UNIVERSIDADE REGIONAL DE BLUMENAU - FURB
CENTRO DE CIÊNCIAS EXATAS E NATURAIS
DEPARTAMENTO DE SISTEMAS E COMPUTAÇÃO
DISCIPLINA: Algoritmos e Estruturas de Dados

Prof. Aurélio Hoppe

TRABALHO PARCIAL 02 - COMPRESSÃO DE DADOS

[codificação de Huffman]

Objetivos

O objetivo deste trabalho é utilizar o algoritmo de compactação de mensagens de Huffman para criar arquivos binários (ou texto) compactados.

Descrição

O algoritmo de Huffman sugere um esquema de codificação para o alfabeto de uma mensagem em função da frequência de cada símbolo. Essa codificação é representada através de uma árvore binária.

Código Huffmann

O ASCII é um esquema de codificação de caracteres dito de tamanho fixo, já que todos os caracteres representados neste código têm o mesmo número de bits. A ideia subjacente ao código Huffman é que a frequência com que aparecem os caracteres num texto dito o tamanho da palavra do seu código, sendo possível associar um código "curto" a caracteres que apareçam com muita frequência e códigos mais "longos" a caracteres menos frequentes. A tabela da figura apresenta a frequência relativa das 26 letras do alfabeto em um texto representativo para o Inglês:

Letra	Frequência	Letra	Frequência
A	77	**N**	67
B	17	**O**	67
C	32	**P**	20
D	42	**Q**	5
E	120	**R**	59
F	24	**S**	67
G	17	**T**	85
H	50	**U**	37
I	76	**V**	12
J	4	**W**	22
K	7	**X**	4
L	42	**Y**	22
M	24	**Z**	2

Pela tabela acima temos a confirmação de que determinadas letras, por exemplo as vogais, possuem maior frequência que outras, as consoantes J, Z ou Q. Como usar este conhecimento para construir códigos com tamanho de palavra variável?

O código Huffman de cada letra é derivado de uma árvore binária designada por árvore binária de Huffman ou simplesmente por árvore de Huffman. Cada folha numa árvore Huffman corresponde a uma letra.

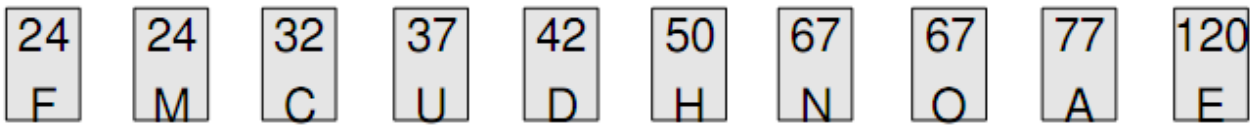
Considera se que pretendemos a construção de uma árvore Huffman para as letras:

A	C	D	E	F	H	M	N	O	U
77	32	42	120	24	50	24	67	67	37

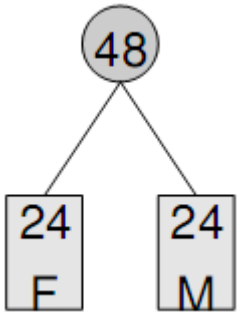
Oordenando as letras por ordem crescente do seu peso (frequência) tem-se:

F	M	C	U	D	H	N	O	A	E
24	24	32	37	42	50	67	67	77	120

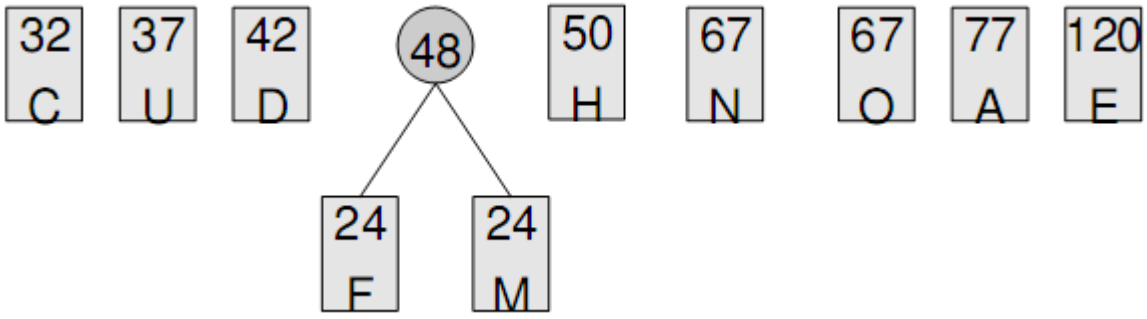
É possível visualizar cada uma das letras como sendo uma árvore Huffman:



