

**START OF QUIZ**

**Student ID:**

**19756840,Wang,Rennan**

## Question 1

Topic: Lecture 1

Source: Lecture 1

Unsupervised Learning typically tries to find structure in unlabeled data. Give two reasons why we might want to still annotate a small dataset to use with our algorithm. (1)

## Question 2

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 3

Topic: Lecture 2

Source: Lecture 2

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

## Question 4

Topic: Lecture 2

Source: Lecture 2

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

## Question 5

Topic: Lecture 3

Source: Lecture 3

In your own words, explain the Markov assumption, and how it is used in language modeling.

(1)

## Question 6

Topic: Lecture 1

Source: Lecture 1

Let's consider a variant of the string alignment problem where instead of aligning characters, we're aligning sequences of characters (maybe we're doing machine translation...). What would need to be modified to handle a situation where we likely have a much higher vocabulary, and there's a lot less copying going on? What assumptions would we be making about the data? Would any of these assumptions make Levenstein distance inappropriate?

(2)

## Question 7

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 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 1

Do you think that auto-correct has a bias for where in a word an error occurs (ie, the index of the mistake)? If so, how might you approach fixing this problem? If not, explain why the position doesn't matter. As always, list any assumptions you're making. (3)

**END OF QUIZ**