START OF QUIZ
Student ID:
33836685,LAM,HUI YIN

Question 1

Topic: Lecture 8
Source: Lecture 8

What is the reasonining behind substituting TF-IDF with Okapi BM25? (1)

Question 2

Topic: Lecture 5
Source: Lecture 5

The Frobenius norm looks very similar to a distance metric we've already observed. Explain which one. (1)

Question 3

Topic: Lecture 7
Source: Lecture 7

Explain why the cosine similarity between a document and query vector is roughly equivalent to adding up the TF-IDF scores of each word in the document that occurs in the query. (2)

Question 4

Topic: Lecture 5
Source: Lecture 5

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

Question 5

Topic: Lecture 8
Source: Lecture 8

In class, I mentioned that high k value for BM25 TF weigthing rewards documents with many, many instances of a term in them. Explain why that's the case. (2)

Question 6

Topic: Lecture 6
Source: Lecture 6

In class, we talked about bookstores and streaming algorithms classifying books / movies. How can we tell that they don't use a topic modeling algorithm (or, if you think they do, what would be some clues)? (1)

Question 7

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 8

Topic: Lecture 7
Source: Lecture 7

Explain why boolean filtering is usually insufficient for retrieval, and why we normally need some way of scoring the documents. (2)

Question 9

Topic: Long
Source: Lecture 7

Imagine that we have 2 information retrieval systems, and we are evaluating on the same test set, which has 10 relevant documents. The first system returns them in positions [1, 5, 7, 15, 25, 50, 60, 70, 71, 90]. The second returns the documents at positions [2, 3, 6, 8, 10, 62, 80, 83, 91, 95]. Make an argument for each system being better, and provide support for both. Explain which system you would rather use, and why. If there are any other considerations, list them. (3)

END OF QUIZ