# START OF QUIZ
# Student ID:
# 36304153,Kang,David

## Question 1

Topic: Lecture 8
Source: Lecture 8

Give 2 reasons to use a zip file. (1)

## Question 2

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 3

Topic: Lecture 5
Source: Lecture 5

There are two ways of matching a pattern against the start of a string. Describe them. (1)

## Question 4

Topic: Lecture 7
Source: Lecture 7

What might the training data for a sentence segmenter look like? Do you think it would be easy or hard to train? Explain briefly. (1)

## Question 5

Topic: Lecture 5
Source: Lecture 5

Imagine you have a block of text with paragraphs separated by blank lines. How would you use regex to find the start of each paragraph? What assumptions would you make about the formatting of the text? (1)

## Question 6

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 7

Topic: Lecture 6
Source: Lecture 6

Consider using XML to represent a machine learning model's architecture. What XML tags might be useful for representing layers, activation functions, and connections between layers (you don't need to describe a deep-learning architecture - describe one you're familiar with)? If this doesn't seem possible, explain why not. (2)

## Question 8

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 9

Topic: Long
Source: Lecture 7

Suppose you're building a text classification model for a highly inflected language like Finnish. How might you approach preprocessing tasks such as lemmatization or stemming? Would you perform these tasks before or after feature extraction, and why? Discuss how the choice of sequence may impact the quality of the features and model accuracy. Would you make the same decision for sentiment analysis? (3)

# END OF QUIZ