START OF QUIZ
Student ID: 98055874,Tseng,Agnes

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

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

Question 2

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 3

Topic: Lecture 6
Source: Lecture 6

How would we find all images in an HTML document? (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 5
Source: Lecture 5

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

Question 6

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 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 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 satisties the rules of NQAXML. (3)

END OF QUIZ