

**START OF QUIZ**

**Student ID:**

**16551426,Cao,Tianhao**

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 7

Source: Lecture 7

What might the training data for a sentence segmenter look like? Do you think it would be easy or hard to train? Explain briefly. (1)

## Question 2

Topic: Lecture 6

Source: Lecture 6

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

### Question 3

Topic: Lecture 8

Source: Lecture 8

If you were to encounter an alien text, which encoding might you want to use to digitize it?  
Explain briefly. (1)

## Question 4

Topic: Lecture 7

Source: Lecture 7

In class, we built a POS tagger that tries to give a majority tag to a word; if it's out-of-vocabulary, it backs-off to Regexes. This is clearly overly simplistic. List two assumptions that are being violated by this model. (1)

## Question 5

Topic: Lecture 5

Source: Lecture 5

There are two ways of matching a pattern against the start of a string. Describe them. (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 5

Source: Lecture 5

Imagine you are processing a text document where dates are written in multiple formats, such as "12-05-2024", "05/12/2024", or "12 December 2024". How would you write a regex to capture these date formats (just the logic)? What assumptions would you make? (2)

## Question 9

Topic: Long

Source: Lecture 6

You've been hired by a company that is working with their own version of XML that they call "NQAXML" (Not-Quite-As-eXtensible Markup Language). It provides stronger restrictions on tag names (they must be all uppercase, and no longer than 10 characters long), and it doesn't allow nested spans with identically-named tags. Like HTML, it also has a set of tags that must appear in every document. Describe your process for creating a data validator that takes an XML file, and ensures that it satisfies the rules of NQAXML. (3)

# END OF QUIZ