E/L/TX

Reg. No:

VIT
Vellore Institute of Technology

Final Assessment Test – November 2024

Course: **BCSE409L - Natural Language Processing**

Class NBR(s): **1855/1858/1864/1879/1884/1890** Slot: **E1+TE1**

Time: **Three Hours** Max. Marks: **100**

- **KEEPING MOBILE PHONE/ANY ELECTRONIC GADGETS, EVEN IN 'OFF' POSITION IS TREATED AS EXAM MALPRACTICE**
- **DON'T WRITE ANYTHING ON THE QUESTION PAPER**

Answer ALL Questions

(10 X 10 = 100 Marks)

1. Outline the major stages of the NLP pipeline and choose a specific NLP application (sentiment analysis, or machine translation, or chatbot development) and illustrate how each stage of NLP contributes to the overall process in that application.

2. a) Can you apply the rules of derivational morphology to form new words from the base word 'happy'? Consider how prefixes and suffixes alter the meaning and grammatical category of the base word. What new words can be formed, and how does each derivational morpheme affect the word's function, part of speech, and meaning? Can you also explore cases where multiple affixes are applied simultaneously (e.g., unhappiness), and discuss how the resulting words relate to their base form? **[4]**

   b) Design a Finite state Transducer to implement the following k – insertion rule. **[6]**

   Rule: If a verb ends with vowel + -c add –k while changing it to past tense.

   Example:

   - panic + ed = panicked
   - picnic + ed = picnicked

3. Consider a sequence of words ["rabbit", "jumps", "quickly"] to find the most likely sequence of tags (Noun, Verb, Adverb) using the Viterbi algorithm. Then, compare the sequence (Noun → Noun → Verb) with a lower probability path.

Transition, emission, and start probabilities are given:

Start Probabilities:

P(Noun|Start)=0.5

P(Verb|Start)=0.3

P(Adverb|Start)=0.2

Transition Probabilities:

P(Verb|Noun)=0.7

P(Adverb|Verb)=0.8

Emission Probabilities:

P(rabbit|Noun)=0.6

P(jumps|Verb)=0.7

P(quickly|Adverb)=0.9

4. Consider the following CFG that describes a simple language:
Grammar:
S → NP VP
NP → Det N | Det N PP
VP → V NP | V NP PP
PP → P NP
Det → 'the' | 'a'
N → 'man' | 'dog' | 'telescope'
V → 'saw' | 'ate'
P → 'with' | 'in'

Derive the parse tree for the sentence "the man saw the dog with the telescope" that can be derived from the given context free grammar (CFG). Provide the different parse trees for the ambiguous sentence and explain the different interpretations. Brief the strategies that can be employed to resolve such ambiguities in NLP applications.

5. Explain the concept of dependency parsing and its role in syntactic analysis of text. Discuss the different types of dependency parsing algorithms in brief with appropriate example.

6. Compare and contrast the supervised and unsupervised approaches to Word Sense Disambiguation (WSD), highlighting their key advantages, disadvantages, and common application areas. Which approach would be more suitable for building a WSD system in a specialized domain, such as medicine or law, and why? Give specific examples to represent the different WSD tasks for which each approach is ideal.

7. a) Discuss concepts of term frequency (TF) and inverse document frequency (IDF) and elaborate the significance of TF-IDF in identifying important words in a document relative to a collection of documents, and how it helps in feature extraction for NLP tasks. [5]

b) Given the following set of documents: [5]
Document 1: "Cats are great pets."
Document 2: "Dogs are loyal animals."
Document 3: "I have both cats and dogs as pets."
Calculate the term frequency for "cats" in each document, inverse document frequency for "cats" across all documents, and TF-IDF for "cats" in each document.

8. Consider the following statement corpora:
<s> Business pays the best in investment </s>
<s> Investment is a good plan</s>
<s> Investment increases at fast rate </s>
<s> Education is the good man in life </s>
<s> Education is a real asset </s>
<s> Karna best in sports</s>
<s> The smile of children is the best in life</s>

<s> Sports are the most important things in life </s>
<s> Karna is a real good investor </s>
<s> Karna likes to walk in rain </s>
<s> Karna is a real good man in the world</s>

Using a bi-gram model calculate the Perplexity of the following sentence:
<s> KARNA IS A REAL GOOD MAN IN LIFE </s>

9.a) Design and discuss in detail about a hybrid text summarization system that combines both extractive and abstractive methods. What architectural components would you include, and how would you ensure the system optimally balances coherence and informative.

OR

9.b) GPT models may sometimes generate biased or inappropriate answers. How would you detect and mitigate bias in answers generated by a question-answering system? Provide strategies for bias detection, filtering, and rephrasing in sensitive domains such as hiring, law, or education with proper explanation.

10.a) Design an Neural Machine Translation (NMT) system that adapts to user preferences over time, such as translating formal or informal language styles or adhering to specific terminology preferences for businesses? Explore techniques for personalization in translation models, challenges related to maintaining generality, and balancing user-specific adaptations with overall system performance.

OR

10.b) Critically assess the ethical implications of using sentiment analysis to gauge public opinion during an election cycle? Discuss in detail about the potential biases in the data collected from social media and how they might affect the analysis with neat diagram and methodology.

⇔⇔⇔ **E/L/TX** ⇔⇔⇔