START OF QUIZ
Student ID:
71936512,Jimenez,Daniel

Question 1

Topic: Lecture 6
Source: Lecture 6

In class, we mentioned a few different file types that are actually XML (such as .html, .doc, and .ipynb). Do you think that you could represent a Python library as an XML document? If so, what kind of tags might you need to cover some of the syntactic rules of Python? If not, why not? (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 are 2 benefits of .py files over .ipynb files, and 2 benefits of .ipynb files over .py files? (1)

Question 4

Topic: Lecture 7
Source: Lecture 7

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

Question 5

Topic: Lecture 8
Source: Lecture 8

If you were working with an unknown language, which encoding would be most appropriate? Briefly explain. (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 5
Source: Lecture 5

List one advantage that regular expressions have over string comparison, and one disadvantage to using them. (1)

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