Universidade Federal do Espírito Santo

Departamento de Computação e Eletrônica - CEUNES
Estrutura de Dados I
Prof. Oberlan Romão

Exercício Programa 3

Verificador Ortográfico

O objetivo desse EP é comparar a utilização de *árvores binárias de busca* (balanceada vs. não balanceada). Para isso, você deve completar/implementar um programa que faça a leitura de um dicionário de palavras em português e um texto. Seu programa deve "marcar" as palavras escritas incorretamente, ou seja, que não estão no dicionário fornecido. Em seguida, ele deve listar essas palavras (sem repetí-las) e exibir algumas sugestões (como ocorre nos editores de texto modernos).

As palavras do dicionário devem ser carregadas e armazenas em uma árvore binária de busca (convencional ou AVL, dependendo do argumento passado ao programa). Já o conjunto de palavras escritas incorretamente, podem ficar em qualquer estrutura de sua escolha, mas a exibição delas deve obedecer a ordem de ocorrência, ou seja, se no texto a palavra ABC aparece antes da palavra XYZ, na exibição das palavras incorretas, ABC deve vir antes de XYZ. Além disso, na listagem de palavras incorretas **não** deve haver repetição. A figura abaixo exibe um exemplo de execução do programa. Note que a ordem das palavras incorretas é a mesma do texto e que a palavra "dicionario" (sem acento) aparece apenas uma vez na listagem de palavras incorretas.

```
Arquivo  Editar  Ver  Pesquisar  Terminal  Ajuda
❯ ./EP3 -d DicionarioEP3.txt -t EP3.txt
O objetivo desse EP é comparar a utilisação de arvores binarias de busca (balanceada vs. não balanceada). Para isso, você deve completar/implementar um programa
que faça a leitura de um dicionario de palavra em português e um texto. Seu programa deve "marcar" as palavras escritas incoretamente, ou seja, que não estão no
dicionario fornecido. Em seguida, ele deve listar essas palavras (sem repetí-las) e ezibir algumas sugestoes (como ocorre nos editores de texto modernos).



----------------------------------------
-      Número de palavras lidas: 74
- Número de palavras incorretas: 10
Palavra(s) incorreta(s) e sugestão(ões)
----------------------------------------
EP: AL, AC, AI, AC, AK, AR, APS, APL, AZ, BS, BC, BP, CPR, CA, CB, CEQ, CEQ, CO, CPA, CZ, CT, DB, DEC, DC, DES, De, Dr, EDP, ECG, EDT, EEG, EGA, EUA, EPA, EST, E
d, Eli, Ely, Ema, Eng, Esp, FM, Eva, Eva, Fl, FPC, Fe, Ft, GS, GOP, GM, GA, GE, GNP, GPO, GU, HP, Hz, IA, ID, IEE, IM, Il, IR, IQ, Io, Jo, Jr, KY, KS, Ku, M, MD,
 MN, MO, MPH, Mt, Mr, Ms, NC, NJ, NH, NM, NW, NP, NV, NY, OK, OEM, OS, Oz, P, PC, PM, PDP, PR, Ph, Pl, Po, QED, QA, RI, RPM, SC, SD, SE, SW, TV, Sr, TN, TA, T, T
CP, TV, TX, UK, UN, UT, VA, VT, WV, WA, WI, WY, Wu, a, á, ab, ad, ah, ah, ai, am, an, ao, ar, as, as, at, ax, b, be, by, c, cc, cf, d, d, da, de, do, do, du, e,
e, eh, el, em, em, en, et, eu, ex, ex, f, g, go, h, ha, he, hi, ho, ia, i, ie, if, ii, ih, im, in, ir, is, it, iv, ix, j, k, la, l, la, lb, li, lo, lo, ma, ma, m
e, me, mi, ml, mo, mu, my, na, n, nd, ne, no, no, nu, o, o, of, oh, oh, oi, om, on, op, or, os, ou, ow, ox, pa, pi, q, r, rd, re, ri, sa, s, se, si, si, so, st,
te, th, ti, ti, to, tu, u, uh, ui, um, up, us, v, vi, vi, vs, w, we, x, xi, xi, xv, xx, y, ye, yr, z, à, é, ô,
comparar: compara, comparai, comparar, comparam, comparara, comparas, comparer, comprar, conjurar, consagrar, conspirar, contaram, contara, contaras, contatar, c
ooperar,
utilisação: utilização,
binarias: abanarias, afinarias, atinarias, babarias, bailarias, baixarias, bancarias, banharias, banirias, beijarias, beirarias, bicaras, bicaria, bicariam, bica
rias, bigamias, binaries, binárias, bipartas, bipartias, bizarras, blindarias, boiarias, bolarias, botarias, brigarias, brincarias, brindarias, bufarias, buzinar
ias, citarias, danarias, dignarias, ditarias, fiarias, ficarias, filarias, fincarias, findarias, fintarias, fitarias, fixarias, gingarias, girarias, lidarias, li
garias, limarias, lixarias, miarias, mijarias, mimarias, minaria, minaras, minariam, minarias, minorias, mirarias, mixarias, nanarias, ninaras, ninarias, ninaria
, ninariam, ninharias, opinarias, ornarias, penarias, piarias, picarias, pifarias, pingarias, pintarias, pisarias, reinarias, rifarias, rimarias, sanarias, tinir
ias, tirarias, uivarias, urinarias, vincarias, vingarias, virarias, visarias, xingarias,
completar/implementar:
dicionario: adicionarmo, adicionaria, dicionário,
incoretamente: incorretamente, indiretamente,
repetí-las:
ezibir: coibir, elixir, elixir, emitir, erigir, exibia, exibi, exibir, exibira, exibis, exibiu, exigir, eximir, inibir,
sugestoes: sugastes, sugestiones, sugestões,
----------------------------------------
Tempo Total: 3.4104690000 seg
```

