ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN

CẤU TRÚC DỮ LIỆU VÀ GIẢI THUẬT

Nguyễn Gia Minh – 19126054

Lê Thiên Kim – 19126022

Lê Hồng Long – 19126052

Phan Tường Vy –19126072

Thành phố Hồ Chí Minh

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN

| Đề tài |

Challenge 2

| Giáo viên hướng dẫn |

Thầy Văn Chí Nam

Thầy Bùi Huy Thông

Cô Phan Thị Phương Uyên

Thành phố Hồ Chí Minh

LỜI CẢM ƠN

Để có thể hoàn thành đồ án này, chúng em xin gửi lời cảm ơn chân thành tới thầy Văn Chí Nam, thầy Bùi Huy Thông và cô Phan Thị Phương Uyên khoa Công Nghệ Thông Tin, trường Đại học Khoa học Tự nhiên vì đã hỗ trợ tụi em trong suốt quãng thời gian làm đồ án, giúp đồ án này được hoàn thiện nhất có thể.

Do thời gian có hạn cũng như các kiến thức của chúng em vẫn có nhiều hạn chế, rất khó để tụi em tránh khỏi những sai sót khi hoàn thiện đồ án cũng như làm báo cáo. Vì vậy rất mong có thể nhận được những lời góp ý từ các thầy cô để chúng em có thể có thêm kiến thức để hoàn thành những đồ án tiếp theo.

Chúng em xin chân thành cảm ơn.

Mục lục:

BÁO CÁO

I. Thông tin chung:

a. Thông tin thành viên và phân công:

STT	Họ và Tên	MSSV	Nhiệm Vụ
1	Nguyễn Gia Minh	19126054	Báo cáo thuật toán Static Huffman và chỉnh sửa
2	Lê Thiên Kim	19126022	Xử lí Encoding và Decoding của thuật toán
3	Lê Hồng Long	19126052	Báo cáo thuật toán nén LZW
4	Phan Tường Vy	19126072	Xử lí file và tham số dòng lệnh

b. Tiến độ thực hiện:

Chức Năng	Đánh giá
Encode	100%
Decode	70%
Run in CMD	100%
Tham số dòng lệnh	100%

II. Thuật toán nén tĩnh Static Huffman:

a. Cơ sở dữ liệu

Huffman được biểu diễn bằng 1 cây nhị phân:

- Mỗi node lá chứa 1 kí tự.
- Node cha sẽ chứa các kí tự cũng những node con.
- Mỗi node được gán 1 trong 2 :
  + Node lá có chỉ số bằng số lần xuất hiện xuất hiện của kí tự trong file.
  + Node cha có chỉ số bằng tổng chỉ số xuất hiện kí tự của các node con.

b. Tạo cây Huffman:

1) Chọn trong bảng thống kê 2 phần tử x, y có chỉ số thấp nhất để tạo thành node cha z bằng tổng chỉ số của x, y (trong đó node có chỉ số nhỏ nằm bên trái, node có chỉ số lớn nằm bên phải).
2) Đưa x và y ra khỏi bảng.
3) Thêm node z vào bảng.
4) Lặp lại các bước trên cho đến khi còn lại 1 node suy nhất trong bảng.

- **Lưu ý:** Nếu chỉ số bằng nhau, node có kí tự nhỏ nằm bên nhánh trái, node có kí tự lớn nằm bên phải, ưu tiên xử lý các node có kí tự ASCII nhỏ trước.

c. Tính chất cây Huffman:

- Nhánh trái tương ứng với mã hoá bit '0'; nhánh phải tương.
  ứng với mã hoá bit '1'.
- Các nút có tần số thấp nằm ở xa gốc ◊ mã bit dài.
- Các nút có tần số cao nằm ở gần gốc ◊ mã bit ngắn.
- Số nút của cây: **(2n-1)**

d. Biểu diễn mã bit cho các kí tự:

1) Mã của mỗi kí tự được tạo bằng cách duyệt từ nút gốc đến nút lá chưa kí tự đó.
2) Khi duyệt sáng trái, tạo ra 1 bit 0.
3) Khi duyệt sang phải, tạo ra 1 bit 1.

e. Encode Phase:

1) Duyệt file, lập bảng thống kê tuần suất xuất hiện của mỗi ký tự.
2) Xây dựng cây Huffman dựa vào bảng thống kê.
3) Sinh mã Huffman cho mỗi ký tự dựa vào cây Huffman.
4) Duyệt file, thay toàn bộ ký tự bằng mã Huffman tương ứng.
5) Lưu lại cây Huffman (bảng mã) dùng cho việc giải nén. Xuất file đã nén.

