SCC-501 - Introdução à Ciência de Computação II - 2014

Trabalho 4

Professor: Diego Raphael Amancio

DATA LIMITE - 05/12/2014
Entrega no run.codes

Frequência de palavras em textos

Seu programa utilizará uma tabela hash com encadeamento de listas para contar o número de ocorrências de palavras em um arquivo de entrada. O algoritmo executa, para cada palavra:

1. Realiza uma busca na tabela hash
2. Se a palavra não existe na tabela, seu programa insere a palavra com o valor de 1 de ocorrência. Nesse caso, você deverá criar um novo nó e ligá-lo à uma posição da tabela hash.
3. Se a palavra já existe na tabela, incremente o valor de ocorrência e não crie um novo nó.

A escolha de m: a tabela deve armazenar no máximo 3000 palavras distintas, cada uma com nó máximo 64 caracteres. Para escolher o número de compartimentos m em tabelas hash com encadeamento, divida o número máximo de elementos por 5 e tome o número primo próximo do resultado. Exemplo: $3000/5 = 600$, portanto m=599 ou m=601.

A função hash: programe uma função e o mapeamento de compressão conforme achar melhor.

As m listas encadeadas: o programa deve manter cada lista encadeada ordenada alfabeticamente pela palavra. Você poderá escolher como realizar essa ordenação. Isso será importante para obter a saída (ordenada) posteriormente.

Saída

Será composta de todas as palavras no arquivo que possuam 3 ou mais caracteres e 3 ou mais ocorrências no texto, uma por linha, com a contagem logo a seguir de cada palavra. A saída deverá estar em ordem alfabética. Para isso você pode utilizar uma função baseada na intercalação das listas que, conforme gera a saída ordenada, libera a memória ocupada por cada nó de cada lista encadeada.

Entrada

Será composta de uma série de palavras conforme descrito anteriormente. Palavras válidas são compostas por letras e possuem no mínimo 2 caracteres. Não é preciso verificar a corretude da entrada, todas as palavras estarão em letras minúsculas, não serão acentuadas, e o texto não irá conter pontuação, parênteses, hifens ou outros caracteres especiais.

O projeto será avaliado de acordo com:

1. Processamento correto das entradas e saídas do programa;
2. Realização das tarefas descritas;
3. Bom uso das técnicas de programação;
4. Bom uso das estruturas de dados e algoritmos;
5. Bom endentação, clareza e uso de comentários relevantes.

Restrições:

1. Não use variáveis globais.
2. Use hash estático com encadeamento aberto
3. As seguintes funções deverão ser obrigatoriamente implementadas:

   `int hash_code( char* key)`: recebe a chave e retorna o código hash.

   `void insert_ht( Hash table[], char *key)`: verifica se a palavra já existe na tabela. Caso não exista, inserir a palavra de forma ordenada na tabela com 1 ocorrência. Caso já exista, encontrar a palavra e incrementar o seu número de ocorrências.
4. Não poderão ser utilizadas bibliotecas com funções prontas, exceto: `string.h` , `stdio.h` e `stdlib.h`.