START OF QUIZ
Student ID:
36779478,Liao,Spencer

Question 1

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 2

Topic: Lecture 6
Source: Lecture 6

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

Question 3

Topic: Lecture 8
Source: Lecture 8

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

Question 4

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 5

Topic: Lecture 8
Source: Lecture 8

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

Question 6

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 7

Topic: Lecture 7
Source: Lecture 7

Why is part of speech relevant when lemmatizing? Do you think it would be just as important in stemming? (1)

Question 8

Topic: Lecture 6
Source: Lecture 6

What kind of tags do you think would enhance the following sentence? Annotate the sentence with at least two types of tags (they don't have to be accurate - just make sure I can understand the meaning): "Dr. Grant, my dear Dr. Sattler, welcome... to Jurassic Park!" (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