

**START OF QUIZ**

**Student ID:**

**57830358,Yi,Vincent**

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

Imagine you have a block of text with paragraphs separated by blank lines. How would you use regex to find the start of each paragraph? What assumptions would you make about the formatting of the text? (1)

### Question 3

Topic: Lecture 8

Source: Lecture 8

What are two advantages of using .py files over .ipynb files for deployment, and two reasons why .ipynb files are preferred for prototyping or development? (1)

## Question 4

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 5

Topic: Lecture 6

Source: Lecture 6

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

## Question 6

Topic: Lecture 7

Source: Lecture 7

Do you think that we could do lemmatization before machine translation? Provide 1 argument that for why it might help, and one for why it might make things more complicated. List any assumptions that might make your answer more complicated. (2)

## Question 7

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 8

Topic: Lecture 6

Source: Lecture 6

Consider using XML to represent a machine learning model's architecture. What XML tags might be useful for representing layers, activation functions, and connections between layers (you don't need to describe a deep-learning architecture - describe one you're familiar with)? If this doesn't seem possible, explain why not. (2)

## Question 9

Topic: Long

Source: Lecture 8

Imagine that you find an important file buried on a hard drive found in the basement of a university. You are trying to access the data, but realize it is corrupted. Some of the bits have been flipped (switched from 0 to 1, or 1 to 0), and others have been completely deleted. You don't know the encoding, and you don't know the language the data is written in. What are some tests you could run to try to establish and restore at least some of the data? (Hint: remember that a "byte" is 8-bits, and that UTF-8 is 1 byte, or 8 bits, UTF-16 is 2 bytes, or 16 bits, and UTF-32 is 4 bytes, or 32 bits). (3)

# END OF QUIZ