START OF QUIZ
Student ID:
95507984,Li,Qihan

Question 1

Topic: Lecture 6
Source: Lecture 6

Why is XML well-suited to representing linguistic data? (1)

Question 2

Topic: Lecture 5
Source: Lecture 5

In class, I mentioned that we can use regexes to identify typos by finding letters that are close on the keyboard. What functionality of regexes would we exploit to identify these types of characters? Give a brief example. (2)

Question 3

Topic: Lecture 6
Source: Lecture 6

Beautiful Soup parses the children of a tag as a list, why do you think they didn't use a set, instead, given the faster access times? Give 2 reasons, and briefly explain. (1)

Question 4

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 5

Topic: Lecture 5
Source: Lecture 5

Imagine that we had a phonetically-transcribed poem (or song). How could we use regexes to identify the rhyme scheme ((since not all of you are familiar with phonetic transcription, you can just describe the logic)? You can assume that each line is written on a new line, and that it is written in stanzas of 4 lines each. List any assumptions. (2)

Question 6

Topic: Lecture 7
Source: Lecture 7

What implications does correct sentence segmentation have on downstream tasks? List at least one assumption we can make if we can assume that our sentences are correctly segmented. (1)

Question 7

Topic: Lecture 8
Source: Lecture 8

If you have a corpus of 10 billion words stored in a text file tokenized with one word per line, what is best approach to processing the content of the file after it has been opened? (1)

Question 8

Topic: Lecture 7
Source: Lecture 7

What impact does lemmatization or stemming have on TTR? How might that affect our algorithms? (1)

Question 9

Topic: Long
Source: Long

Imagine that you trained a POS tagger on a corpus derived from an XML-annotated document, and you find your model always makes mistakes tagging a certain word. What would be your steps for discovering whether the model or the dataset were responsible for the error? Let's assume that the language is not one that you know. How would you confirm whether the data or the model were responsible? (3)

END OF QUIZ