START OF QUIZ
Student ID:
10237352,Shan,Zhengyi

Question 1

Topic: Lecture 4
Source: Lecture 4

How does semi-supervised learning differ from unsupervised and fully-supervised learning? (1)

Question 2

Topic: Lecture 2
Source: Lecture 2

Explain the purpose of a centroid in K-means clustering, and how we can think of it with respect to its cluster. (1)

Question 3

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 2, 2, and 2, respectively. What is the value we will put in cell (x, y), given that the letters are NOT equal? (1)

Question 4

Topic: Lecture 3
Source: Lecture 3

Describe the noisy channel model, and how it can be used to represent POS-tagging. (1)

Question 5

Topic: Lecture 1
Source: Lecture 1

Why is cosine distance typically a more suitable distance metric for semantic spaces than Euclidean distance? (1)

Question 6

Topic: Lecture 3
Source: Lecture 3

Imagine you were trying to pitch a new version of Scrabble to Hasbro that included "digraphs" (ie, combinations of two consecutive letters, like "th"). Do you think that you could score them as a simple combination of the single letter scores (ie, "th" is worth "t" + "h"), or would you need to do some more complex scoring calculations? Explain. (2)

Question 7

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 8

Topic: Lecture 2
Source: Lecture 2

What kinds of data might be difficult to cluster using k-means? Is it a shortcoming of the algorithm, or does it just need very careful feature engineering and distance calculations? (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