START OF QUIZ
Student ID:
54944541,Liu,Daoming

Question 1

Topic: Lecture 3
Source: Lecture 3

Explain the purpose of padding in language modeling. (1)

Question 2

Topic: Lecture 2
Source: Lecture 2

Describe the purpose of linkage in hierarchical clustering. (1)

Question 3

Topic: Lecture 4
Source: Lecture 4

How is it that EM can arrive at a good solution, even if we have a random initialization of parameters? (1)

Question 4

Topic: Lecture 3
Source: Lecture 3

When you were using Naive Bayes, a bag of words model was sufficient for classification. Why is it too simplistic for language modeling? (1)

Question 5

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 3, 4, and 3, respectively. What is the value we will put in cell (x, y), given that the letters are equal? (1)

Question 6

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 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 1
Source: Lecture 1

What is the primary concern of a semantic vector space (ie, a vector space representing meaning), and how does it relate to our use of cosine similarity to measure word similarity? Can you think of any sorts of words for which it might be very difficult to satisfy this concern? (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