START OF QUIZ
Student ID:
37157856,Wang,Lusha

Question 1

Topic: Lecture 5
Source: Lecture 5

Write a regex pattern that matches any valid email address (i.e., with basic rules like user@domain.com). What challenges might you face in accurately matching all possible email formats? (1)

Question 2

Topic: Lecture 8
Source: Lecture 8

Why do Python programmers like working with (t/c)sv files? When are they appropriate, and what advantages do they provide over .txt files? (1)

Question 3

Topic: Lecture 7
Source: Lecture 7

In class, we built a POS tagger that tries to give a majority tag to a word; if it's out-of-vocabulary, it backs-off to Regexes. This is clearly overly simplistic. List two assumptions that are being violated by this model. (1)

Question 4

Topic: Lecture 7
Source: Lecture 7

What is the difference between a stem and a lemma? What impacts does that have on our algorithms? (1)

Question 5

Topic: Lecture 5
Source: Lecture 5

Imagine you have a block of text with paragraphs separated by blank lines. How would you use regex to find the start of each paragraph? What assumptions would you make about the formatting of the text? (1)

Question 6

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 7

Topic: Lecture 6
Source: Lecture 6

Consider using XML to represent a machine learning model's architecture. What XML tags might be useful for representing layers, activation functions, and connections between layers (you don't need to describe a deep-learning architecture - describe one you're familiar with)? If this doesn't seem possible, explain why not. (2)

Question 8

Topic: Lecture 6
Source: Lecture 6

Suppose you've trained a Named Entity Recognition (NER) model using XML-annotated text data, but it consistently fails to recognize locations. What steps would you take to determine if the problem lies with the model, the training data, or both? What resources would you need to investigate further? (2)

Question 9

Topic: Long
Source: Lecture 8

Imagine that you find an important file buried on a hard drive found in the basement of a university. You are trying to access the data, but realize it is corrupted. Some of the bits have been flipped (switched from 0 to 1, or 1 to 0), and others have been completely deleted. You don't know the encoding, and you don't know the language the data is written in. What are some tests you could run to try to establish and restore at least some of the data? (Hint: remember that a "byte" is 8-bits, and that UTF-8 is 1 byte, or 8 bits, UTF-16 is 2 bytes, or 16 bits, and UTF-32 is 4 bytes, or 32 bits). (3)

END OF QUIZ