Date: 09-09-2024   **Quiz -1**   **Time allowed: 25 mins**

NLP-A   **Total Marks: 18**

Q1. Identify the different types of morphemes (bound & free, or reduplication etc.) in the following words:

(6)

1) sunflower

2) Discussion

3) windmill

4) dilly-dally

5) Beautify

6) womanly

Q2. Using the standard MaxMatch algorithm segment the following unsegmented sentence. Does it have any unknown words?   (5)
Unsegmented sentence: thesistermisjudgedatortoiseagain
Dictionary:
the, gain, misjudged, or, at, sister, an, thesis, to, a, judged, term, is
Segmented Sentence:

Q3. How do BPE and maxmatch differ? and which algorithm is a better tokenizer than the other? Explain with an example.   (2)

Q4. Assume you have the following training corpus, include </s> in your corpus like any other token.   (5)

<s> I love eating pizza and ice cream </s>
<s> My favorite food is sushi and ramen </s>
<s> Eating out is fun and exciting </s>
<s> Pizza is my go to food </s>
Find the *Unigram and Bigram* probability of the given test sentence.
Calculate the perplexity of these both as well.

Test Sentence:
<s> I love eating sushi </s>