START OF QUIZ Student ID: 88179403,Du,Yiyang

Question 1

Topic: Lecture 7
Source: Lecture 7

What is the benefit (in terms of efficiency) of placing the most discriminative search terms first in a boolean search? (1)

Question 2

Topic: Lecture 8
Source: Lecture 8

Why don't we use a higher-order language model to perform IR? (1)

Question 3

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 4

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 5

Topic: Lecture 5
Source: Lecture 5

Why do we need methods like t-SNE? (1)

Question 6

Topic: Lecture 7
Source: Lecture 7

Explain why the cosine similarity between a (TF-IDF-weighted) 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 7

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 8

Topic: Lecture 6
Source: Lecture 6

Imagine we performed LDA on the classes in this block. What might their Theta distributions look like? (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