START OF QUIZ
Student ID:
84574284,Cheng,Yushun

Question 1

Topic: Lecture 8
Source: Lecture 8

If you were working with an unknown language, which encoding would be most appropriate? Briefly explain. (1)

Question 2

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 3

Topic: Lecture 7
Source: Lecture 7

I mentioned in class that POS tagging is often viewed as a pre-processing step for many CL tasks. What assumptions are we making (at least 3) when including it in our NLP pipeline? Do you think these are reasonable assumptions, and if they fail, is it worth the effort to solve the problem, or just ignore POS tagging? (2)

Question 4

Topic: Lecture 5
Source: Lecture 5

Write a short function that uses regexes to identify which case (title, camel, or sentence case) a sentence occurs in. (1)

Question 5

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 6

Topic: Lecture 5
Source: Lecture 5

In the last review set, there was a question about identifying valid floats using string operations. How would you do it with a regex? Explain the logic. (1)

Question 7

Topic: Lecture 8
Source: Lecture 8

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

Question 8

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 9

Topic: Long
Source: Long

Morphological Analysis is a process whereby we recover the lemma and morphologically-informed POS together. For example, the input might be "ran", and the output would be "run + VB;PAST". Do you think it would be best to 1. run tagging first, and then lemmatize using the tag 2. lemmatize first, and then tag, or 3. do both jointly? Why do you think one or the other would be more beneficial, and what information you be leveraging from one to help the other? Do you think this would be harder or easier for inflectionally-rich languages? Justify your answer. As always, state your assumptions. (3)

END OF QUIZ