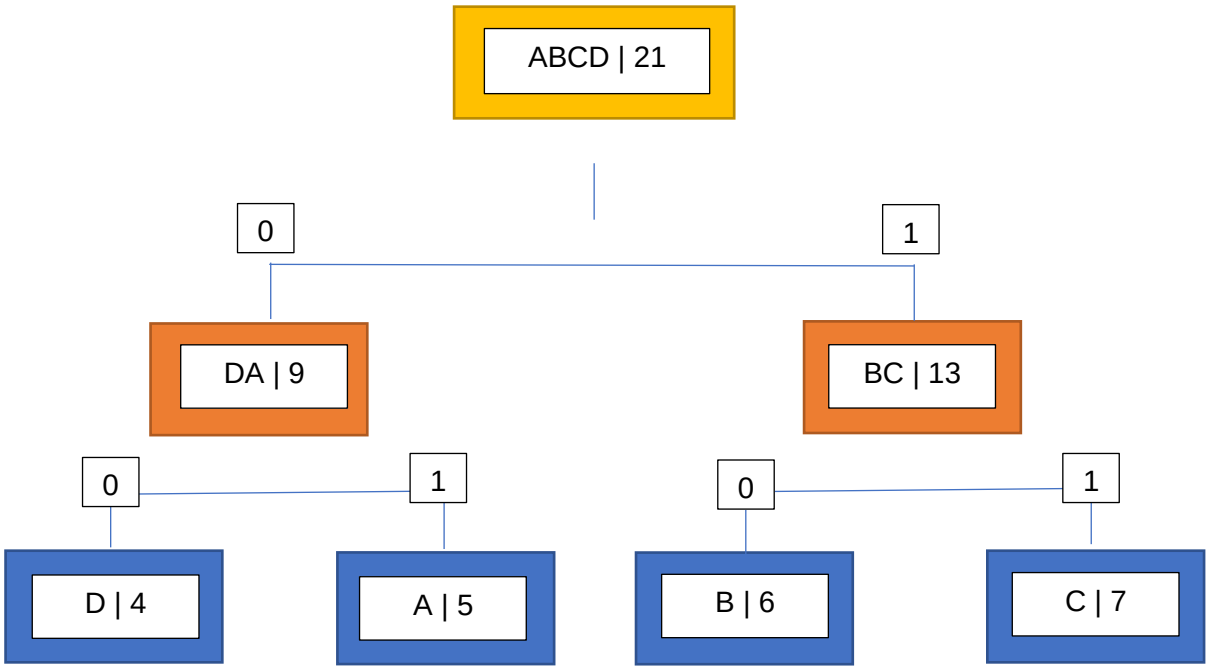
Ví dụ nén dãy: ABCDDCBAAABCDCBABBCDCC

Kí tự	Tần Suất
D	4
A	5
B	6
C	7

Kí tự	Tần Suất
DA	9
B	6
C	7

Kí tự	Tần Suất
DA	9
BC	13

Kí tự	Tần Suất
ABCD	21



Ký Tự	Mã Huffman
D	00
A	01
B	10
C	11

Kết quả: 01101100001110010101101100111001101011001111

Đầu vào kích thước: 22 * 8bits = 176 bits

Đầu ra kích thước: (2 * 4 + 2 * 5 + 2 * 6 + 2 * 7) = 44 bits

Tiết kiệm: 176 – 44 = 132 bits

Tỉ lệ nén: (1 - 44/176) * 100 = 75%

f. Decode Phase:

1) Xây dựng lại cây Huffman từ thông tin giải mã đã lưu
2) Duyệt file nén:
    a. Khởi tạo node hiện hành.

