Khoa Công nghệ thông tin
Đại học Khoa học tự nhiên TP HCM

BÁO CÁO ĐỒ ÁN 2

Cấu trúc dữ liệu và giải thuật

1712078 - Ngô Phan Nhật Lâm
1712212 - Lý Thiên Ân

MỤC LỤC

1 Mức độ hoàn thành

Nội dung	Thành phần	Mức độ hoàn thành
Nén tập tin bằng thuật toán Huffman	Nén tập tin chuỗi (.txt)	100%
	Nén folder gồm nhiều tập tin kiểu chuỗi	100%
	Nén tập tin chứa các kiểu dữ liệu còn lại (.cpp, .jpg,…)	100%
	Nén folder gồm nhiều loại tập tin	100%
Nén tập tin bằng thuật toán nâng cao		

2 Phân chia công việc

MSSV	Họ Tên	Công việc thực hiện
1712078	Ngô Phan Nhật Lâm	• Bài 1: Nén dữ liệu bằng thuật toán Huffman: ○ Tập tin chuỗi (.txt) ○ Tập tin khác (.cpp, .jpg,…) ○ Thư mục tập tin chuỗi ○ Thư mục nhiều loại tập tin
1712212	Lý Thiên Ân	• Bài 2: Nén dữ liệu bằng thuật toán nâng cao:

3 Mô tả các chức năng và cách thực hiện

3.1 Nén dữ liệu bằng thuật toán Huffman:

- Khởi tạo cây nhị phân Huffman để chứa các ký tự (node) trong chuỗi.

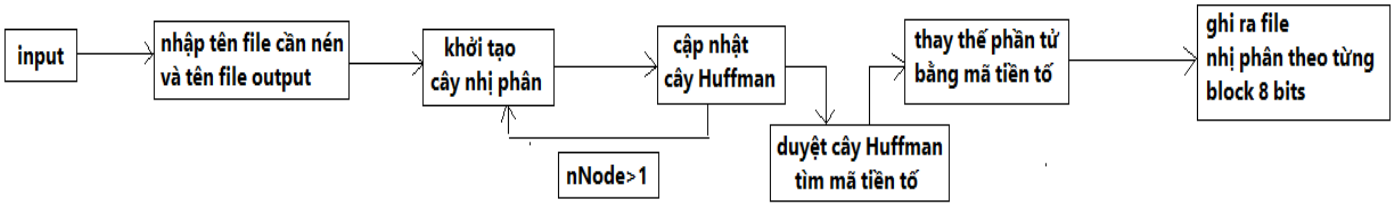
```
struct NODE {
    unsigned char   c;      // ky tu
    int     freq;           // so lan xuat hien
    bool    used;           // danh dau node da xu ly chua
    int     nLeft;          // chi so nut con nam ben trai
    int     nRight;         // chi so nut con nam ben phai
};

NODE    huffTree[MAX_NODE];
MABIT   bangMaBit[256];

void KhoiTao() {
    for (int i = 0; i < MAX_NODE; i++) {
        huffTree[i].c = i;
        huffTree[i].freq = 0;
        huffTree[i].used = false;
        huffTree[i].nLeft = -1;
        huffTree[i].nRight = -1;
    }
}
```

- Thống kê tần suất xuất hiện của các ký tự trong chuỗi. Tìm từng cặp phần tử xuất hiện ít nhất trong bảng thống kê. Xây dựng cây nhị phân dựa vào thuật toán Huffman. Lặp lại quá trình này cho đến khi còn 1 node chưa được xét thì dừng.

```
⊞void ThongKeTanSoXuatHien(char* tenFile){ ... }

⊞void XuatBangThongKe(){ ... }

⊞bool Tim2PhanTuMin(int &i, int &j, int nNode){ ... }

⊞int TaoCayHuffman(){ ... }
```

