Reg. No.:
Name :

VIT
Vellore Institute of Technology

## Continuous Assessment Test I – September 2023

| | | | | | |
|---|---|---|---|---|---|
| Programme | : | **B.Tech CSE** | Semester | : | **Fall 2023-24** |
| Course | : | **Natural Language Processing** | Code | : | **CSE4022** |
| Faculty | : | **Dr. D. Manjula**<br>**Dr. M. Premalatha** | Slot(s) | : | **F1+TF1** |
| | | | Class Nbr(s) | : | **CH2023240100231**<br>**CH2023240100233** |
| Time | : | **1½ Hours** | Max. Marks | : | **50** |

Answer **ALL** the Questions

Consider the following scenario:
"Imagine you are a real estate agent showing a client's *father* a luxuriously designed mansion with stunning architectural **properties** and breath-taking views. As you gracefully walk them through the spacious rooms, the *cleverly* designed kitchen, and the *clear*, panoramic windows, they enthusiastically express their desire to achieve similar levels of **prosperity** in their life. However, as you continue the tour, they humorously remark about how one room appears a bit *fattened* in terms of space utilization, adding a light-hearted touch to the conversation. Finally, they curiously inquire if the previous owner had been a *famously* well-hearing individual, igniting a sense of intrigue in the conversation. How would you analyze the overall sentiment of this interaction?"
Answer the following questions Q1 to Q5 based on the given scenario

1. For the given input text, which type of question answering has to be applied? Justify your answer also by briefly explaining the types of question answering. **[10]**

2. Apply Porter's stemmer algorithm for the words in the first two sentences of the given text. You have to use all the relevant rules and provide an explanation for each of them. **[10]**

3. Choose the 5 words that are underlined font and construct the Finite State Transducer (FST) diagram for the same. **[5]**

4. Choose any 10 words [5 inflectional and 5 derivational morphemes] from the text and deduce the existing morphemes and name their type. **[10]**

5. Apply regular expressions for selection and non-selection for the scenario and answer the queries specified in the following [5*1=5 Marks] **[5]**
   i) The pattern which accepts both forms of two word phrase representations [eg: light-hearted and light hearted]
   ii) The pattern matches ?
   iii) The pattern which doesn't match alphanumeric words.
   iv) The pattern which accepts both analyse and analyse
   v) The pattern which accepts the characters B, c, d, e, f, g

6. Consider bold font words from the text and convert one word into the other word by considering insertion, and deletion costs as 1, and substitution cost as 2 using the dynamic programming approach and perform the following: **[10]**
   i). Write the pseudocode for the algorithm [3 Marks]
   ii). Construct the distance matrix for the words with the directions [4 Marks]
   iii). Specify the path using backtracking and map the characters and its operations involved [3 Marks]

⇔⇔⇔

Reg. No.:
Name :

VIT
Vellore Institute of Technology

**Continuous Assessment Test II – October 2023**

| | | | |
|---|---|---|---|
| Programme | : **B.Tech CSE** | Semester | : **Fall 2023-2024** |
| Course | : **Natural Language Processing** | Code | : **CSE4022** |
| Faculty | : **Dr. D. Manjula**<br>**Dr. M. Premalatha** | Slot(s) | : **F1+TF1** |
| | | Class Nbr(s) | : **CH202324010023D**<br>**CH2023240100233** |
| Time | : **1½ Hours** | Max. Marks | : **50** |

Answer **ALL** the Questions

1. Consider the following scenario:
"As the sun set over the horizon, the golden waves gently kissed the sandy shores, creating a breath-taking spectacle. Suddenly, a group of giggling children appeared, chasing a colourful butterfly through the lush garden. Their laughter filled the air with infectious joy, and the butterflies fluttered away in search of a quieter spot. Meanwhile, a wise old owl perched on a sturdy branch, observing the lively scene below with keen interest. Nightfall brought a sense of tranquillity to the garden, and the moon rose, casting its silvery glow on the slumbering flowers."

a) The scenario has five sentences. Assign a Penn Treebank tagset by applying appropriate parts-of-speech tags for each of the sentences specified in the scenario. [5*2=10] [10]

b) Using the Probabilistic-Context-Free-Grammar, derive the probability of the first sentence of the given scenario by constructing a minimum of two parse trees. [10]
Note: Specify the rules and the lexicons with the assumed probabilities.

2. Consider the sentence "As the sun set over the horizon" and the transition probabilities.

| | | |
|---|---|---|
| P(N\|N) = 0.1 | P(V\|PREP) = 0.3 | P(ADV\|N) = 0.2 |
| P(N\|set) = 0.2 | P(ADJ\|PREP) = 0.2 | P(DET\|V) = 0.4 |
| P(set\|N) = 0.4 | P(DET\|N) = 0.8 | P(DET\|PREP) = 0.3 |
| P(N\|PREP) = 0.4 | P(ADV\|ADJ) = 0.4 | P(Set\|N) = 0.5 |
| P(PREP\|N) = 0.3 | P(DET\|ADJ) = 0.1 | P(Set\|V) = 0.3 |
| P(N\|DET) = 0.7 | P(ADJ\|V) = 0.3 | P(ADJ\|N) = 0.6 |
| P(V\|N) = 0.5 | P(N\|ADJ) = 0.3 | P(N\|V) = 0.4 |
| P(PREP\|V) = 0.2 | | |

a) Disambiguate the word "set" using the transition probabilities. [5]

b) What is the probability of the word "set" being identified as a verb or noun, given "sun" being identified as a noun and "over" being identified as the preposition? Use the chain rule to clarify the same. [5]
Note: Assume the transition probabilities if any of the probabilities are not available.

