Enunciado do Exercício 3 de PPP

O CLion, com a opção add_compile_options(-Wall -Wextra), não deve gerar qualquer aviso sobre o código fonte, exceto os que não possam ser resolvidos com a matéria lecionada. Para fazer esta verificação o CLion deve ter o clang-tidy ativado na configuração standard, ou o clang-tidy deve ser executado na linha de comandos "clang-tidy ficheiro.c" (ficheiro.c é o ficheiro que está a ser verificado) tendo previamente colocado o ficheiro .clang-tidy nessa diretoria. O ficheiro .clang-tidy pode ser obtido de https://git.dei.uc.pt/snippets/31

O interface das estruturas de dados deve ser bem definido, para haver abstração - os programas que usam as estruturas de dados não podem manipular diretamente as variáveis de suporte às estruturas de dados, só os módulos que implementam as estruturas de dados.

A biblioteca lib-utf8 pode ser usada livremente.

O trabalho deve usar corretamente ficheiros header sempre que um programa seja constituído por vários ficheiros fonte.

Nos programas elaborados, as funções auxiliares do programa principal devem estar num ficheiro fonte autónomo, com um interface bem definido.

As funções que concretizam cada uma das estruturas de dados usadas devem estar em ficheiro fonte autónomo.

Só os ficheiros fonte, bem como o ficheiro CMakeList.txt (ou equivalente), devem ser submetidos como resultado do trabalho.

O objetivo do trabalho é permitir procurar palavras num texto em português, codificado em utf-8, e mostrá-las em contexto.

Há duas fases, cada uma delas implementada por um programa autónomo:

A primeira para ler o ficheiro, isolar as palavras e criar um ficheiro com as palavras encontradas.

A segunda para ler o ficheiro produzido na primeira fase e criar um índice em memória, que permite mostrar em contexto as palavras que o utilizador queira analisar. Mostrar em contexto significa mostrar a palavra em conjunto com palavras que a precedem e lhe sucedem. O utilizador também deve poder pedir para serem listadas todas as palavras que começam por uma determinada letra, ou intervalo de letras.

Primeira fase:

O programa obtém o nome do ficheiro de texto a processar como parâmetro da linha de comandos

Percorre depois o ficheiro de texto, isolando cada palavra e anotando a posição (número de bytes desde o início do ficheiro da primeira letra da palavra) em que essa palavra surge no texto.

As palavras são constituídas apenas por letras, algarismos e hífenes; todos os outros carateres devem ser ignorados e considerados separadores.

Só deve considerar palavras com mais de três carateres.

À medida que identifica uma palavra vai escrevendo num ficheiro essa palavra (preservando as maiúsculas/minúsculas, acentos e cedilhas) acompanhada da sua posição no texto.

O nome do ficheiro onde escreve as palavras é igual ao ficheiro que contém o texto, mas com o prefixo "tab_"

O ficheiro é um ficheiro de texto, com um par palavra - posição por linha, separados por ponto e vírgula. Por exemplo, 'trabalho;347' se houver uma ocorrência da palavra 'trabalho' a começar no tricentésimo quadragésimo sétimo byte do ficheiro de texto.

Segunda fase:

O programa pede ao utilizador o nome do ficheiro de texto.

O ficheiro com os pares palavra-posição tem um nome construído como na primeira fase.

O programa lê os pares palavra-posição desse ficheiro e constrói um índice, sob a forma de uma árvore binária. Como cada palavra pode surgir várias vezes no mesmo texto, em cada nó da árvore binária é guardada a raiz para uma lista ligada que contém todas as posições em que a mesma palavra surge.

Considera-se que se trata da mesma palavra se a única diferença forem acentos, cedilhas, ou maiúsculas/minúsculas.

O índice, ou seja, a árvore binária e a lista associada a cada nó, deve ser construído com recurso a memória dinâmica.

Por cada palavra que o utilizador escolha devem ser mostradas todas as ocorrências, dentro do seu contexto, por ordem decrescente da posição em que ocorrem no ficheiro.

O contexto são 10 palavras para trás e 5 palavras para a frente.

O texto não deve ser lido integralmente para memória: a posição de cada palavra, contida no índice, deve ser usada para ler do ficheiro de texto, diretamente, a parte a mostrar como contexto de cada palavra. .l. .l. .l. .l.

O utilizador deve ainda poder pedir para serem listadas todas as palavras que comecem por uma letra, ou gama de letras. Neste caso devem ser listadas apenas as palavras, sem repetições nem contexto.