ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
UIT-HCM

THUẬT TOÁN HUFFMAN VÀ SHANNON-FANO

Thành viên: Lê Đặng Đăng Huy 19521612
Tô Thanh Hiền 19521490
Trần Vĩ Hào 19521482

Nội dung

I. Giới thiệu
II. Ý tưởng chính của thuật toán
III. Thực nghiệm
IV. Phân công công việc

I. Giới thiệu

Huffman Vs Shannon-Fano

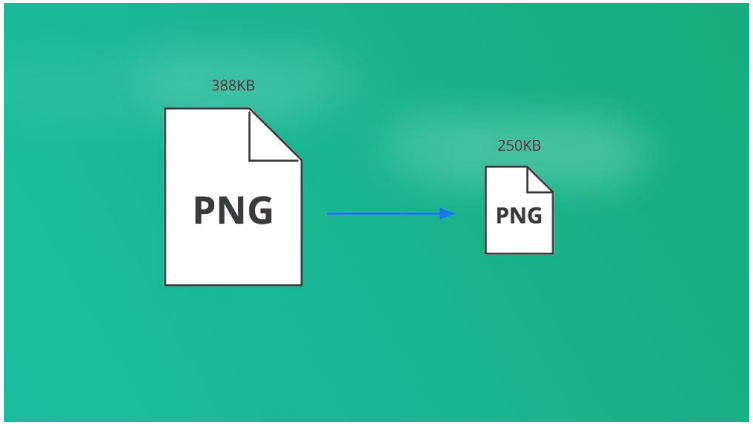
I. Giới thiệu

Nén tệp văn bản

- **Input:** Một tập tin văn bản có định dạng là *.txt.
- **Output:** Một file *.txt chứa thông tin đã mã hóa của tệp văn bản đầu vào với kích thước file nhỏ hơn.

- **Input:** Một ảnh màu có định dạng là *.jpg.
- **Output:** Một file *.txt chứa thông tin đã mã hóa của ảnh đầu vào với kích thước file nhỏ hơn.

Nén hình ảnh



II. Ý tưởng chính

1. Huffman

- Dựa trên bảng tần suất xuất hiện các kí tự cần mã hóa để xây dựng một bộ mã nhị phân cho các kí tự đó sao cho dung lượng (số bit) sau khi mã hóa là nhỏ nhất.
- Giảm số bit biểu diễn 1 ký tự.
- Dùng chuỗi bit ngắn hơn để biểu diễn ký tự xuất hiện nhiều lần.
- Sử dụng mã tiền tố để phân cách các ký tự.

II. Ý tưởng chính

2. Shannon

- Xây dựng mã tiền tố dựa trên một tập hợp các ký hiệu và xác suất của các ký tự.
- Chọn mã tiền tố trong đó biểu tượng nguồn được cung cấp độ dài từ mã

$$l_i=[-\log_2 p_i].$$

3. Fano

- Xây dựng mã tiền tố dựa trên một tập hợp các ký hiệu và xác suất của các ký tự.
- Chia các ký hiệu nguồn thành hai bộ (0 và 1) với xác suất càng gần ½ càng tốt. Sau đó, các tập hợp đó tự được chia làm hai, và cứ tiếp tục như vậy, cho đến khi mỗi tập hợp chỉ chứa một ký hiệu. Từ mã cho biểu tượng đó là chuỗi 0 và 1 ghi lại nửa số chia mà nó nằm trên.

III. Thực nghiệm

1. Dữ liệu

Nén tệp văn bản

File	Kích thước (bytes)	Chủ đề
1.txt	715.332	Đoạn văn tiếng Anh.
2.txt	1.115.033	Dataset bán hàng trên Kaggle.
3.txt	251.705	Những bài văn tiếng Việt.
4.txt	78.674	Dữ liệu crawl được từ bài tập Lab1 và Lab2 của môn học này.
5.txt	2.167.737	Văn bản tiếng Anh.