3. Use the word "close" and frame 5 different sentences where in each sentence, the parts-of-speech tag of the word "close" is unique. [5]

4. Consider the following corpus [15]
<s>I went for a run yesterday<s>
<s>I run every morning</s>
<s>Drink a tea</s>
By considering the sentences specified above, find the probability for the sentence "I drink tea every morning" using Bigram estimates and apply appropriate smoothing.

⇔⇔⇔

VIT Vellore Institute of Technology

Reg. No. : 20BCE1707

**Final Assessment Test (FAT) - November/December 2023**

| Programme | **B.Tech.** | Semester | **FALL SEMESTER 2023 - 24** |
|---|---|---|---|
| Course Title | **NATURAL LANGUAGE PROCESSING** | Course Code | **CSE4022** |
| Faculty Name | **Prof. Premalatha M** | Slot | **F1+TF1** |
| | | Class Nbr | **CH2023240100233** |
| Time | **3 Hours** | Max. Marks | **100** |

**SECTION A (10 X 10 Marks)**
**Answer all questions**

01. Identify and explain the thematic roles of highlighted phrases in the following sentences: **[10]**
(5*2=10 Marks)
a) John broke the window with a rock
b) Doris gave the book to Edward
c) A fork ate the noodles
d) Usually John agrees with Mary on everything
e) Sales fell to $25 million from $27 million

02. Apply porter stemmer on the words of following sentence and specify all possible rules that are applied: ***[10]***
*According to the Cambridge Dictionary, a sentence is defined as "a group of words, usually containing a verb that expresses a thought in the form of a statement, question, instruction, or exclamation"*
Note: If a word itself a stem and no rules are required to be applied, specify the same

03. What is the tagging of the sentence "computers process programs accurately" with the following HMM tagger: **[10]**
(part of) lexicon:
computers N 0.123
process N 0.1
process V 0.2
programs N 0.11
programs V 0.15
accurately Adv 0.789
(part of) transitions:
P(N|V)=0.5 P(N|Adv)=0.12 P(V|Adv)=0.05
P(V|N)=0.4 P(Adv|N)=0.01 P(Adv|V)=0.13
P(N|N)=0.6 P(V|V)=0.05

04. Identify and explain the type of lexical semantic relations [hyponymy/ hypernymy/ synonymy/ antonymy/ polysemy/ homonymy/ meronymy]. of the underlined word(s) in each of the following corpus to analyze the meanings of the words based on their relationships with one another: [5*2=10 Marks] **[10]**
Corpus 1:

Which flights serve breakfast?

Does America West serve Philadelphia?

Corpus 2:

I saw a fly on the wall

Corpus 3:

The wind can be strong enough to wind a kite string.

Corpus 4:

An eagle is a type of bird.

Corpus 5:

The wheels of the car were badly worn.

05. A company active in automatic recognition of hand-written documents needs to improve the quality of their recognizer. This recognizer produces sets of sequences of correct English words, but some of the produced sequences do not make any sense. For instance the processing of a given hand-written input can produce a set of transcriptions like: "A was salmon outer the does", "It was a afternoon nice sunny", and "I Thomas at mice not the spoon". Identify the errors in these types of sentences and suggest the levels of NLP that can be applied to interpret the same. Justify your answer. **[10]**

06. Create a transducer(s) for the consonant doubling spelling rule in English. For example, the surface form of "beg" + "ing" is "begging". Note that there are exceptions; for example, the surface form of "sing" + "ing" is "singing", not "singging". All strings should be accepted. If the end of the string consists of a morpheme ending in a consonant followed by 'ing', then the output of the transducer should be the input with the appropriate consonant doubled. Show the sequence of states your machine goes through to accept the string and explain them in detail. **[10]**

07. Apply Penn Treebank POS tagging for the following: [5*2=10 Marks] **[10]**

a) John broke the window with a rock

b) Doris gave the book to Edward

c) A fork ate the noodles

d) Usually John agrees with Mary on everything

e) Sales fell to $25 million from 27$ million

08. Consider the grammar, lexicons, and their respective probabilities as specified below. **[10]**

| Rules | Probability | Rules | Probability |
|---|---|---|---|
| S → NP VP | 1.0 | NP → Astronomers | 0.40 |
| PP → P NP | 1.0 | NP → ears | 0.18 |
| VP → V NP | 0.7 | NP → saw | 0.04 |
| VP → VP PP | 0.30 | NP → telescopes | 0.1 |
| NP → NP PP | 0.4 | P → with | 1.0 |
| NP → stars | 0.18 | V → saw | 1.0 |

Generate a sentence with 5/6 lexicons and find the probability using the probabilistic context-free grammar method.

Note: Use two parse trees for calculating the probability of a sentence

09. Identify the phrases in the following sentences according to CONLL 2000 chunking task definition. (4*2.5=10 Marks) **[10]**

a) Tokens are usually words but I noticed that some words get split into multiple tokens.

b) A tagger is a piece of software that reads text in some language and assigns parts of speech to each word such as nouns, verbs, adjectives, etc.

c) The Brill tagger and Treebank parser both need to have sentences broken down into tokens separated by spaces.

d) The tagger uses a dictionary of tags and a trained model to apply parts of speech tags to each token in a sentence.

10. Apply Edit Distance to convert the word "cereal" to the word "serial". Illustrate with the distance matrix and specify the backtracking path. Use the insertion cost as 1, the deletion cost as 1, and the substitution cost as 2. Explain the same with a pseudocode. **[10]**