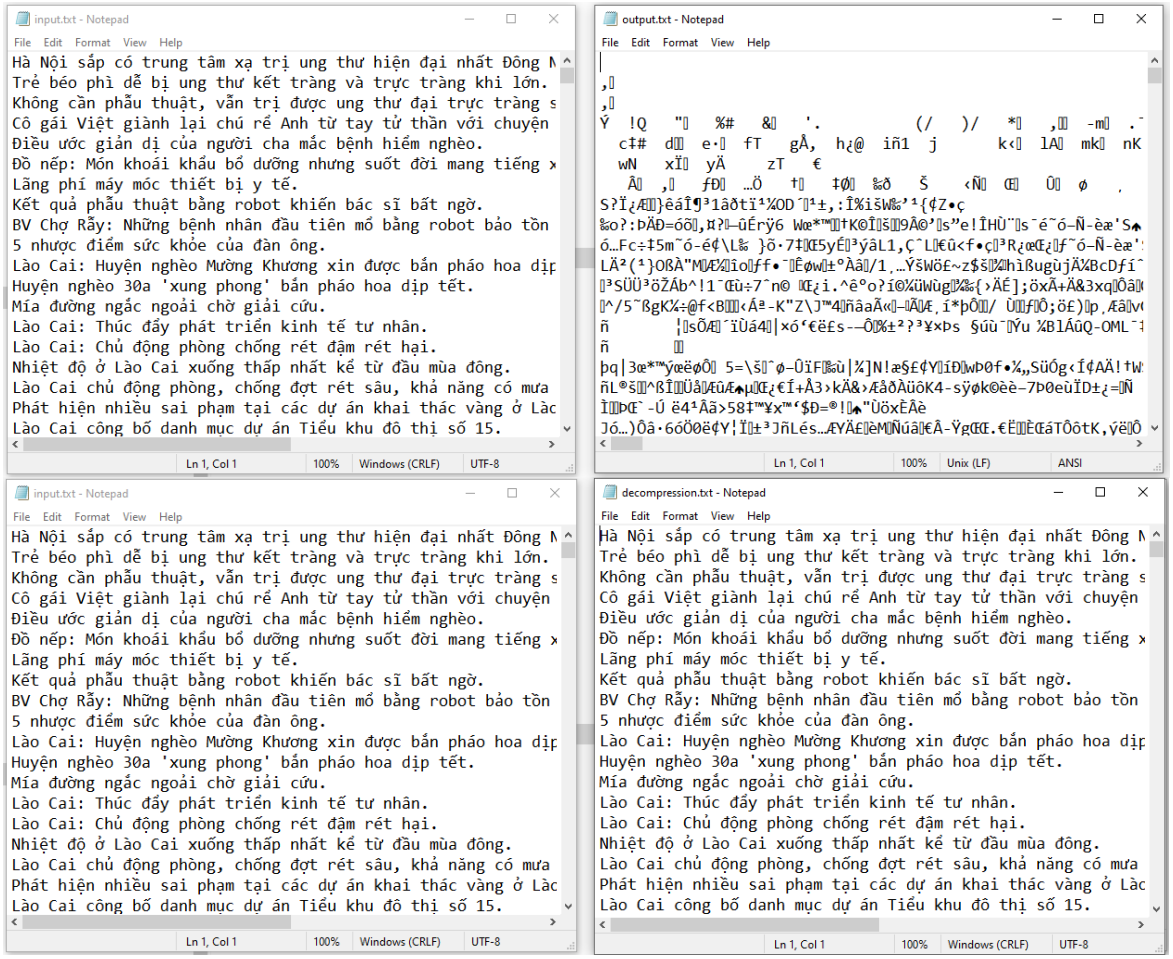
- Duyệt cây Huffman từ nốt gốc để tìm mã tiền tố cho mỗi ký tự theo quy ước rẽ trái gắn bit 0, rẽ phải gắn bit 1. Thay thế mỗi phần tử bằng mã tiền tố tương ứng cho đến hết chuỗi. Ghi ra file nhị phân từng block 8 bits.

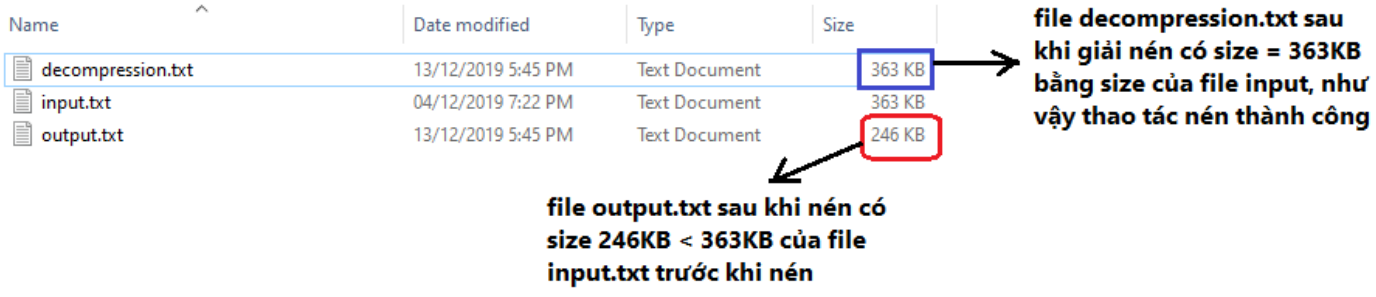
Lưu đồ thuật toán:



