START OF QUIZ
Student ID: 91905067,Nong-bualang,Bon

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 5
Source: Lecture 5

Explain the logic behind the IDF part of TF-IDF (ie, why does it give higher weights to more ”interesting” words?). (1)

Question 3

Topic: Lecture 5
Source: Lecture 5

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

Question 4

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 5

Topic: Lecture 6
Source: Lecture 6

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

Question 6

Topic: Lecture 6
Source: Lecture 6

In some ways, we could consider Beta / Theta distributions themselves to be an embedding of a topic / document. Explain, and explain how we might be able to leverage that. (2)

Question 7

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 8

Topic: Lecture 8
Source: Lecture 8

What are some assumptions that we make when we are interpolating between a document and a corpus? When should we trust the corpus more, and when should we trust the document more? (2)

Question 9

Topic: Long
Source: Lecture 6

Imagine that we have a Beta distribution for each document, and a Theta distribution for each topic. We are at the Maximization state of EM write a short function that calculates the probability of a document, given these distributions. Pay special attention to edge cases and special considerations... (3)

END OF QUIZ