W15 - Multiple Topics

Due 28 Apr at 23:59 **Points** 20 **Questions** 16
Available 17 Apr at 5:00 - 28 Apr at 23:59 **Time limit** None
Allowed attempts Unlimited

Instructions

Content and Background

This quiz relates to the following contents:

1. Counting Binary Trees

2. Intro to IR

3. Postings List

4. tf-idf weighting

5. Tries

6. Vector Space Model

It may also draw upon supporting knowledge and skills expected from a CS sophomore. Please make sure that you are up to date on the coursework before attempting the quiz.

Difficulty

This quiz is equivalent to an in-class exercise. Have pen and paper ready and be prepared to work on challenging problems.

Discussion

Please use Discussion forums to discuss any of the questions. Do not reveal your answers.

This quiz was locked 28 Apr at 23:59.

Attempt history

	Attempt	Time	Score
LATEST	Attempt 1	20 minutes	20 out of 20

Correct answers are hidden.

Score for this attempt: **20** out of 20

Submitted 28 Apr at 12:33

This attempt took 20 minutes.

Question 1

1 / 1 pts

The tallest Binary Tree with 63 nodes would have a height of

62 .

The shortest Binary Tree with 63 nodes would have a height of

5 .

Answer 1:

62

Answer 2:

5

Question 2

1 / 1 pts

I am a Full Binary Tree with 31 nodes. The following are two truths and a lie about me. Spot the lie.

- I have 16 leaves
- I have a height of 8
- I am symmetric

Question 3

1 / 1 pts

If we consider information retrieval using the index at the back of a book, what constitutes the corpus and a document for this task?

corpus — the book

document — a page in the book

Question 4

1 / 1 pts

An inverted index is necessary to perform information retrieval.

- True
- False

Question 5

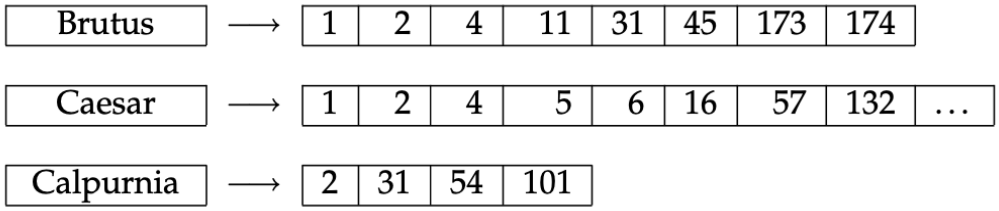
1 / 1 pts

In the absence of an index, what strategy would we employ to find a term in the book?

- Inspect every word in every page of the book.

Inspect a random page of the book and repeat if the query term is not found.

The next few questions refer to the the postings list below.



Question 6

1 / 1 pts

What are the returned results for the query: Calpurnia OR Brutus

Provide a comma separated list of sorted document IDs without spaces.

1,2,4,11,31,45,54,101,173,174

Question 7

1 / 1 pts

What are the returned results for the query: Brutus AND Caesar

Provide a comma separated list of sorted document IDs without spaces.

1,2,4

The next few questions relate to the corpus composed of the following documents.

Doc 1	breakthrough drug for schizophrenia
Doc 2	new schizophrenia drug
Doc 3	new approach for treatment of schizophrenia
Doc 4	new hopes for schizophrenia patients

Question 8

2 / 2 pts

Provide the postings list of each of the terms below.

drug: 1,2

schizophrenia: 1,2,3,4

Provide a comma separated list of sorted numerical document IDs without spaces.

Answer 1:

1,2

Answer 2:

1,2,3,4

Question 9

1 / 1 pts

Imagine a document containing terms of the following type. For each term, indicate the amount of boost that the document will get in the results if that term is queried for.

A term that occurs rarely in the corpus but frequently in the document.	High
A term that occurs rarely in the corpus and rarely in the document.	Medium
A term that occurs frequently in the corpus but rarely in the document.	Low
A term that occurs frequently in the corpus and frequently in the document.	Medium

Question 10

2 / 2 pts

We saw that the inverse document frequency (idf) score is defined as $\mathrm{idf}_t = \log \frac{N}{\mathrm{df}_t}$ where N is the size of the corpus. We want to see the value of this score for a term which has the maximum document frequency (df).

What is the maximum value of df_t : N

For that value, what is the idf score, $\mathbf{idf}_t$: 0

Answer 1:

N

Answer 2:

0

Question 11

1 / 1 pts

The formula for $\mathbf{idf}_t$ above is prone to error when $\mathbf{df}_t$ is 0. Why does that not prove to be a problem?

- We use a special condition to guard against that case.
- There is no problem with the formula when that occurs.
- This situation will never arise because such a term will never have to be indexed.

Imagine the set of strings,

S = {"bit", "byte", "bite", "bits", "bytes", "bites"}

and a delimiter/terminal character to mark the end of string.