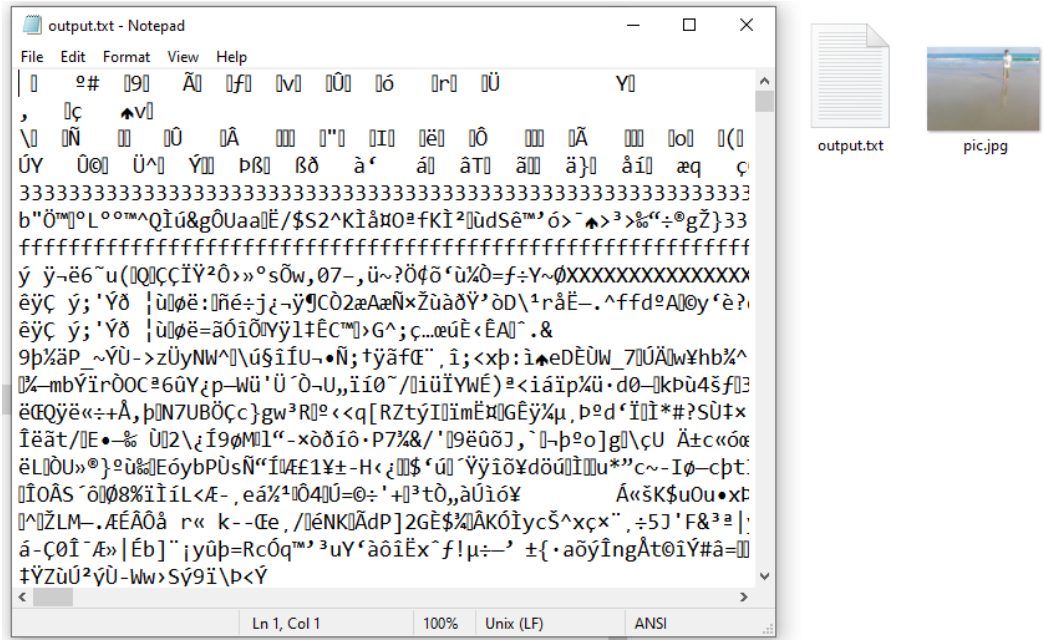
Chạy thử chương trình:

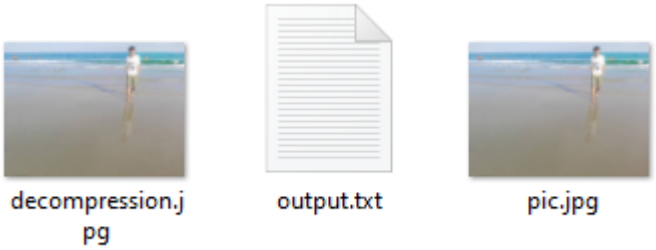
1. Nén file chuỗi (.txt)





2. Nén file ảnh (.jpg)





3. Nén file .cpp

Compression.cpp	07/12/2019 3:54 PM	C++ Source	9 KB
decompression.cpp	13/12/2019 6:00 PM	C++ Source	9 KB
output.txt	13/12/2019 6:00 PM	Text Document	6 KB

4. Test trên file dữ liệu mẫu

file output.txt đã được nén từ file corpus-title.txt có size nhỏ hơn file gốc

corpus-title.txt	12/02/2018 11:20 PM	Text Document	591,449 KB
output.txt	14/12/2019 11:44 AM	Text Document	398,749 KB

file decompression.txt được giải nén từ file output.txt có size bằng file gốc

corpus-title.txt	12/02/2018 11:20 PM	Text Document	591,449 KB
decompression.txt	14/12/2019 11:49 AM	Text Document	591,449 KB

3 Tham khảo & Video Demo

Video Demo:
https://www.youtube.com/watch?v=2vbQsF3a5PY

Link github:
https://github.com/nhatlamitus99/compression

Tài liệu tham khảo:
http://diendan.congdongcviet.com/threads/t2381::thuat-toan-nen-du-lieu-ly-thuyet-nen-du-lieu-huffman-phan-1-y-tuong.cpp