

**START OF QUIZ**  
**Student ID:**  
**46280228,Guo,Gilly**

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 8

Source: Lecture 8

Why should you get into the habit of using "with open()"? Are there any downsides? (1)

## Question 2

Topic: Lecture 6

Source: Lecture 6

Beautiful Soup parses the children of a tag as a list. Why do you think they didn't use a set, instead, given the faster access times? Give 2 reasons, and briefly explain. (1)

## Question 3

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 4

Topic: Lecture 6

Source: Lecture 6

How would we find all images in an HTML document? (1)

## Question 5

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 6

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 7

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 8

Topic: Lecture 5

Source: Lecture 5

Imagine we have a spell-checker that can identify common misspellings of words by replacing certain letters with a capture group that contains letters that are nearby on the keyboard. How aggressive of a regex would we want to write for this (ie, how many letters in the word would we want to replace with a group)? Explain. (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**