



## Trabalho prático de ECO029

Entregar: 10/04/2014 às 23:59h valor: 15 pontos

---

Neste primeiro trabalho iremos construir uma máquina de busca utilizando **hash com tratamento de colisão**.

### O trabalho

O sistema deverá ser capaz de criar um hash em um diretório contendo vários arquivos e apresentar uma interface contendo todas as possibilidades de busca que serão listadas abaixo.

### Descrição do trabalho

Como ocorre na web, os arquivos do diretório que são permitidos a realização de busca estão listados dentro do arquivo "robots.txt". O sistema deve processar os arquivos listados em "robots.txt" e construir, a partir dos arquivos processados, o hash (**os arquivos originais não podem ser alterados**). Em seguida, o sistema deve apresentar uma interface de busca contendo as seguintes opções:

- Busca por palavra
- Busca por frase

Após realizar a busca, o sistema deve apresentar os resultados ordenados pela melhor relevância, contendo **um número**, o **nome do arquivo** e um **trecho do texto** (uma linha) que contenha a palavra ou a frase consultada. **OBS: o resultado da busca deve ser apresentado utilizando o arquivo original.**

Abaixo temos dois exemplos:

#### Resultado da busca "casa"

1 – imóveis.txt

Uma **casa** ou uma residência, é no sentido

2 – casamento.txt

Todo casal recém-casado deseja ter uma **casa**

#### Resultado da busca "casa praia"

1 – imóveis.txt

Alugasse uma **casa de praia** no valor de R\$150,00

2 – casamento.txt

Todo casal que viaja para a **praia** aluga uma **casa**

### Busca por palavra

O sistema deve receber uma palavra e recuperar do hash todas as ocorrências. O resultado da busca deve ser apresentado de forma ordenada pela **melhor relevância**. No caso de busca por palavra, os arquivos com maior relevância são os que possuem **a maior incidência da palavra**.

### Busca por frase

O sistema deve recuperar do seu hash todas as ocorrências de todas as palavras que compõem a frase. O resultado da busca deve apresentar somente os textos que contenham todas as palavras da frase e de forma ordenada pela **melhor relevância**. Abaixo apresentamos as regras de relevância para busca por frase na ordem do mais relevante para o menos relevante.

1. Relevância alta: textos que possuem a sequência de palavras na mesma ordem da frase.
2. Relevância baixa: textos que possuem as palavras fora de ordem.

## O arquivo robots.txt

Neste trabalho iremos utilizar um arquivo simples do tipo “robots.txt” que conterá somente os nomes dos arquivos que podem ser utilizados para busca dentro do diretório.

### Exemplo de um arquivo robots.txt

casa.txt carros.txt
------------------------

## O processamento dos arquivos

O processamento deve colocar todo o texto em minúsculo, deve retirar todos os acentos das palavras e deve substituir todos os caracteres que não pertencem ao alfabeto (  $\Sigma = \{a, b, c, \dots, z\}$  ) pelo caractere vazio.

A ideia de tirar o acento das palavras é permitir realizar busca com e sem acento da mesma palavra.

### Hash

O Hash deve ser implementado utilizando lista encadeada como tratamento de colisão. Cada célula do hash irá armazenar uma palavra e as informações necessárias para realizar as buscas propostas neste trabalho.

## Entrega do trabalho

Deve ser enviado para o email **rafafic@gmail.com** até o dia 10/04/2014 às 23:59 h o relatório em PDF, os códigos fontes criados na linguagem C/C++ e o diretório compactado contendo os arquivos de teste (mínimo de 5 arquivos e cada arquivo deve conter no mínimo 10 linhas). **Os arquivos de teste não podem estar processados, pois o seu sistema deve fazer isso ao construir o hash.** Dica: retire textos da internet. **Não é para entregar o relatório impresso.**

## Relatório

1. Capa;
2. Descrição do trabalho;
3. Descrição das estruturas de dados utilizadas;
4. Explicação de como foi implementado o sistema;
5. Testes realizados (no mínimo 5 com busca por palavra e 5 com busca por frase). Apresente os testes através de prints da tela ou através de cópias do cmd.
6. Conclusão;
7. Bibliografia utilizada.

## Entrevista do trabalho

A entrevista acontecerá no dia 11/04/2014 (sexta-feira) durante a aula.

**Lembrete:** Cópia de qualquer parte do trabalho é 0 (ZERO) para quem copiou e para quem emprestou o trabalho.