Trabalho Prático 1

Clara Costa da Fonseca - 2021031971

1. Introdução:

O trabalho descrito a seguir foi realizado na linguagem C++, seguindo como referência a página da Wikipédia fornecida na descrição do trabalho e o a página do site medium How Data Compression Works: Exploring LZ78 | by Dhanesh Budhrani | The Startup | Medium.

2. Divisão de Arquivos:

Foram necessários apenas dois arquivos, um .hpp contendo a forma da estrutura de dados e quais funções seriam usadas e um .cpp com a implementação do main e das funções.

a. A Estrutura de Dados:

A estrutura de dados foi uma Trie chamada No* contendo um char para armazenar os caracteres do padrão, um int para armazenar o índice do nó anterior, de quem o nó atual é filho, e um vector de No* com os filhos do nó atual.

b. As Funções:

Foram criadas quatro funções para a implementação do algoritmo. As funções **insere** e **busca** foram apenas uma forma de auxiliar a função **comprime** e tornar o código mais organizado.

A função **comprime** é a que de fato realiza a compressão do texto, nela, a raiz da árvore é declarada e um No* auxiliar (**dicio**) é criado para realizar o caminhamento da árvore. O texto é lido caractere a caractere, **busca** retorna para um segundo No* auxiliar (**dicio2**) se encontrou um outro No* que contém o caractere lido ou então, nullptr se não o encontrou. O primeiro caso leva **dicio** a se igualar a **dicio2** e caminhamento na árvore em busca do último No* onde o caractere foi inserido continua; o segundo caso gera o chamado de **insere** e o caractere e seu índice (inicializado em uma variável igual a 1 e é acrescido de + 1 a cada inserção) são passados. Cada tupla <índice, caractere> inserida é passada para um vector auxiliar (**texto**) que, ao final da leitura do texto, será utilizado para escrever o texto comprimido no arquivo de saída. Após todo o texto ser lido, o índice da raiz, que contém um caractere vazio como padrão, é passado para o arquivo.

A função **descomprime** lê o arquivo de saída gerado pela função comprime, lê os números das tuplas, lê seus caracteres em seguida e guarda em um vector auxiliar (**texto**) o caractere de texto na posição do número lido mais o caractere lido, a medida que se anda no texto, a palavra original é reconstruída.

A função **main** é utilizada apenas para ler a linha de comando, preparar os arquivos de leitura/saída e chamar **comprime** e **descomprime**, segundo o comando passado — se é "-c" ou "-x".

3. Exemplos:

Foram utilizados 13 textos para testar o código, dentre eles os três fornecidos juntos da descrição do tp. De modo geral, o algoritmo foi capaz de comprimir os arquivos, no entanto — como é o caso de três dos exemplos —, ele se mostrou ineficiente para arquivos pequenos.

Uma hipótese para tanto é que a eficiência diz respeito à existência de padrões repetidos, não somente ao tamanho do arquivo. O arquivo com a melhor taxa de compressão não foi o maior arquivo, mas sim o da Constituição, afinal ela é um texto extremamente padronizado, no qual, a título de exemplo, a palavra "Artigo" se repete milhares de vezes. Ademais, o arquivo **filipinas.txt** tinha 158 KB e o **grecia.txt,** 164 KB, o primeiro permaneceu com 158 KB após a compressão, mas o segundo aumentou de tamanho, para 169 KB. Mas vale ressaltar que a descompressão obteve êxito em todos os casos.

a. brasil.txt:

Tamanho original: 58 KB; tamanho comprimido: 67 KB
Taxa de compressão: 67/58 ≅ 1,15

b. constituicao1988.txt:
Tamanho original: 637 KB; tamanho comprimido: 435 KB
Taxa de compressão: 435/637 ≅ 0,68

c. dom_casmurro.txt:
Tamanho original: 401 KB; tamanho comprimido: 362 KB
Taxa de compressão: 362/401 ≅ 0,90

d. filipinas.txt:
Tamanho original: 158 KB; tamanho comprimido: 158 KB
Taxa de compressão: 1

e. grecia.txt:
Tamanho original: 164 KB; tamanho comprimido: 169 KB
Taxa de compressão: 169/164 ≅ 1,03

f. guarani.txt:
Tamanho original: 364 KB; tamanho comprimido: 327 KB
Taxa de compressão: 327/364 ≅ 0,89
g. guerra.txt:
Tamanho original: 381 KB; tamanho comprimido: 347 KB
Taxa de compressão: 347/381 ≅ 0,91

h. ingles.txt:
Tamanho original: 774 KB; tamanho comprimido: 641 KB
Taxa de compressão: 641/774 ≅ 0,82

i. norte.txt:
Tamanho original: 834 KB; tamanho comprimido: 691 KB
Taxa de compressão: 691/834 ≅ 0,82

j. novelas.txt:
Tamanho original: 312 KB; tamanho comprimido: 291 KB
Taxa de compressão: 291/312 ≅ 0,93

k. os_lusiadas.txt:
Tamanho original: 337 KB; tamanho comprimido: 318 KB
Taxa de compressão: 318/337 ≅ 0,94

l. policarpo.txt:
Tamanho original: 425 KB; tamanho comprimido: 387 KB
Taxa de compressão: 387/425 ≅ 0,91

m. tempo.txt:
Tamanho original: 326 KB; tamanho comprimido: 301 KB
Taxa de compressão: 301/326 ≅ 0,92