

START OF QUIZ

Student ID:

**70152863, Khan, Muhammad
Mujtaba**

Question 1

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 2

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 3

Topic: Lecture 7

Source: Lecture 7

What is the difference between a stem and a lemma? What impacts does that have on our algorithms? (1)

Question 4

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 5

Topic: Lecture 6

Source: Lecture 6

XML can be opened by most plain-text text editors. Name a benefit and a disadvantage of this feature. (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 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 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 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