

START OF QUIZ

Student ID:

49919301,Keigan,Jonathan

Question 1

Topic: Lecture 6

Source: Lecture 6

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

Question 2

Topic: Lecture 5

Source: Lecture 5

Write a regex pattern that matches any valid email address (i.e., with basic rules like user@domain.com). What challenges might you face in accurately matching all possible email formats? (1)

Question 3

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 4

Topic: Lecture 7

Source: Lecture 7

Can you think of any classes of words in English where the stem and the lemma will always be identical? Why is that of little interest to us? (1)

Question 5

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 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 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 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