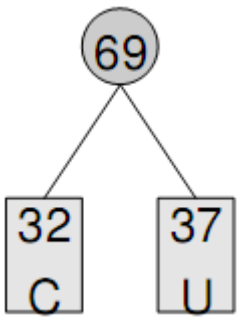
Estas árvores isoladas podem ser reunidas para formar uma nova árvore Huffman usando o seguinte algoritmo: pegue as primeiras duas árvores da lista e agrupe-as em uma árvore cujas folhas são as árvores selecionadas e o nó pai desta folhas é um nó interno de peso igual à soma do peso dos seus filhos:



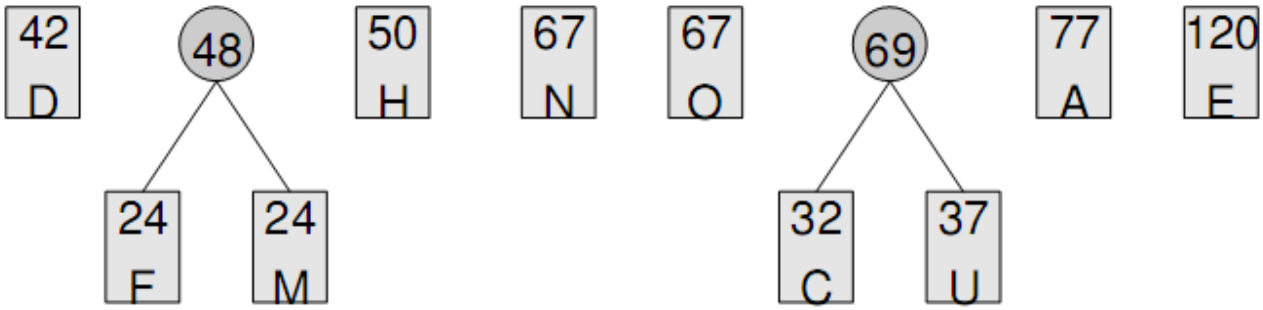
A nova árvore será reintroduzida na lista, na posição correspondente (respeitando a ordenação) e o processo será repetido novamente para as duas primeiras árvores da lista:



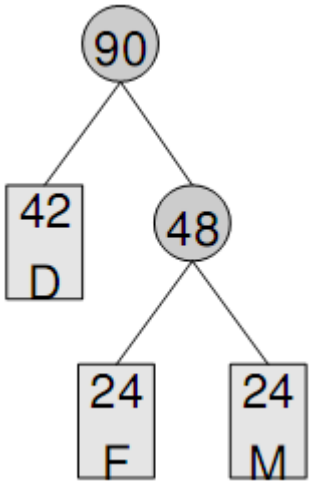
Aplicando novamente o passo do algoritmo temos como árvore resultante:



