START OF QUIZ
Student ID:
19668508,Li,Julian

Question 1

Topic: Lecture 2
Source: Lecture 2

What is the impact of choosing a poor value for k in k-means clustering? How can we determine a more appropriate k? (1)

Question 2

Topic: Lecture 1
Source: Lecture 1

Suppose we are filling the table for the Levenshtein distance algorithm. We are in cell (x, y). The values of cell (x-1, y-1), (x-1, y), and (x, y-1) are 4, 3, and 3, respectively. What is the value we will put in cell (x, y), given that the letters are NOT equal? (1)

Question 3

Topic: Lecture 3
Source: Lecture 3

Describe the noisy channel model, and how it can be used to represent Machine Translation. (1)

Question 4

Topic: Lecture 2
Source: Lecture 2

Desribe the concept of cluster homogeneity, and how it relates to precision. (1)

Question 5

Topic: Lecture 1
Source: Lecture 1

When is dynamic programming more efficient than brute force programming? (ie, what assumptions do we make about a problem when we use dynamic programming?) (1)

Question 6

Topic: Lecture 3
Source: Lecture 3

Imagine that we have a trigram model that encounters a trigram where none of the tokens are in the vocabulary. How do you think that might impact our probability calculation for the sentence? How might we go about finding a solution? (2)

Question 7

Topic: Lecture 4
Source: Lecture 4

Imagine that we are doing OCR (optical character recognition; ie, the translation of handwritten text into digital text) instead of POS tagging. Do you think we could use an HMM? If so, what would the states, transitions, and emissions be? If not, describe why it's an inappropriate tool for the task. (2)

Question 8

Topic: Lecture 4
Source: Lecture 4

Let's imagine we're modifying our HMM to handle 2nd-order Markov operations (ie, consider the previous two states). Does anything in the model fundamentally change? Describe which aspects of the forward/Viterbi algorithm would need to be modified, if any. (2)

Question 9

Topic: Long
Source: Lecture 2

Imagine you are tasked with clustering social media posts to identify trends or topics. You have access to a large amount of unstructured text data. What kind of features do you think would be helpful, how would you preprocess the data, and how would you verify that the clustering is a good one? (3)

END OF QUIZ