GSI024 – Organização e Recuperação de Informação
Lista de exercícios 3

0.1

Descreva com detalhes a estrutura da web.

0.2

Como fazer um bombardeio de links bem sucedido? Como um sistema ORI pode evitar bombardeio de links?

0.3

O que é PageRank? Porquê o PageRank é útil para rankear documentos da web?

0.4

Apresente um exemplo de cadeia de Markov ergódica com 5 estados.

0.5

Como fazer para transformar um grafo representando a web em uma cadeia de Markov ergódica?

0.6

Use a seguinte lista de links de documentos para obter o PageRank.

```
d0 -> d1 d8 d3 d6
d1 -> d3 d4 d2
d2 -> d1 d8 d5
d3 -> d5
d4 -> d4 d7
d5 -> d6 d8
d6 -> d4 d5 d8 d1
d7 -> d0
d8 -> d4
```

0.7

Quais são os potenciais problemas do uso do PageRank para a web?

0.8

Como um spammer pode fazer para tentar aumentar o PageRank de suas páginas?

0.9

Como funciona o algoritmo HITS?

0.10

Quais são as similaridades e diferenças entre o algoritmo PageRank e o HITS?

0.11

Porquê uma boa página hub e quase sempre uma boa página de autoridade?

0.12

Use a seguinte lista de links de documentos para obter o ranking de autoridades e hubs para uma consulta que retornou inicialmente dos documentos $d0$, $d1$ e $d2$.

```
d0 -> d1 d8 d3 d6
d1 -> d3 d4 d2
d2 -> d1 d8 d5
d3 -> d5
d4 -> d4 d7
d5 -> d6 d8
d6 -> d4 d5 d8 d1
d7 -> d0
d8 -> d4
```

0.13

Escreva um parágrafo que explique como funciona um buscador Web para uma pessoa leiga em computação mas que já usou um buscador antes.

0.14

O que é um crawler?

0.15

Quais são os problemas do crawler descrito pelo seguinte algoritmo?

```
filaUrls := (urls de inicialização bem escolhidas)
enquanto filaUrls não está vazia:
  url := filaUrls.pegarUltimoEremover()
  pagina := url.capturar()
  urlsCapturados.adicionar(url)
  novosUrls := pagina.extrairUrls()
  para cada url em novosUrls:
    se url não está em urlsCapturados e nem em filaUrls:
      filaUrls.adicionar(url)
  fim para
  adicionarParaIndiceInvertido(pagina)
fim enquanto
```

0.16

Como fazer uma armadilha web para um crawler simples descrito no exercício anterior?

0.17

Porque pode ser importante fazer um crawler distribuído?

0.18

Para que serve o arquivo `robots.txt`?

0.19

Como funciona a fronteira de URLs de Mercator? Descreva o funcionamento das filas dianteiras, traseiras e da heap.

0.20

Porquê é importante utilizar Aprendizado de Ranking?

0.21

Sendo a função de pontuação para um par j de documento e consulta definida por $pontuacao(d_j, q_j) = g \cdot s_T(d_j, q_j) + (1 - g) \cdot s_B(d_j, q_j)$, função de erro $\epsilon(g, \Phi_j) = (r(d_j, q_j) - pontuacao(d_j, q_j))^2$ e a função de erro total $\sum_j \epsilon(g, \Phi_j)$. Encontre o parâmetro g para minimizar o erro total segundo os seguintes exemplos de treino.

Exemplo Φ_j	s_T	s_B	julgamento
1	0	1	1
2	1	0	1
3	1	1	1
4	1	1	0
5	0	1	0
6	0	0	0
7	1	0	0
8	1	0	1

0.22

Descreva cada elemento da anatomia do sistema de busca do protótipo do Google.