Nén hình ảnh

File	Kích thước gốc (bytes)	Kích thước sau khi chuyển đổi thành mảng (bits)	Size
1.jpg	480.273	68.836.691	1200x600
2.jpg	114.644	59.081.540	549x960
3.jpg	2.137.052	213.530.513	1920x1080
4.jpg	431.489	227.492.125	1920x1080
5.jpg	120.999	59.284.232	691x777

III. Thực nghiệm

2. Kết quả

Nén tệp văn bản

File		Kích thước sau khi nén (bytes)	Tỷ số nén	Thời gian nén (giây)	Thời gian giải nén (giây)
1.txt	H	394.017	1,81549	0,38568	0,85217
	S-F	446.830 – 400.490	1,6009 – 1,78614	0,38883 – 0,39092	1,35034 – 1,22663
2.txt	H	558.512	1,99644	0,55016	1,19519
	S-F	610.344 – 571.243	1,82689 – 1,95194	0,56058 – 0,55436	1,81574 – 1,75459
3.txt	H	118.259	2,12842	0,12282	0,2659
	S-F	132.112 – 119.654	1,90524 – 2,10361	0,13127 – 0,12323	0,43577 – 0,37388
4.txt	H	52.576	1,49639	0,04489	0,10977
	S-F	56.792 – 53.839	1,3853 – 1,46128	0,05273 – 0,05597	0,19472 – 0,1689
5.txt	H	1.145.521	1,89236	1,15352	2,38875
	S-F	1.232.335 – 1.180.838	1,75905 – 1,83576	1,12055 – 1,1209	3,69857 – 3,62304

III. Thực nghiệm

2. Kết quả

Nén ảnh

File		Kích thước sau khi nén (bits)	Tỷ số nén	Thời gian nén (giây)	Thời gian giải nén (giây)
1.jpg	H	32.946.601	2,08934	21,11621	87,65538
	S-F	43.480.367 – 35.589.154	1,58317 – 1,9342	22,05799 – 23,11391	121,25229 – 94,00442
2.jpg	H	29.634.697	1,99366	18,30662	80,19416
	S-F	41.422.532 – 31.131.113	1,42631 – 1,89783	20,11017 – 19,55467	107,30088 – 82,86994
3.jpg	H	106.141.666	2,01175	74,42264	301,52266
	S-F	148.499.611 – 109.822.600	1,43792 – 1,94432	76,21462 – 78,42185	388,5325 – 290,18915
4.jpg	H	113.470.111	2,00486	79,82168	322,04064
	S-F	158.240.326 – 117.531.798	1,43764 – 1,93558	75,68447 – 79,59545	395,44557 – 383,99727
5.jpg	H	29.722.538	1,99459	18,04945	82,94047
	S-F	42.015.277 – 30.688.090	1,41102 – 1,93183	21,25767 – 19,73568	104,35448 – 77,21478

III. Thực nghiệm

3. Kết luận

- Từ kết quả nhận được thông qua quá trình thực nghiệm trên cả hai tác vụ nén tệp văn bản và nén ảnh, ta có thể nhận thấy Huffman tỏ ra vượt trội so với Shannon-Fano trên bộ dữ liệu mà chúng tôi đã chuẩn bị.
- Huffman không chỉ có tỷ số nén cao hơn mà còn có thời gian nén lẫn thời gian giải nén ngắn hơn Shannon-Fano.
- Ở tác vụ nén tệp văn bản, ta có thể thấy sự khác biệt nhưng vẫn chưa đủ rõ ràng. Cho đến khi tác vụ nén ảnh được đưa vào thực nghiệm thì Shannon thực sự thua kém Huffman một cách rõ rệt và chỉ còn Fano là đủ sức cạnh tranh với Huffman.
- Ngoài ra, trong quá trình thực nghiệm, phương pháp Fano cho kết quả tốt hơn phương pháp Shannon. Trên cùng một tệp văn bản hay ảnh, qua mỗi lần nén thì thuật toán Shannon-Fano lại cho ra một tỷ số nén khác nhau (trong khi Huffman thì không gặp hiện tượng này), điều đó nói lên việc Shannon-Fano chưa thực sự cho ra mã tối ưu và cố định.

III. Thực nghiệm

4. So sánh

Huffman	Shannon-Fano
Hiệu quả và tối ưu.	Kết quả tạo ra không tối ưu.
Độ dài từ dự kiến $H(X) + 1 - p_{min}$.	Độ dài từ dự kiến $H(X) + 1$.
Xây dựng cây nhị phân Bottom up.	Xây dựng cây nhị phân Top down.

IV. Phân công công việc

Họ và tên	MSSV	Công việc	Mức độ hoàn thành
Trần Vĩ Hào	19521482	Huffman + Báo cáo	100%
Tô Thanh Hiền	19521490	Fano + Code	100%
Lê Đặng Đăng Huy	19521612	Shannon + Slide	100%

Cảm ơn!