

START OF QUIZ

Student ID:

92642370,Chen,Tiffany

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

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

Question 3

Topic: Lecture 3

Source: Lecture 3

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

Question 4

Topic: Lecture 4

Source: Lecture 4

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

(1)

Question 5

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 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 2

Source: Lecture 2

Imagine we were using k-means to cluster misspellings around their correct spellings. How many clusters would we need, and what would be a good distance function? Explain. (2)

Question 8

Topic: Lecture 4

Source: Lecture 4

Imagine that we are doing OCR (optical character recognition; ie, the translation of hand-written 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 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