START OF QUIZ
Student ID: 78097441,Tampubolon,Juan

Question 1

Topic: Lecture 7
Source: Lecture 7

From a processing perspective, what is one benefit structured data has over unstructured data, and vice versa. (1)

Question 2

Topic: Lecture 6
Source: Lecture 6

Why can't we just run an HMM over documents to discover the latent states like we do for POS-tagging? (1)

Question 3

Topic: Lecture 8
Source: Lecture 8

In class, I mentioned that TF-IDF unfairly punishes words that appear in a lot, but not all, of the documents in our corpus. Explain how Okapi BM25 attempts to fix this. (1)

Question 4

Topic: Lecture 6
Source: Lecture 6

Why don't we just use k-means to cluster document-vectors (sparse or dense)? (1)

Question 5

Topic: Lecture 5
Source: Lecture 5

What impact do sparse matrices have on similarity metrics like cosine similarity? (1)

Question 6

Topic: Lecture 7
Source: Lecture 7

When doing information retrieval, bag-of-words (and even just indicator functions) typically work very well. Explain why context is less important if we have a well-designed query. You may also want to explain your assumptions about a "well-designed" query. (2)

Question 7

Topic: Lecture 5
Source: Lecture 5

Why can we represent a rank-m matrix as the sum of m rank-1 matrices or the product of an n x m matrix and an m x n matrix (ie, what is matrix multiplication doing that we can take advantage of?)? Explain. (2)

Question 8

Topic: Lecture 8
Source: Lecture 8

In class (and in the lab) you saw some examples of using a language model for IR. How do you think we could incorporate an LLM into the IR pipeline? In what ways do you think an n-gram lm might be more appropriate? (2)

Question 9

Topic: Long
Source: Lecture 5

Imagine that we are working with a language other than English, such as Indonesian, with significant agglutinative morphology (words are inflected through the concatenation of affixes to a lemma). How do you think that this would impact our various vector space models? Which of them would be most affected, and which would be least affected? How might we go about adapting our model to solve these problems? Explain. (3)

END OF QUIZ