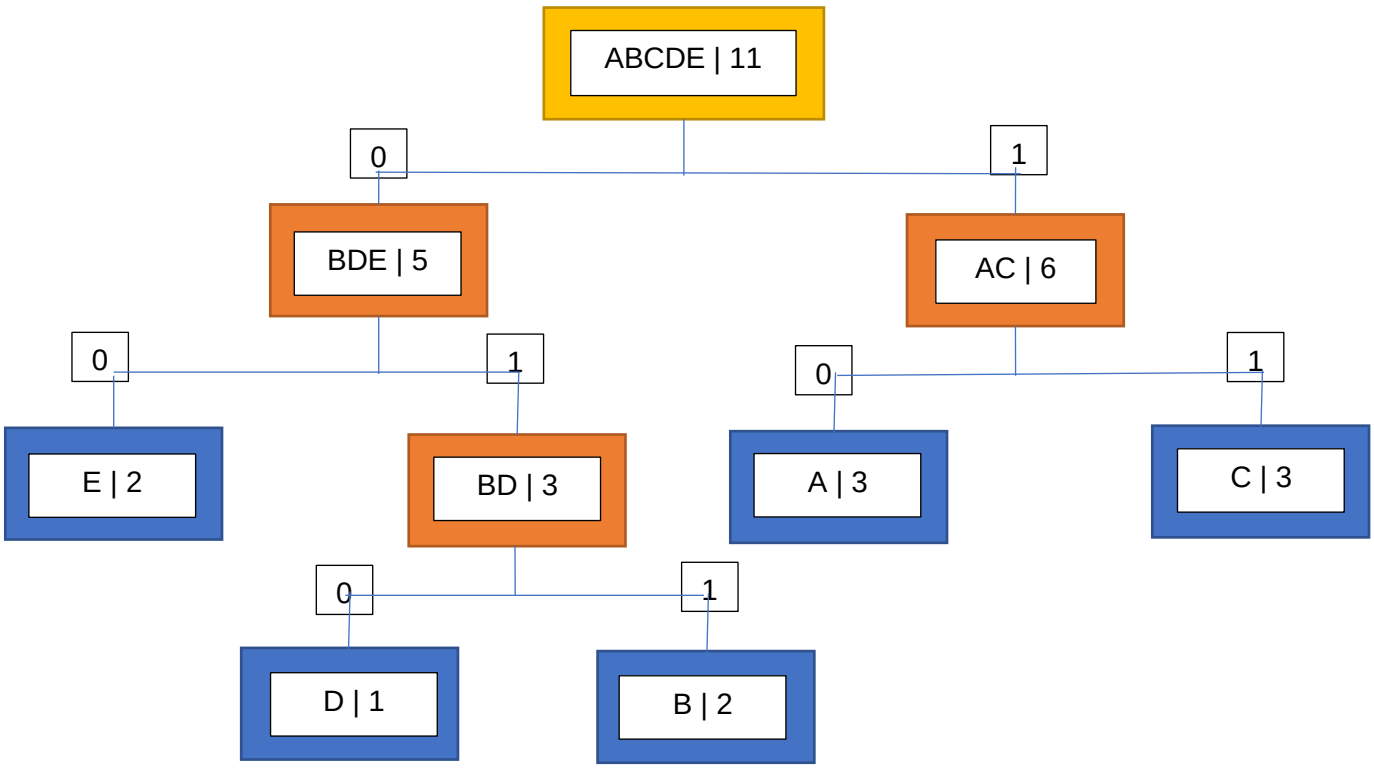
    b. Đọc lần lượt từng bit trong file.

    c. Duyệt cây: Nếu bit = 0 thì xét bên trái của Node hiện hành còn bit = 1 thì xét bên phải của Node hiện hành.
3) Xuất ký tự tương ứng cho đến khi duyệt đến nút lá.
4) Thực hiện bước 2, bước 3 cho đến duyệt hết file.
5) Dừng thuật toán. Xuất file đã giải nén.

Ví dụ giải nén: 101100011010011 1100100011

Kí tự	Mã bit	Chỉ số xuất hiện
A	11	3
B	10	2
C	00	3
D	011	1
E	010	2

- **Cây Huffman**:



- **Duyệt từng bit trong dãy của file nén:**

Mức	1	0	1	1	0	0	0	1	1	0	1	0	0	1	1	1	1	0	0	1	0	1	1	1	0
1	X		X		X		X			X			X			X		X		X		X		X		
2		X		X		X		X			X			X			X		X		X		X		X
3									X			X			X											
Kết quả		A		C		E			B			D			B		C		E		A		C		A

- Kết quả: ACEBDBCEACA

g. Ưu – Nhược Điểm:

Ưu điểm:

- Hệ số nén tương đối cao.
- Phương pháp thực hiện khá đơn giản.
- Tốn ít bộ nhớ để lưu trữ.

Nhược điểm:

- Mất 2 lần duyệt file khi nén.
- Phải lưu trữ thông tin giải mã vào file nén.
- Phải xây dựng lại cây huffman khi giải nén.

h. Ứng dụng thực tế:

- Sử dụng trong tất cả các định dạng nén thường gặp như: GZIP, PKZIP, WINZIP, …
- Hầu hết các tệp hình ảnh (JPEG, MPEG) hay tệp âm thanh (MP3).
- Dùng trong mã hóa chuỗi HPACK (kỹ thuật nén tiêu đề của http/2).
- Là "back-end" cho một số phương pháp nén như DEFLATE (thuật toán của PKZIP).

