

**START OF QUIZ**

**Student ID:**

**76415876,Zhou,Wei**

Academic honesty is essential to the continued functioning of the University of British Columbia as an institution of higher learning and research. All UBC students are expected to behave as honest and responsible members of an academic community. Failure to follow the appropriate policies, principles, rules, and guidelines of the University with respect to academic honesty may result in disciplinary action.

I agree that all answers provided are in my own words, and that I will not discuss the contents of this quiz with any of my fellow students until after the exam period has completed for everyone. Furthermore, any response that used generative AI tools has been rephrased into my own interpretation, and has been appropriately cited.

Signature: \_\_\_\_\_

## Question 1

Topic: Lecture 5

Source: Lecture 5

Describe the purpose of the various types of brackets in regexes, and how they differ. (1)

## Question 2

Topic: Lecture 5

Source: Lecture 5

There are two ways of matching a pattern against the start of a string. Describe them. (1)

### Question 3

Topic: Lecture 7

Source: Lecture 7

Can you think of any classes of words in English where the stem and the lemma will always be identical? Why is that of little interest to us? (1)

## Question 4

Topic: Lecture 6

Source: Lecture 6

Why is XML well-suited to representing linguistic data? (1)

## Question 5

Topic: Lecture 8

Source: Lecture 8

Why do Python programmers like working with (t/c)sv files? When are they appropriate, and what advantages do they provide over .txt files? (1)

## Question 6

Topic: Lecture 6

Source: Lecture 6

Suppose you've trained a Named Entity Recognition (NER) model using XML-annotated text data, but it consistently fails to recognize locations. What steps would you take to determine if the problem lies with the model, the training data, or both? What resources would you need to investigate further? (2)

## Question 7

Topic: Lecture 8

Source: Lecture 8

Imagine that you're working with a linguist who is not very good with technology. They store all of their data in .docx files, scattered across their desktop. What arguments would you make for them to convert to .tsv or .json, and how would you alleviate their worries that they wouldn't be able to access or modify their information (no, you can't teach them Python)? (2)

## Question 8

Topic: Lecture 7

Source: Lecture 7

I mentioned in class that POS tagging is often viewed as a pre-processing step for many CL tasks. What assumptions are we making (at least 3) when including it in our NLP pipeline? Do you think these are reasonable assumptions, and if they fail, is it worth the effort to solve the problem, or just ignore POS tagging? (2)

## Question 9

Topic: Long

Source: Lecture 7

Suppose you're building a text classification model for a highly inflected language like Finnish. How might you approach preprocessing tasks such as lemmatization or stemming? Would you perform these tasks before or after feature extraction, and why? Discuss how the choice of sequence may impact the quality of the features and model accuracy. Would you make the same decision for sentiment analysis? (3)

# END OF QUIZ