

# START OF QUIZ

Student ID: 75287953, Apparaju, Ananya

## Question 1

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 2

Topic: Lecture 7

Source: Lecture 7

What is the purpose of an inverted index? (1)

### Question 3

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  $n \times m$  matrix and an  $m \times n$  matrix (ie, what is matrix multiplication doing that we can take advantage of)? Explain. (2)

## 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 6

Source: Lecture 6

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

## Question 6

Topic: Lecture 8

Source: Lecture 8

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

## Question 7

Topic: Lecture 8

Source: Lecture 8

Why do we not simply take the probability of a word given its document (maybe with smoothing added in)? (1)



## Question 8

Topic: Lecture 5

Source: Lecture 5

Why can we be confident that a low-rank approximation of a matrix contains the most important information in a document? (1)

## Question 9

Topic: Coding

Source: Coding

Write a function that returns the most likely  $n$  documents given a term-document matrix, a smoothing parameter, and a query. (3)

**END OF QUIZ**