Sugestões de palavras

Como você deve ter observado na figura anterior, para cada palavra escrita errada, é exibida uma lista de sugestões (algo comum em muitos editores de textos). Para o EP, usaremos o algoritmo de Edição de Distância (Edit distance) para definir as sugestões. Esse algoritmo explora uma técnica chamada Programação Dinâmica (vista com detalhes na disciplina de Projeto e Análise de Algoritmos). Esse algoritmo tem complexidade $O(n \times m)$, onde n e m são os tamanhos dos strings. No arquivo `Util.c` você encontra uma implementação do algoritmo.

De forma resumida, a função `int distanciaEdicao(char *a, char *b)` recebe dois `strings` e retorna a "distância" entre eles, ou seja, quantas mudanças (inserções, substituições e remoções) são necessárias para que um `string` fique igual ao outro. Por exemplo, `distanciaEdicao("uma", "uva")` retornaria 1 (basta trocarmos o caractere `"m"` por `"v"`). Dessa forma, quanto menor o valor retornado mais "próximo" dois `strings` estão. Para o EP, ao listar as sugestões, você deve definir um limite para que uma palavra seja considerada "sugestão" de uma palavra errada. Claro que quanto maior esse limite, maior será o número de palavras sugeridas.

Perceba que para cada palavra errada, seu programa deve percorrer toda a árvore procurando por sugestões. Pense em como limitar o número de chamada da função `distanciaEdicao` (chamando-a apenas quando fizer sentido). Você também pode dar maior prioridade as palavras que tenham uma distância menor a palavra errada e limitar o número de palavras sugeridas (por exemplo, exibir, no máximo, as 5 "melhores" sugestões).

Tarefa

Você deve fazer uma análise empírica de desempenho do seu programa, comparando a utilização de uma árvore binária de busca sem balanceamento e a árvore AVL. Utilize diferentes arquivos de textos com um número variado de palavras erradas. Além disso, você deve fazer um relatório técnico com a metodologia adotada para a realização dos experimentos e análise e discussão dos resultados. Faça tabelas e/ou gráficos para te ajudar na análise.