III. Phương Pháp Nén LZW:

a. Cơ sở dữ liệu:

- Thuật toán LZW hoạt động theo nguyên lý tạo ra một dãy mã
- Mã từ 0 – 255 miêu tả 1 dãy ký tự thay thế cho ký tự 8 bit
- Mã từ 256 – 4095 được tạo bên trong 1 từ điển cho trường hợp lặp chuỗi
- Mỗi bước nén, byte nhập vào được tập hợp lại thành 1 chuỗi

b. Encode Phase:

1) Khởi tạo "Từ điển" chứa tất cả chuỗi có 1 kí tự.
2) Tìm chuỗi W dài nhất trong từ điểm đối chiếu với dữ liệu nhập hiện tại.
3) Xuất vị trí từ điển cho W ra file output và xóa W khỏi dữ liệu nhập.
4) Thêm W và ký tự tiếp theo trong dữ liệu nhập vào "Từ điển".
5) Quay lại bước 2.

Ví dụ mã hóa chuỗi "ABCBCABCABCD"

1) Cho T = NULL;
2) Đọc kí tự thứ K trong chuỗi
3) Nếu TK tồn tại trong "Từ điển", mã hóa ra T = K
4) K = K + 1

Count	T	K	TK	Ký tự	Index	Output
0	NULL	A	A			
1	A	B	AB	AB	258	65
2	B	C	BC	BC	259	66
3	C	B	CB	CB	260	67
4	B	C	BC			
5	BC	A	BCA	BCA	261	259
6	A	B	AB			
7	AB	C	ABC	ABC	262	258
8	C	A	CA	CA	264	67
9	A	B	AB			
10	AB	C	ABC			
11	ABC	D	ABCD	ABCD	264	262
12	D	NULL	D			68

Đầu vào kích thước: 12 * 8bits = 96 bits

Đầu ra kích thước: 5 * 8 + 3 * 9 = 67 bits

Tiết kiệm: 96 – 67 = 29 bits

Tỉ lệ nén: (1 - 67/96) * 100 = 30,2%

c. Decode Phase:

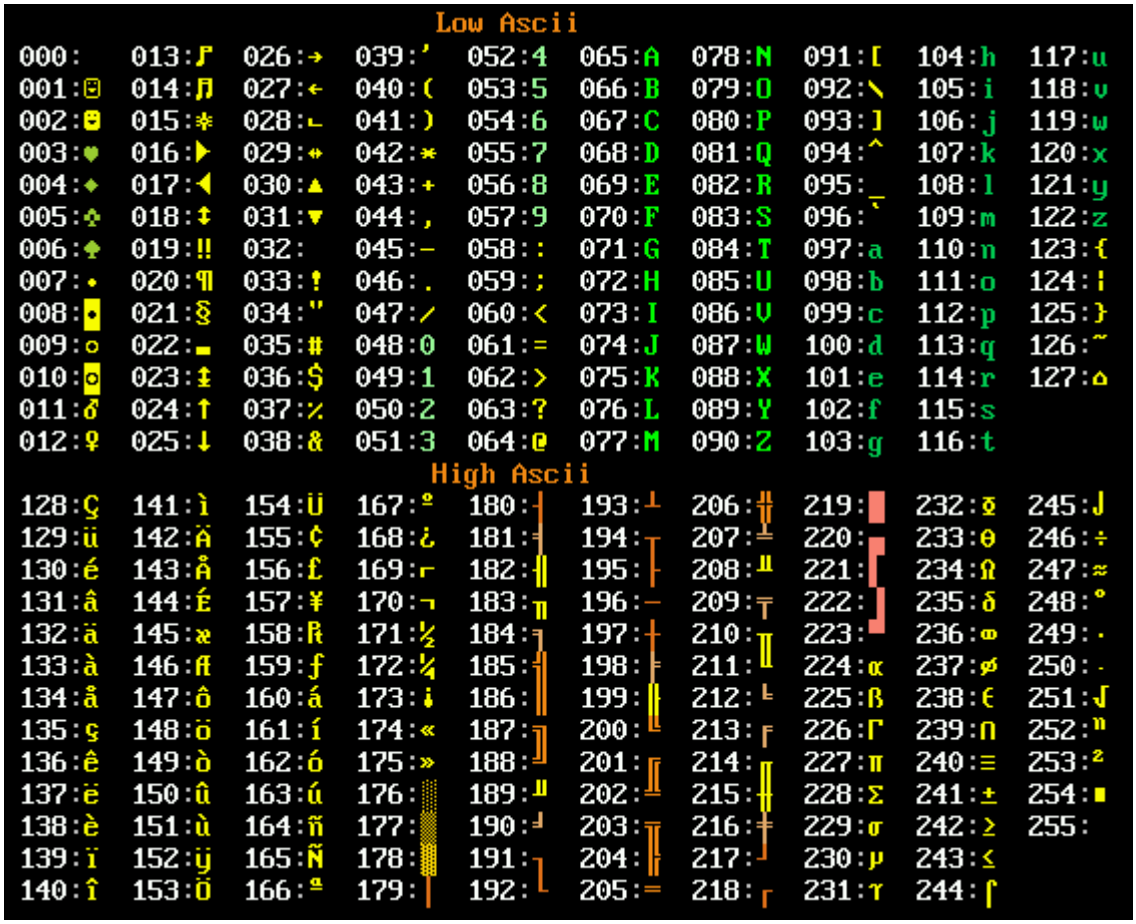
1) Đọc giá trị từ dữ liệu nhập đã mã hóa và xuất ra chuỗi tương ứng từ " Từ điển" đã được khởi tạo.

