START OF QUIZ
Student ID:
92642370,Chen,Tiffany

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 5
Source: Lecture 5

Describe the purpose of the various types of brackets in regexes, and how they differ. (1)

Question 3

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 4

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 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

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 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 8

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