==========

Fourth Year Natural Language Processing Question Paper

==========

SECTION 1: MULTIPLE CHOICE QUESTIONS (10 Questions, 1 Marks each)

Bloom's Taxonomy Level: 3

Topic: Basic Text Processing and Morphology

Subtopic: Tokenization (word token, word type)

Q1: Which of the following sentences has a higher word type to word token ratio?

A) The cat sat on the mat. The cat sat on the mat.

B) The quick brown fox jumps over the lazy dog. (1 Marks)

Q2: what isd my name (1 Marks)

Q3: What is the difference between a word token and a word type in the context of tokenization?

a) There is no difference; they are synonyms.

b) A word token is an instance of a word in a text, while a word type is a unique word in the vocabulary.

c) A word token is a unique word in the vocabulary, while a word type is an instance of a word in a text.

d) Word tokens are always capitalized, while word types are not. (1 Marks)

Q4: Which of the following sentences would result in the fewest unique word *types* after tokenization?

a) "The cat sat on the mat."

b) "The dog chased the ball."

c) "The quick brown fox jumps over the lazy dog."

d) "The cat sat on the mat, the mat was red." (1 Marks)

Q5: Given the sentence "She sells seashells by the seashore.", how many word *types* are present, considering case-insensitive tokenization?

a) 6

b) 7

c) 8

d) 9 (1 Marks)

Subtopic: Word segmentation

Q6: Which of the following languages poses the GREATEST challenge for unsupervised word segmentation algorithms due to the lack of explicit word boundaries?

a) English
b) Spanish
c) Chinese
d) German (1 Marks)

Q7: A text contains the phrase "applesauceisdelicious". Which word segmentation approach would MOST likely correctly segment this phrase without relying on a dictionary?

a) Maximum likelihood estimation
b) Rule-based segmentation using hyphenation rules
c) A probabilistic approach based on character n-gram frequencies
d) A dictionary-based lookup (1 Marks)

Q8: You are developing a word segmentation algorithm for a low-resource language. Which of the following is the MOST important factor to consider when choosing a suitable approach?

a) Computational speed
b) Availability of a large, annotated corpus
c) The complexity of the language's morphology
d) The popularity of the language (1 Marks)

Q9: Which of the following is NOT a common evaluation metric for word segmentation?

a) Precision
b) Recall
c) F1-score
d) Root Mean Squared Error (RMSE) (1 Marks)

Q10: A word segmentation algorithm consistently mis-segments compounds in a German text. Which improvement would be MOST likely to address this issue?

a) Increasing the size of the training corpus.
b) Incorporating morphological information, such as prefixes and suffixes, into the algorithm.
c) Switching to a purely rule-based approach.
d) Reducing the complexity of the algorithm. (1 Marks)