START OF QUIZ
Student ID:
26835850,Fu,Infinity

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 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 3

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 4

Topic: Lecture 5
Source: Lecture 5

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

Question 5

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 6

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 7

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 8

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 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