Classificação de palavras em ordem alfabética utilizando o método de Ordenação Externa

Nikolas Machado Corrêa

nmcorrea@inf.ufsm.br

1. Introdução

O algoritmo implementado tem por objetivo classificar palavras (representadas por strings) de um arquivo de texto fornecido como entrada em ordem alfabética. Para isso, foi utilizado um método de ordenação externa do tipo "divisão e conquista", que consiste em particionar o arquivo original em sub-arquivos e ordená-los aos poucos, uma vez que a implementação supõe que a entrada não possa ser carregada de uma vez só na memória principal do dispositivo.

2. Algoritmo

Para a implementação do programa fez-se necessário um estudo dos métodos de ordenação interna e externa, visto que o código utiliza ambos; por isso, primeiramente foi desenvolvida uma versão de teste baseando-se em dados do tipo *int,* visto que cada dado possui um tamanho fixo de 4 bytes na memória, o que facilita o particionamento e a divisão destes em sub-arquivos. A linguagem de programação escolhida foi C++.

Para o processo de implementação com strings foram criadas funções para extração e armazenamento temporário das palavras, descartando caracteres especiais, pontuação e acentos, que por sua vez, foram ordenadas internamente utilizando a função *std::sort* da biblioteca <*algorithm*> e agrupadas em arquivos, obedecendo o limite de dados por partição definido por 200 bytes. Vale ressaltar que os sub-arquivos não possuem um número fixo de strings uma vez que o tamanho de cada uma delas é variável, pois cada caractere conta como 1 byte, diferentemente do tratamento de dados que é feito com números inteiros.

Após a separação e divisão em subproblemas, utilizou um método de ordenação externa que: abre dois dos arquivos temporários e compara duas palavras (uma de cada um) por vez, unindo, ao final da ordenação, os dois arquivos abertos em um só; isso foi realizado sucessivamente até que todos os arquivos temporários acabassem fundidos em uma única saída, gravada como *"txt_ordenado.txt"* na pasta dos códigos-fonte. Este processo baseado em *dividir-para-conquistar* não utilizou a recursão pois o objetivo principal é ocupar apenas parte da memória, algo que não aconteceria devido às chamadas recursivas.

As funções e dados comuns a mais de uma dela foram armazenados em uma *struct,* declarada no arquivo *"ordenacao.hpp"*.

Detalhando melhor o programa, foram criadas 11 funções no total, responsáveis por:

(1) abrir o arquivo de entrada;

(2) efetuar a **ordenação interna** utilizando:

(3) extração de palavras,

(4) criação de sub-arquivos ordenados;

(5) efetuar a **ordenação externa** utilizando:

(6) abertura das subpartições,

(7) comparação das palavras de cada uma, duas a duas,

(8) gravação dos dados nestas e unificação;

Também foram utilizadas funções auxiliares para:

(9) verificar se arquivo está aberto e não chegou ao fim,

(10) calcular tamanho da partição em bytes,

(11) remoção dos arquivos temporários.

Observações:

1) Não foram utilizados comentários nas funções do código pois utilizou-se uma técnica de *Clean Code* que consiste em utilizar nomes autoexplicativos para dispensar a necessidade de explicá-los; o número de funções também é devido a uma técnica, pois o ideal é que cada função faça apenas uma coisa;

2) A linha para compilação encontra-se no arquivo *ordenacao.hpp* assim como a forma de execução para abrir o arquivo de texto que será utilizado como entrada.

3. Aplicações

Todos os algoritmos de ordenação têm por finalidade facilitar a recuperação de dados posteriormente por meio de uma classificação; a ordenação externa é utilizada em casos em que se trabalha com um grande volume de dados e que se tornam inviáveis de serem organizados internamente, por isso, o algoritmo se aplica a vários problemas que necessitam dessa organização, como, por exemplo, a organização das palavras em uma espécie de dicionário (fichário de nome); recuperação de nomes de uma lista telefônica, etc.

Referências

Stack Overflow – *Website* - https://stackoverflow.com/

C++ Reference – *Website* - http://www.cplusplus.com/

Notas de aula – *PDF* – https://github.com/joao-lima/notas-de-aula-POD-UFSM

Ordenação Externa – *Vídeo* - https://www.youtube.com/watch?v=sVGbj1zgvWQ