

START OF QUIZ

Student ID:

11337250,Oloomi,Sina

Academic honesty is essential to the continued functioning of the University of British Columbia as an institution of higher learning and research. All UBC students are expected to behave as honest and responsible members of an academic community. Failure to follow the appropriate policies, principles, rules, and guidelines of the University with respect to academic honesty may result in disciplinary action.

I agree that all answers provided are in my own words, and that I will not discuss the contents of this quiz with any of my fellow students until after the exam period has completed for everyone. Furthermore, any response that used generative AI tools has been rephrased into my own interpretation, and has been appropriately cited.

Signature: _____

Question 1

Topic: Lecture 6

Source: Lecture 6

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

Question 2

Topic: Lecture 5

Source: Lecture 5

List one advantage that regular expressions have over string comparison, and one disadvantage to using them. (1)

Question 3

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 4

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 5

Topic: Lecture 8

Source: Lecture 8

What are two advantages of using .py files over .ipynb files for deployment, and two reasons why .ipynb files are preferred for prototyping or development? (1)

Question 6

Topic: Lecture 5

Source: Lecture 5

Imagine we have a spell-checker that can identify common misspellings of words by replacing certain letters with a capture group that contains letters that are nearby on the keyboard. How aggressive of a regex would we want to write for this (ie, how many letters in the word would we want to replace with a group)? Explain. (2)

Question 7

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 8

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 9

Topic: Long

Source: Lecture 5

In class, we've taken a brief look at both prefixes and suffixes, but there are other ways of inflecting words. "circumfixes" wrap around a word, such as the German past participle marker "ge-t" ("ich spiele" - "I play"; "ich habe gespielt" - I have played). Likewise, "infixes" occur inside of a word - "cupful" + Plural -> "cupsful", or in Tagalog: "bili" -> "to buy"; "bumili" -> "X is buying". Finally, "reduplication" occurs when part or all of a token is repeated to indicate some feature, such as repetition or future intent in Tagalog: "aray" -> "day"; "arayaray" -> everyday; "basa" -> "to read"; "babasa" -> "will read (in the future)". Which of these are best suited for regexes, and which features of regexes are they exploiting? Are there any that are mostly unsuited to regexes? Why? (3)

END OF QUIZ