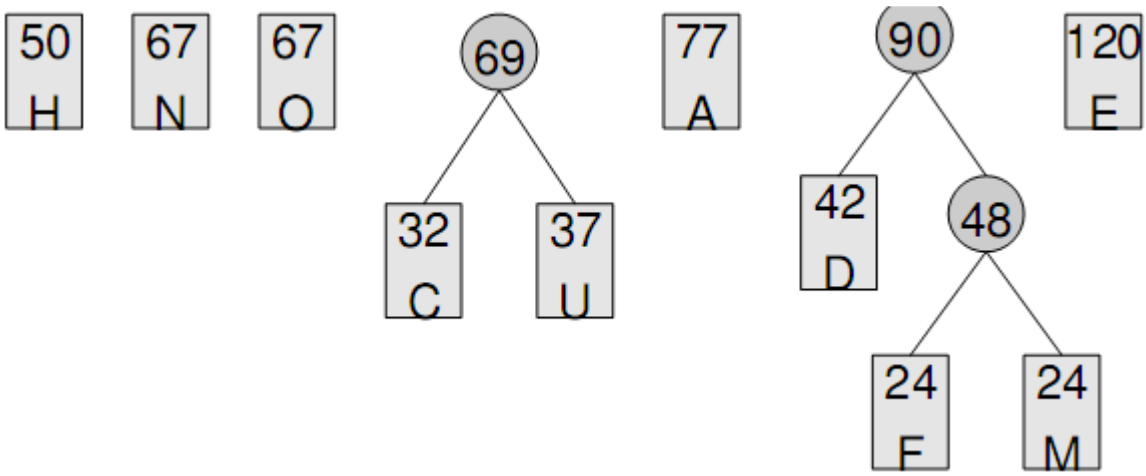
E introduzindo a nova árvore na lista temos:



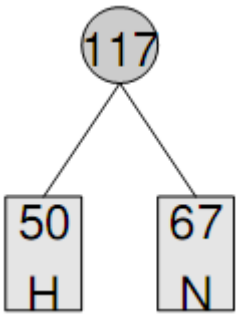
No próximo passo é gerada a árvore:



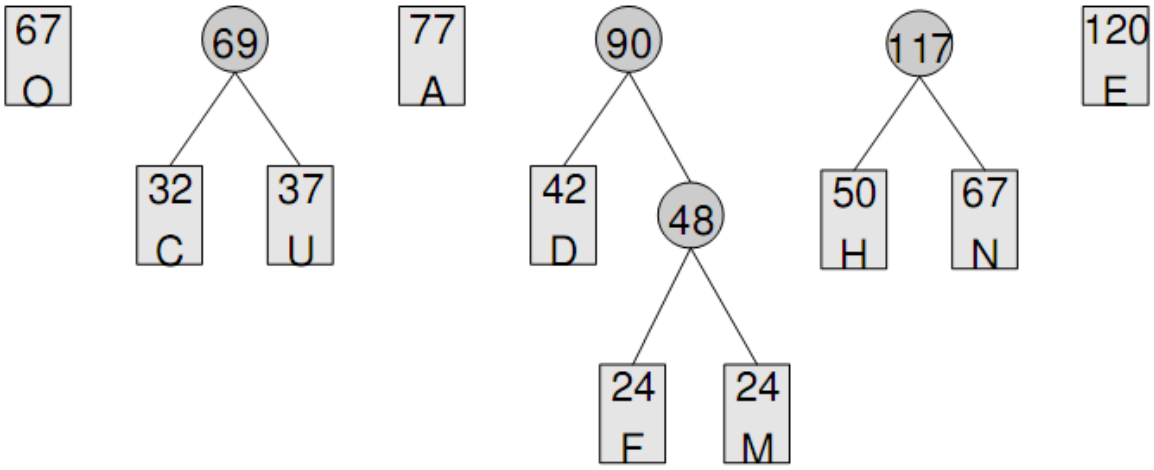
E a lista:



Agora a aplicação do passo seguinte do algoritmo gera a árvore:

