START OF QUIZ
Student ID: 34157719,Philip,Reshmi

Question 1

Topic: Lecture 8
Source: Lecture 8

Why do Python programmers like JSON files so much? (1)

Question 2

Topic: Lecture 6
Source: Lecture 6

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

Question 3

Topic: Lecture 5
Source: Lecture 5

How would you use regexes to do sentence segmentation? Do you think it could correctly identify all cases? Explain. (1)

Question 4

Topic: Lecture 8
Source: Lecture 8

What is the purpose of an archive (2 reasons). (1)

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 6
Source: Lecture 6

Imagine that you're building a web scraper, and you find that most of the information presented on the front page is just a collection of links to other pages, so you can't just parse it with an XML parser. What extra functionality would you have to build into your scraper to actually get all the XML data? (2)

Question 8

Topic: Lecture 7
Source: Lecture 7

Do you think it would be easy or hard to build a training set for a sentence segmenter? Explain. Do you think it would be easier or harder to build a training set for a word tokenizer? What kind of assumptions would you be making about the difficulty of the task? You don't need to worry about the ML tool used - this is about building the training set. (2)

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