START OF QUIZ Student ID: 43887546,Kumar,Abhi

Question 1

Topic: Lecture 6
Source: Lecture 6

Imagine that you're building a web scraper, and you find that most of the information presented on the front page is just a collection of links to other pages, so you can't just parse it with an XML parser. What extra functionality would you have to build into your scraper to actually get all the XML data? (2)

Question 2

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 3

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 4

Topic: Lecture 6
Source: Lecture 6

Why is XML well-suited to representing linguistic data? (1)

Question 5

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 6

Topic: Lecture 8
Source: Lecture 8

What is the purpose of an archive (2 reasons). (1)

Question 7

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 8

Topic: Lecture 7
Source: Lecture 7

What implications does correct sentence segmentation have on downstream tasks? List at least one assumption we can make if we can assume that our sentences are correctly segmented. (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