

START OF QUIZ

Student ID:

76324177,Chiu,Hayden

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 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 3

Topic: Lecture 8

Source: Lecture 8

If you have a corpus of 10 billion words stored in a text file tokenized with one word per line, what is best approach to processing the content of the file after it has been opened? (1)

Question 4

Topic: Lecture 5

Source: Lecture 5

How would you use regexes to do sentence segmentation? Do you think it could correctly identify all cases? Explain. (1)

Question 5

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 6

Topic: Lecture 8

Source: Lecture 8

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

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

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

Question 9

Topic: Long

Source: Long

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)? (3)

END OF QUIZ