2) Tại cùng 1 thời điểm nó thu được giá trị tiếp theo từ dữ liệu nhập, thêm vào từ điển xích chuỗi của chuỗi xuất. Và kí tự đầu tiên của chuỗi nhận được khi mã hóa kí tự tiếp theo.

3) Trình giải nén xử lý giá trị nhập tiếp theo, quá trình đó lặp lại đến khi dữ liệu nhập không còn, tại thời điểm giá trị nhập cuối cùng được mã hóa không còn bất kì giá trị nào thêm vào từ điển.

Ví dụ giải nén: 75 73 77 76 79 78 71 86 129 73 78 72

- Từ điển:



Duyệt từng dữ liệu

+ Bước 1: 75 tương ứng với chữ K

+ Bước 2: 73 tương ứng với chữ I, thêm KI vào từ điển vị trí 256

+ Bước 3: 77 tương ứng với chữ M, thêm IM vào từ điển vị trí 257

+ Bước 4: 76 tương ứng với chữ L, thêm ML vào từ điển vị trí 258

+ Bước 5: 79 tương ứng với chữ O, thêm LO vào từ điển vị trí 259

+ Bước 6: 78 tương ứng với chữ N, thêm ON vào từ điển vị trí 260

+ Bước 7: 71 tương ứng với chữ G, thêm NG vào từ điển vị trí 261

+ Bước 8: 86 tương ứng với chữ V, thêm GV vào từ điển vị trí 267

+ Bước 9: 129 tương ứng với chữ IM, thêm VIM vào từ điển vị trí 268

+ Bước 10: 73 tương ứng với chữ I, thêm IMI vào từ điển vị trí 269

+ Bước 10: 78 tương ứng với chữ N, thêm IN vào từ điển vị trí 270

+ Bước 11: 72 tương ứng với chữ H, thêm NH vào từ điển vị trí 271

+ Bước 12: KI được nối chuỗi trùng với IM, thêm KIM vào từ điển vị trí 272

…

+ Bước 57: KIMLONGVIMIN được nối chuỗi trùng với IMLONGVIMINH, thêm KIMLONGVIMINH vào từ điển vị trí 317

+ Bước 58: Không còn giá trị nào được thêm vào từ điển, trả về kết quả cuối cùng

- Kết quả: KIMLONGVIMINH

d. Ưu – Nhược Điểm:

Ưu điểm:

- Hệ số nén tương đối cao.
- Từ lưu trữ trong từ điển xuất hiện thường xuyên trong văn bản thì mức độ nén cao hơn.
- Trong tập tin nén không cần phải chứa bảng mã.
- Decode có thể tự xây dựng bảng mã không cần bên nén gửi bảng mã
- Hạn chế lãng phí bộ nhớ
- Khắc phục được sự cứng nhắc của thuật toán nén, góp phần giúp thuật toán mềm dẻo hơn.

