Name: _____ Roll: _____

**CSE 556: Natural Language Processing (NLP) — QUIZ - 1**

**Date:** 7th Feb 2025 **Duration:** 30 mins. **Max Marks: 20**

---

**1.** Given a tag set T = [**NN** (noun), **VB** (verb), **JJ** (adjective), **RB** (adverb), **IN** (preposition), **PRP** (pronoun), **O** (Others)], find the correct PoS tag sequence for the following sentences. **[3*2 marks]**

| An | Argentine | international | , | Messi | is | his | country's | all-time | leading | goal | scorer | . |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| O | JJ | NN | O | NN | VB | PRP | NN | JJ | JJ | NN | NN | O |

| The | complex | houses | married | and | single | soldiers | and | their | families | . |
|---|---|---|---|---|---|---|---|---|---|---|
| O | NN | VB | JJ | O | JJ | NN | O | PRP | NN | O |

| The | old | man | the | boat | . |
|---|---|---|---|---|---|
| O | NN | VB | O | NN | O |

***No partial marking for any mistake.**

**2.** Differentiate between type and token. Give the count of each for the following sentence. **[2]**

| Unarguably, Federrer, Nadal, and Djokovich are the best tennis players ever and they are leading the grand slam trophies in Wimbledon (8), French Open (14), and Australian Open (9), respectively. |
|---|
| Tokens: A character sequence which has a specific meaning. Includes valid words and punctuations.<br>Types: Unique tokens. |
| Tokens: 43<br>Types: 29 |

**3.** Define $N_C$ in Good-Turing Smoothing. **[1]**

| Number of words with frequency C. |
|---|

**4.** For a vocabulary size $|V|$, context size $C$, and embedding dimension $H$, how many parameters (without biases) does a Skipgram-Word2Vec model will have at the? **[2]**

| | |
|---|---|
| **a.** Input-to-hidden layer | $\|V\| \times \|H\|$ |
| **b.** Hidden-to-output layer | $2C \times \|H\| \times \|V\|$ |

**5.** For the given merge rules [r$, er$, ew, new, lo, low, newer$, low$], tokenize the word "**newest**" using BPE. Show steps. **[2]**

n-e-w-e-s-t-$ ⇒ n-ew-e-s-t-$ ⇒ new-e-s-t-$

No marks without steps.

**6.** Write the equations for an LSTM unit at timestep *t*. Define all necessary variables appropriately. **[7]**

Check slides

2 marks for each gate. An incorrect equation will fetch zero for that gate.
1 mark for defining all variables.