START OF QUIZ
Student ID: 90412503,Yin,Ting

Question 1

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 2

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 3

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 4

Topic: Lecture 6
Source: Lecture 6

How would we find all links in an HTML document? (1)

Question 5

Topic: Lecture 8
Source: Lecture 8

In class, I mentioned that we always want to close a file correctly. Beyond freeing up system resources, it also ”flushes the buffer”, which ensures that any current read or write operations that are in the job queue, but haven’t yet been processed, are completed. Knowing what you do about encodings, what is a possible ramification of not flushing the buffer? Explain at least 2. (2)

Question 6

Topic: Lecture 8
Source: Lecture 8

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

Question 7

Topic: Lecture 6
Source: Lecture 6

XML can be opened by most plain-text text editors. Name a benefit and a disadvantage of this feature. (1)

Question 8

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 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