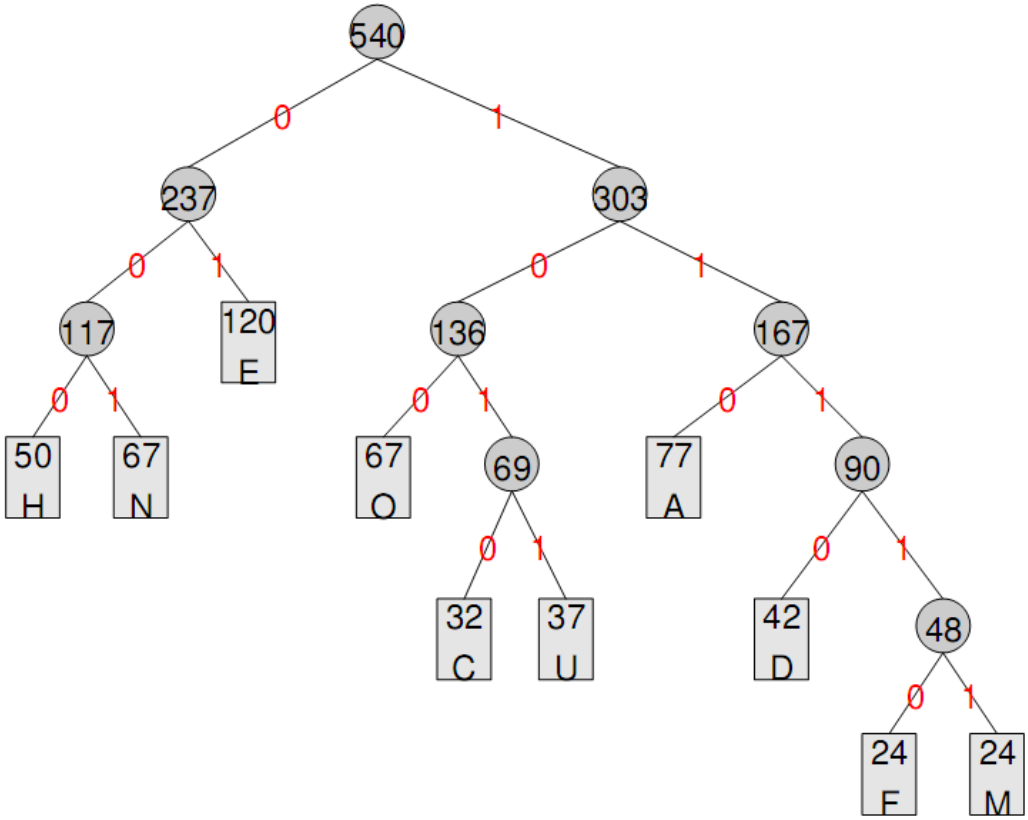


E a lista:



Letra	Código Huffman	Frequência
H	000	50
N	001	67
E	01	120
O	100	67
C	1010	32
U	1011	37
A	110	77
D	1110	42
F	11110	24
M	11111	24

O processo será repetido até à obtenção da árvore de raiz com peso 540 (Figura abaixo). A atribuição de códigos aos caracteres após a construção da árvore é um processo simples: começa-se pela raiz e atribui-se a cada ramo um '0' ou um '1' consoante se trate de um ramo esquerdo ou direito respectivamente. O código Huffman de uma letra é o binário correspondente ao caminho na árvore Huffman desde a raiz até à folha correspondente. A tabela ao lado apresenta as letras usadas para a construção da árvore, o código Huffman e a respectiva frequência. É possível constatar que tal como o inicialmente anunciado, letras com menos frequência têm códigos de menor comprimento e vice-versa



A codificação de mensagens usando os códigos Huffman é possível substituindo cada caracter da mensagem inicial pelo respectivo código. Assim, por exemplo “Huffman” será codificado por:

000101111110111011111110001

A descodificação de mensagens faz-se percorrendo os bits da mensagem a descodificar da esquerda para a direita até à exaustão da mesma e percorrendo a árvore de Huffman usando o caminho “dado” pela mensagem: quando atingimos uma folha temos uma letra da messagem descodificada. Recomeça-se novamente até esgotarmos todos os bits da mensagem. Por exemplo para descodificar 1000011110110 toma-se o ramo direito da árvore, esquerdo e esquerdo novamente, atingindo-se a folha correspondente à letra “O”. Seguidamente pega-se os ramos esquerdo, esquerdo e direito, atingindo-se a folha do “N”. A descodificação da mensagem resulta em “ONDA”.

1. Implemente um programa que faça a compressão de dados baseado na codificação de Huffman, conforme descrito no material disponibilizado pelo professor.
2. Faça um relatório contendo uma tabela com o tempo e a taxa de compressão de cada arquivo de entrada. Ao final, faça uma análise e/ou justifique os resultados obtidos, indicando motivos, vantagens e desvantagens.

Observações:

1. o trabalho é individual. A interpretação do enunciado faz parte da avaliação;
2. o principal instrumento de avaliação do relatório será a discussão sobre os resultados obtidos;
3. os programas-fonte e o relatório devem ser postados no AVA até o dia **10/05/2014.**