Nhược điểm:

- Tốn nhiều bộ nhớ
- Khó thực hiện với mảng đơn giản (nhỏ hơn 64KB)

e. Ứng dụng thực tế:

- Dùng để nén dạng ảnh PNG không làm mất chất lượng ảnh, được hỗ trợ trong suốt nên nó là định dạng tuyệt vời cho đồ họa Internet.
- Có thể nén định dạng GIF, thuật toán nén GIF xây dựng một bảng màu, mỗi màu được kết hợp với 1 pixel nên hình ảnh có vùng màu càng lớn file sẽ được nén lại càng nhỏ.

IV. Tài liệu tham khảo:

- https://l.facebook.com/l.php?u=https%3A%2F%2Fwww.dcode.fr%2Flzw-compression%3Ffbclid%3DIwAR3MGdniy_E37r445Ox98PZztxza97E6A2BsBiTQbfEceUf2dITED3WxaaA%230&h=AT3Sf8uCSgsQO3jNvJN51jrsvtwI4NFfVErEde3JzdV4Vo1XghzIq7cst9r_rjxf

T4s39qugIhL3V9NvHMi1lWrB9MfqrmBYwrSV7YLhqTvQWj9BgoaUcKWIzeun6CEzULM8DSyj1M90IZ58u3XO4Q

- https://l.facebook.com/l.php?u=https%3A%2F%2Fwebhome.cs.uvic.ca%2F~nigelh%2FPublications%2FimprovingLZW.pdf%3Ffbclid%3DIwAR0Avf-W0rHEhOFLux3suF7cScJfS3g-Fbpl3lNBZJrh-tsmgjvMVTp24xM&h=AT3Sf8uCSgsQO3jNvJN51jrsvtwI4NFfVErEde3JzdV4Vo1XghzIq7cst9r_rjxfT4s39qugIhL3V9NvHMi1lWrB9MfqrmBYwrSV7YLhqTvQWj9BgoaUcKWIzeun6CEzULM8DSyj1M90IZ58u3XO4Q
- https://l.facebook.com/l.php?u=https%3A%2F%2Flib.hpu.edu.vn%2Fbitstream%2Fhandle%2F123456789%2F18056%2F1_TrinhThiThuHa_CT901.pdf%3Ffbclid%3DIwAR0CACaE0vSzKOOpo63nbsF_vtvNq5iZG5o7qARagvm-ZVp165tWo_R-Fiw&h=AT3Sf8uCSgsQO3jNvJN51jrsvtwI4NFfVErEde3JzdV4Vo1XghzIq7cst9r_rjxfT4s39qugIhL3V9NvHMi1lWrB9MfqrmBYwrSV7YLhqTvQWj9BgoaUcKWIzeun6CEzULM8DSyj1M90IZ58u3XO4Q
- https://l.facebook.com/l.php?u=http%3A%2F%2Fdulieu.tailieuhoctap.vn%2Fbooks%2Fcong-nghe-thong-tin%2Fthe-loai-khac%2Ffile_goc_768102.pdf%3Ffbclid%3DIwAR2XTOOPqUHP4YDfqEFtfpHIXzADZZj6RFRRv9NVhqtEmMb3HXuO1vucun8&h=AT3Sf8uCSgsQO3jNvJN51jrsvtwI4NFfVErEde3JzdV4Vo1XghzIq7cst9r_rjxfT4s39qugIhL3V9NvHMi1lWrB9MfqrmBYwrSV7YLhqTvQWj9BgoaUcKWIzeun6CEzULM8DSyj1M90IZ58u3XO4Q
- https://l.facebook.com/l.php?u=https%3A%2F%2Fwww.it-swarm-vi.tech%2Fvi%2Falgorithm%2Fcac-ung-dung-trong-gioi-thuc-cua-ma-hoa-huffman-la-gi%2F968426967%2F%3Ffbclid%3DIwAR0ySBqn-fwdk5kEA4zf3wk5_gNAtDlDooQXY9UlZPJJ7cAj2-Ng4m4wlLk&h=AT3Sf8uCSgsQO3jNvJN51jrsvtwI4NFfVErEde3JzdV4Vo1XghzIq7cst9r_rjxfT4s39qugIhL3V9NvHMi1lWrB9MfqrmBYwrSV7YLhqTvQWj9BgoaUcKWIzeun6CEzULM8DSyj1M90IZ58u3XO4Q
- http://vi.uwenku.com/question/p-ggeiczqn-r.html
- http://luanvan.co/luan-van/ma-hoa-lzw-lempel-ziv-wech-30500/
- https://www.youtube.com/watch?v=PlNVtQg-FWg
-