 20
b. Describe what *N-gram language models* are and how they estimate the probability of a word sequence. What is a significant practical issue that arises when estimating N-gram probabilities using simple Maximum Likelihood Estimates (MLE)?

4 a. Explain the concept of perplexity as an intrinsic evaluation metric for language models. What does a lower perplexity score indicate about a language model? 5 + 5 + 10 = 20
b. Describe the basic idea behind the *gradient descent* algorithm used for training Logistic Regression models. What does the *learning rate (η)* control in this process?
c. Explain the difference between *Stochastic Gradient Descent, Batch Gradient Descent*, and *Mini-batch Gradient Descent*.

5 a. What are Activation Functions in Neural Networks? Explain any three activation functions. 10 + 10 = 20
b. Explain with a toy example of classifying whether a given image is a Car or Bike using Artificial Neural Network.

6 Write short notes any four 5 x 4 = 20
a. Word Sense Disambiguation
b. Generative vs. Discriminative Classifiers
c. Back propagation in ANN
d. Sentiment Analysis
e. Manner of Articulation in Speech
f. Levenshtein Edit Distance Algorithm

xxXxx