

START OF QUIZ

Student ID:

19756840,Wang,Rennan

Question 1

Topic: Lecture 6

Source: Lecture 6

How would we find all images in an HTML document? (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 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 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 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 7

Topic: Lecture 5

Source: Lecture 5

Imagine we have a spell-checker that can identify common misspellings of words by replacing certain letters with a capture group that contains letters that are nearby on the keyboard. How aggressive of a regex would we want to write for this (ie, how many letters in the word would we want to replace with a group)? Explain. (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 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 satisfies the rules of NQAXML. (3)

END OF QUIZ