Question 12

1 / 1 pts

How many internal nodes (nodes not containing a delimiter) are there in a standard trie on S?

11

Question 13

1 / 1 pts

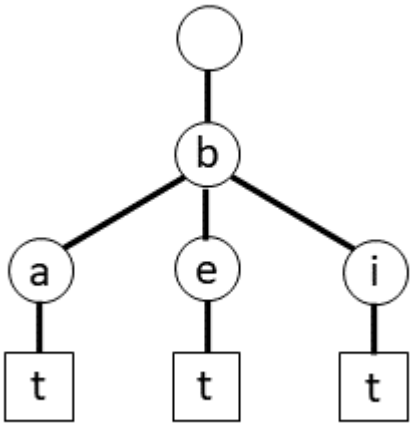
Including nodes with a delimiter, what is maximum number of children of any node in the standard trie on S?

3

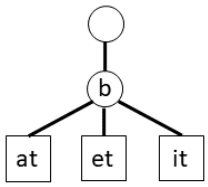
Question 14

2 / 2 pts

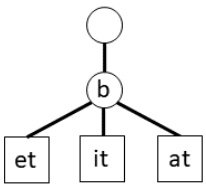
Given the following standard trie (T),



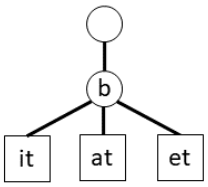
which of the following options correctly shows the compressed trie of T.



(a)



(b)



(c)

Mark all that apply.

- [x] Option (a)
- [x] Option (b)
- [x] Option (c)
- [ ] None of the options.

Question 15

1 / 1 pts

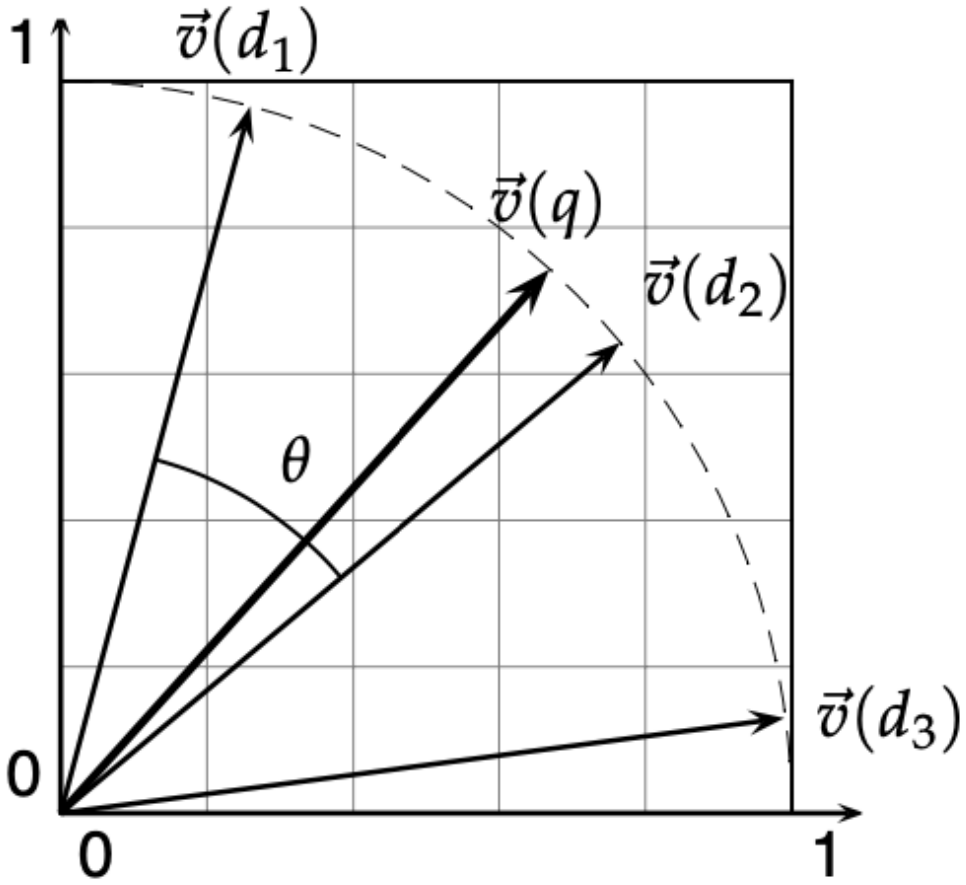
In a corpus, $C = \{d_1, d_2, d_3, \ldots, d_n\}$, given any arbitrary document, d_i , which is the document most similar to it?

d_i

Question 16

2 / 2 pts

In the following vector space illustration, why do the tips of all the vectors lie on a circular arc?



- It is just for illustration purposes.
- Because each vector represents a document.
- Because all the vectors are normalized to unit length.
- Because we are interested in the cosine similarity.

Quiz score: **20** out of 20