# START OF QUIZ
# Student ID:
# 30542179,Yang,Yimei

## Question 1

Topic: Lecture 5
Source: Lecture 5

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

## Question 2

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 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

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 6
Source: Lecture 6

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

## Question 7

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 8

Topic: Lecture 7
Source: Lecture 7

What impact does lemmatization or stemming have with respect to the Zipfian curve? How might that affect our algorithms? (1)

## 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