Dicionários. Você pode utilizar os dicionários (em português e inglês) disponibilizados no arquivo `EP3.zip` ou procurar por outros na internet. Sinta-se livre para manipular os dados dos dicionários (por exemplo, embaralhar ou reordenar as palavras), mas relate, no relatório, o que e como foi feito e o porquê dessa modificação. Caso utilize outros dicionários, envie-os na entrega do EP. Não se esqueça de discutir a complexidade das principais funções usadas no EP. Exemplos de discussões que podem/devem ser feitas no relatório:

- Por quê a construção/inserção na árvore considerando um determinado arquivo com palavras do dicionário demora tanto tempo?
- O que poderia ser feito no arquivo para melhorar esse tempo? O quanto isso melhora? A complexidade muda?
- Usar uma árvore balanceada melhora o tempo de execução? Porquê?

Textos. Teste seu programa com textos de tamanhos variados. Envie também os arquivos usados nos experimentos. Uma boa fonte de arquivos para testar o seu programa é o Projeto Gutenberg.

O que entregar

Você deve entregar, pelo **AVA**, um arquivo compactado contendo todos os códigos da sua implementação, uma arquivo `Makefile` (basta completar o que está no arquivo `EP3.zip`) e um relatório (no formato `.pdf`). Envie também os textos e dicionário utilizados.

Data de entrega: até às 6h do dia 04/10/2021.

Observações:

1. Sinta-se livre para usar outros dicionários ou modificar o disponibilizado, mas deixe claro o que foi feito/utilizado no relatório;
2. Você também pode (e deve) exibir outros dados no programa para te auxiliar no relatório. Por exemplo, o tempo para construir a árvore, o tempo para listar as palavras com as sugestões, a altura da árvore gerada, etc. Seja criativo;
3. Você deve adicionar outros arquivos ao projeto. Pense em como organizá-lo em diferentes TADs, cada um com seus arquivos `.h` e `.c`;
4. Nos nós das árvores ou de outras estruturas que contiverem `strings`, você deve utilizar alocação dinâmica (`char *`). Algo assim:

```
typedef struct no {
    char *<nomeCampo>;
    // Outros campos
} No;
```

   Códigos que não respeitem essa restrição (utilizando alocação estática) receberão 70% da nota da implementação;
5. Preferencialmente, use o Linux (ou o CS50 IDE) para a implementação. No Windows, você pode usar o Code::Blocks para auxiliar na compilação do projeto;
6. Seu código deve ser compilado com as flags:

   ```
   -O0 -std=c11 -Wall -Werror -Wextra -Wno-sign-compare -Wno-unused-parameter -Wno-unused-variable -Wshadow
   ```
7. Códigos com erros de sintaxe (que não compilem) receberão nota 0;
8. **A compilação será feita usando o `Makefile`, se você não enviá-lo ou ele estiver incompleto, sua nota na implementação será 0;**
9. Código com vazamento de memória e/ou falha de segmentação, valerá 70% da nota da implementação;
10. Código que não segue o Guia de Estilo, valerá 90% da nota da implementação;
11. Em caso de plágio, será atribuído 0 a todos os envolvidos.

Critérios de avaliação

A nota geral do EP será da seguinte forma:

- **Implementação**: 5.0 pontos;
  - Organização: 1.0 ponto (separação dos códigos em `.h` e `.c`);
  - Implementação da Árvore Binária de Busca (não balanceada): 1.5 pontos;
  - Implementação da Árvore AVL: 1.5 pontos;
  - Outras estruturas de dados: 1.0 ponto;
- **Relatório técnico**: 5.0 pontos (caprichem). O relatório deve conter:
  - Capa;
  - Introdução
    - Faça uma breve introdução sobre o problema abordado e como ele está sendo resolvido;
  - Metodologia adotada para os experimentos;
  - Resultados:
    - Nessa seção você deve fazer uma análise e discussão dos experimentos feitos;
    - Faça a análise para diferentes textos com número variado de palavras corretas e incorretas;
    - Analise e discuta a complexidade das principais funções usadas no EP;
    - Discuta o que pode ser feito para deixar a execução do programa mais eficiente;
    - Inclua tabelas e/ou gráficos (preferencialmente, feitos com o Matplotlib) comparando os resultados;
    - Faça outras discussões e comparações que achar interessante.
  - Conclusão.