

GSI024 – Organização e Recuperação de Informação

Lista de exercícios 1

0.1

O que é recuperação de informação?

0.2

O que é o modelo booleano?

0.3

Qual a diferença entre modelo booleano e ordenação de resultados?

0.4

O que é índice invertido? Dê dois exemplos de índice invertido que não são encontrados em computadores.

0.5

Porque utilizar um índice invertido em vez de uma matriz de incidência termo-documento?

0.6

O que é um token, um termo e um documento?

0.7

Quais são e descreva os passos e operações necessários para realizar uma busca booleana na forma `casa AND carro`?

0.8

Qual é a ordem de complexidade do uso de uma operação `AND` em uma consulta booleana?

0.9

O que é um operador de proximidade? Quais modificações são necessária para um índice invertido ser capaz de processar um operador de proximidade?

0.10

Como é o índice posicional e quais são as vantagens de seu uso.

0.11

A seguir está mostrada uma parte de um índice posicional no formato:

`termo: docId: [pos1, pos2]; docId: [pos1, pos2];`

`tolos:2: [1, 17, 74, 222]; 4: [8, 78, 108, 458]; 7: [3, 13, 23, 193];`
`medo:2: [87, 704, 722, 901]; 4: [13, 43, 113, 433]; 7: [3, 13, 23, 193];`
`em:2: [3, 37, 76, 444, 851]; 4: [10, 20, 110, 470, 500]; 7: [5, 15, 25, 195];`
`correm:2: [2, 66, 194, 321, 702]; 4: [9, 69, 149, 429, 569]; 7: [4, 14, 404];`
`para:2: [47, 86, 234, 999]; 4: [14, 24, 774, 944]; 7: [199, 319, 599, 709];`
`caminhar:2: [57, 94, 333]; 4: [15, 35, 155]; 7: [20, 320];`
`onde:2: [67, 124, 393, 1001]; 4: [11, 41, 101, 421, 431]; 7: [15, 35, 200, 735];`

Liste os documentos obtidos para as seguintes consultas:

1. tolos AND correm AND "em medo"
2. caminhar AND "para onde"
3. "tolos correm"
4. "correm tolos"
5. correm AND tolos

0.12

Porque saber o idioma em que um texto foi escrito é importante em um sistema de organização e recuperação de informação? Quais são os cuidados que devemos ter quando processarmos diferentes idiomas?

0.13

Dados os seguintes caracteres, discuta o problema potencial de tokenização de cada um deles.

- Hewlett-Packard
- guarda-civil
- co-autoria
- base de dados
- senhor sabe-tudo
- São Paulo
- Los Angeles-based company
- torneio Rio-São Paulo
- Atlético e Cruzeiro

0.14

O que são stop-words? Porque elas podem ser um problema para indexação? Em quais situações devemos considerá-las?

0.15

O que é uma classe de equivalência de tokens?

0.16

O que é lematização? Porque ela é importante?

0.17

O que são biwords? Porque eles são necessários e úteis para os usuários de um sistema de recuperação de informação?

0.18

O que são consultas com caractere curinga? Quais são as principais formas de implementar tais consultas?

0.19

Quando é possível usar hashes em vez de árvores para implementar busca tolerante?

0.20

O que é um índice permuterm e como utilizá-lo para possibilitar buscas do tipo `car*o`. Para que serve o símbolo \$?

0.21

O que é um índice k-grama, como é o seu funcionamento para a busca a seguir `cartucho impressora`. Como fica essa busca após ser transformada em 3-grama? Como fica um índice invertido com 3-grama? Qual é a diferença desse índice invertido para o índice invertido de documentos?

0.22

Quais são as vantagens de utilizar correção ortográfica para recuperação de informação?

0.23

Como funciona a distância de Levenshtein porque utilizá-la para correção ortográfica?

0.24

Compute a matriz de edições de Levenshtein para as palavras "carro" e "onibus". Descreva passo a passo as edições realizadas para transformar a primeira palavra na segunda.

0.25

Para que servem os algoritmos BSBI e SPIMI. Quais são suas diferenças e semelhanças?

0.26

Porque usar construção de índice distribuída? Qual é o principal custo que será distribuído? Quais são as atividades que podem ser paralelizadas?