MAC–IME–USP

Carlos Eduardo Ferreira
Sala 108C Tel.: 3091 6135
e-mail cef@ime.usp.br
monitor Bruno Klava

MAC 323 – Estruturas de Dados

Primeiro semestre de 2005

Exercício-Programa 4 – Entrega: **23 de junho de 2005**

Corretor ortográfico

O objetivo deste exercício-programa é construir um programa que funcione como um corretor ortográfico, e verificar empiricamente o funcionamento das estruturas de dados aprendidas na sala de aula.

Corretores ortográficos são programas que recebem como entrada um texto e um dicionário e verificam quais palavras deste texto não se encontram no dicionário. Um programa deste tipo, escrito pelo D. Knuth, pode ser visto na página:

`http://www.ime.usp.br/~yoshi/2001i/mac323/EPs/EP2/wordtest/`

O programa está escrito em CWEB, a linguagem de programação predileta de Knuth. Se você quiser aprender mais de CWEB, veja as excelentes notas de aula do Prof. Feofiloff:

`http://www.ime.usp.br/~pf/algoritmos_em_grafos/`

Você certamente pode estudar o programa de Knuth para fazer este EP.

A organização do dicionário é a parte fundamental deste EP, que pretendemos analisar com mais detalhes. Uma palavra é uma seqüência maximal de letras ou "-" (para permitir palavras como "guarda-chuva"). O dicionário deverá ser implementado de pelo menos três formas (fique à vontade para fazer de outras formas também):

- árvore de busca binária, sem balanceamento (as palavras entram na árvore em ordem aleatória);
- árvore de busca binária balanceada (use a estrutura do EP3);
- tabela de hash (implemente a sua! Teste várias funções de hash).

Faça uma análise empírica dos resultados de buscar as palavras de um texto grande no seu dicionário organizado conforme estas estruturas. Entregue junto com seu EP um pequeno relatório com os textos feitos e resultados obtidos.

Seu EP deverá manipular textos grandes (use páginas da rede como entrada de seu programa). Como dicionário para o português você pode usar o que está disponível em

`http://www.ime.usp.br/~ueda/br.ispell/`

O `ispell` é um aplicativo do unix para correção ortográfica, e sua versão para o português do Brasil é mantida por Ricardo Ueda. Na página acima você obtém informações de como obter uma lista de palavras do português (aprox. 2.8M). É importante notar que este dicionário está disponível com a licença GNU GPL. Assim, qualquer trabalho produzido com base nele deve ficar também disponível para distribuição.