START OF QUIZ
Student ID:
80059942,Leier,Kara

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 impact does lemmatization or stemming have with respect to the Zipfian curve? How might that affect our algorithms? (1)

Question 2

Topic: Lecture 6
Source: Lecture 6

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

Question 3

Topic: Lecture 6
Source: Lecture 6

What kinds of tags might be useful in the following text (describe at least two): ”But you liked Rashomon!” ”That’s not how I remember it!” (1)

Question 4

Topic: Lecture 5
Source: Lecture 5

List one advantage that regular expressions have over string comparison, and one disadvantage to using them. (1)

Question 5

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 6

Topic: Lecture 8
Source: Lecture 8

In class, I mentioned that we always want to close a file correctly. Beyond freeing up system resources, it also "flushes the buffer", which ensures that any current read or write operations that are in the job queue, but haven't yet been processed, are completed. Knowing what you do about encodings, what is a possible ramification of not flushing the buffer? Explain at least 2. (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 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 9

Topic: Long
Source: Lecture 7

Suppose you're building a text classification model for a highly inflected language like Finnish. How might you approach preprocessing tasks such as lemmatization or stemming? Would you perform these tasks before or after feature extraction, and why? Discuss how the choice of sequence may impact the quality of the features and model accuracy. Would you make the same decision for sentiment analysis? (3)

END OF QUIZ