UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
Bacharelado em Ciência da Computação e
Engenharia da Computação
INF 01124 – Classificação e Pesquisa de Dados

ÍNDICE INVERTIDO

Considere os documentos abaixo sobre a *Copa do Mundo de 2010* para resolver os exercícios a seguir.

DOC 01 Fabricante se diz surpresa com críticas à bola da Copa
DOC 02 Felipe Melo põe Argentina atrás do Brasil na Copa e vira espelho de Dunga
DOC 03 Seleção completa quatro jogos em três treinos e mostra Dunga "carrasco"
DOC 04 Seleção fica no 0 a 0 em primeiro coletivo na África; gols só na bola parada

01. Mostre a *tokenização* para os documentos apresentados.

Em amarelo a tokenização.

DOC 01 Fabricante se diz surpresa com críticas à bola da Copa

DOC 02 Felipe Melo põe Argentina atrás do Brasil na Copa e vira espelho de Dunga

DOC 03 Seleção completa quatro jogos em três treinos e mostra Dunga "carrasco"

DOC 04 Seleção fica no 0 a 0 em primeiro coletivo na África; gols só na bola parada

02. Mostre a eliminação das *stop words* (termos discriminates) para os documentos apresentados.

Stop words eliminadas em vermelho

DOC 01 Fabricante ~~se~~ diz surpresa ~~com~~ críticas ~~à~~ bola ~~da~~ Copa
DOC 02 Felipe Melo põe Argentina atrás ~~do~~ Brasil ~~na~~ Copa ~~e~~ vira espelho ~~de~~ Dunga
DOC 03 Seleção completa quatro jogos ~~em~~ três treinos ~~e~~ mostra Dunga "carrasco"
DOC 04 Seleção fica ~~no~~ 0 ~~a~~ 0 em primeiro coletivo na África; gols só ~~na~~ bola parada

Para quem eliminou os zeros também está correto

A eliminação de stop words tem como objetivo filtrar palavras com valores discriminatórios baixos para a tarefa de recuperação de informação. Ou seja, diminui o numero de palavras que serão usadas para montar o índice.

03. Cite 5 exemplos de normalização que poderiam ser feitos para a construção do índice invertido (em documentos em geral e não somente com base nesse exemplo).

- Eliminação de acentos
- Deixar tudo em letras minúsculas
- Eliminar erros de ortografia
- Eliminação de pronomes
- Eliminação de caracteres especiais

.

04. Faça a análise de freqüência para os 6 primeiros termos.

Fabricante : DOC1

Diz : DOC1

Surpresa : DOC1

Criticas : DOC1

Bola : DOC1, DOC4

Copa : DOC1, DOC2

05. Desenhe o índice invertido para a coleção de documentos apresentada (utilize somente os 6 primeiros termos).

Fabricante 1

Diz 1

Surpresa 1

Criticas 1

Bola 2

Copa 2

06. Desenhe o dicionário de termos para a coleção de documentos apresentada (utilize somente os 5 primeiros termos).

Abaixo um exemplo – a solução pode ser diferente

Palavras do índice	Freqüência	Apontador para a localicao física da lista que contem os documentos onde a palavra aparece
Fabricante	1	001
Diz	1	002
Surpresa	1	003
Criticas	1	004
Bola	2	005

Arquivo físico (em disco) onde está armazenada a lista com o indice

Termos do indice	Lista de documentos
Fabricante	DOC1
Diz	DOC1
Surpresa	DOC1
Criticas	DOC1
Bola	DOC1, DOC4

07. Considerando o seu índice invertido, quais documentos retornariam as seguintes consultas? Qual o caminho percorrido no índice para responder as essas consultas?

a. Copa busca seqüencial no índice até encontrar a palavra copa e acesso a lista de documentos que retornaria DOC1 e DOC4
b. Brasil busca seqüencial no índice até encontrar a palavra brasil e acesso a lista de documentos que retornaria DOC2

c. Copa AND Brasil

busca seqüencial no índice até encontrar as palavras copa e brasil e acesso a lista de documentos dessas duas palavras. Somente o DOC 2 é retornado porque tem o AND

d. Gols OR Dunga

busca seqüencial no índice até encontrar as palavras gols e dunga e acesso a lista de documentos dessas duas palavras. A consulta retorna os DOC2, DOC3 e DOC4 porque tem um OR

e. Brasil AND Argentina

busca seqüencial no índice até encontrar as palavras Brasil e argentina e acesso a lista de documentos dessas duas palavras. Somente o DOC 2 é retornado porque tem o AND

08. Qual a importância da ordenação dos termos do índice? Dê um exemplo.

Permite que se faça busca binária no índice ao invés de seqüencial. Mais eficiente.

09. Qual a importância da ordenação dos documentos na lista invertida? Dê um exemplo.

Permite exibir os documentos em uma determinada ordem. Uma solução mais inteligente seria ordena-los pela ordem decrescente de freqüência. Assim, os documentos possivelmente mais "relevantes" apareceriam primeiro.

10. Faça uma análise comparativa entre armazenar a freqüência dos termos na lista de termos ou armazenar a freqüência de termos para cada documento? Como essas informações poderiam ser utilizadas?

Armazenar a freqüência na lista permite saber qual a palavra é mais freqüente.

Armazenar a freqüência de termos em cada documento permite saber qual o documento pode ser mais "relevante" por conter mais vezes a palavras em questão.

11. Qual a vantagem (utilidade) de se armazenar a posição em que cada termo ocorre dentro dos documentos? Como isso poderia ser implementado (Estrutura de dados)?

Vantagem: pode mostrar parte do documento e não o documento inteiro. Mais difícil de implementar porque é necessário além da tokinização dividir o documento em pedaços e esse tamanho deve ser decidido no momento da implementação.

12. Quais as vantagens e desvantagens do uso de índice invertido?

Vantagem: facilita a recuperação de documentos em aplicações de busca por palavra-chave

Desvantagem (ou dificuldade): a cada novo documento que entra para a coleção o indice precisa ser refeito