

CHALLENGE 3: NÉN DỮ LIỆU

1 Nội dung

1.1 Tìm hiểu Nén dữ liệu là khoa học biểu diễn thông tin dưới dạng rút gọn. Nén dữ liệu đã và đang là một trong những công nghệ hỗ trợ quan trọng cho cuộc cách mạng kỹ thuật số diễn ra trong nhiều thập kỷ. Challenge này đề cập đến hai thuật toán nén kinh điển: **Nén Huffman tĩnh (Static Huffman Encoding)** và **Lempel-Ziv-Welch(LZW)**. Tham khảo [tài liệu sau](#) và hoàn thành các yêu cầu dưới đây

1. Định nghĩa các thuật toán **Huffman tĩnh** và **LZW**, cần nêu rõ:
 - Pha nén và giải nén của từng thuật toán.
 - Mỗi pha cần biểu diễn theo từng bước, không trình bày mã nguồn.
 - Các cấu trúc dữ liệu có liên quan cần được thể hiện, đồng thời nêu rõ ảnh hưởng của các cấu trúc dữ liệu này lên các thuật toán nén.
 - Khuyến khích việc sử dụng ví dụ minh họa cho các yêu cầu ở trên.
2. Chỉ ra các điểm mạnh và điểm bất lợi của từng thuật toán nêu trên.
3. Chỉ ra một số ứng dụng thực tế và các phiên bản cải tiến của 2 thuật toán được nêu.

1.2 Lập trình

- Sinh viên được yêu cầu cài đặt một chương trình 2 chiều: Nén và Giải nén file văn bản dùng thuật toán nén **LZW**.
- Chương trình cần được biên dịch thành file thực thi có định dạng `".exe"` nhận các tham số dòng lệnh sau:

`A.exe Action InputPath OutputPath OutputInfo`

,trong đó:

- * **A.exe**: File thực thi ".exe".
- * **Action**:
 - **-e**: Pha nén. Nhận vào file input có định dạng (.txt) and và trả về file output nhị phân có định dạng (.lzw)
 - **-d**: Pha giải nén. Nhận vào file input nhị phân có định dạng (.lzw) và trả về file text có định dạng (.txt).
- * **InputPath**: Đường dẫn đến file input.
- * **OutputPath**: Đường dẫn đến file output.
- * **OutputInfo**: Các thông tin về quá trình thực thi, cần được biểu diễn trên màn hình console.
 - **-i**: Kích thước của file input, kích thước của file output, tỉ lệ (%) không gian bộ nhớ tiết kiệm được.
 - **-d**: Từ điển LZW được tạo ra trong quá trình nén/giải nén.
 - **-ind**: Cả 2 thông tin trên.

Ví dụ:

Text file	Compressed file	-i	-d
Text file: "Data.txt"	Binary file: "Compressed.lzw"	Input size: bits Output size: bits Space saved: %	new word 1: 256 new word 2: 257 ...
WYS*WYGWYS*WYSWYSG	87 89 83 42 256 71 256 258 262 262 71	Input size: 144 bits Output size: 110 bits Space saved: 23.6%	WY: 256 YS: 257 S*: 258 *W: 259 WYG: 260 GW: 261 WYS: 262 S*W: 263 WYSW: 264 WYSG: 265

2 Yêu cầu và Quy định nộp bài

2.1 Yêu cầu

- Phần Challenge là phần bài tập không bắt buộc. Điểm số của bài Challenge này sẽ được cộng thêm vào điểm tổng kết môn học.
- Sẽ có 3 bài Challenge xuyên suốt quá trình học. Sinh viên làm các bài Challenge này theo nhóm 3 người.
- Một bài Challenge hợp lệ (Được chấm và tính điểm cộng) cần thỏa mãn các điều kiện sau:
 - Nằm trong số 13 bài nộp đầu tiên (tính theo mốc thời gian của lần nộp bài cuối cùng). Link nộp bài sẽ tự động đóng khi đã đủ nhóm nộp hoặc hết thời hạn nộp bài.
 - Bài làm cần đảm bảo thể hiện đầy đủ các yêu cầu được giao. Không chấp nhận các bài làm thiếu sót một hoặc nhiều phần dưới bất kì hình thức nào.
 - Trình bày và thực hiện đầy đủ các quy định ở phần **Quy định nộp bài** bên dưới.
 - Đối với mỗi Challenge, sinh viên cần lập một nhóm 3 người mới, nghĩa là, 2 sinh viên bất kì chỉ được chung nhóm tối đa 1 Challenge.
- Giảng viên lý thuyết có quyền yêu cầu một trong các nhóm có bài nộp Challenge hợp lệ trình bày trước lớp trong buổi học lý thuyết bất kì.

2.2 Quy định nộp bài

- Mã nhóm là chuỗi kết hợp của 3 số cuối của mã số sinh viên của các thành viên trong nhóm, ngăn cách bởi dấu "_". Thành viên khóa lớn hơn đứng trước, nếu cùng khóa thì MSSV lớn hơn đứng trước. Ví dụ:
 - 1812999, 20120995, 20120994 \rightarrow 999_995_994
 - 20120999, 20120995, 20120994 \rightarrow 994_995_999

- Nhóm sinh viên nộp bài trên Moodle dưới dạng [**Mã_nhóm.rar**] hoặc [**Mã_nhóm.zip**], là file nén của thư mục [**Mã_nhóm**]. Thư mục này bao gồm:
 - File báo cáo được trình bày dưới dạng văn bản [**Report.pdf**] hoặc dạng slide [**Report.pptx**]. File này chứa nội dung tìm hiểu ở phần **1.1** và trình bày hướng giải quyết vấn đề lập trình ở phần **1.2**.
 - * Nếu bài nộp file báo cáo là dạng slide, cần giải thích cặn kẽ từng slide ở phần *Note* của slide đó.
 - * Trang đầu (hoặc slide đầu) của file báo cáo cần ghi rõ thông tin (Họ và tên, MSSV) của từng thành viên trong nhóm.
 - * File báo cáo cần trình bày có cấu trúc, logic, rõ ràng, mạch lạc, trả lời thẳng vào trọng tâm của câu hỏi. Độ dài của bài nộp không vượt quá 15 trang đối với file văn bản, và 30 trang đối với slide.
 - Thư mục lập trình **Source**. Thư mục này chỉ chứa các file `".cpp"` và `".h"`, nếu sinh viên có nộp kèm file có định dạng khác, cần giải thích rõ ràng công dụng của file này trong báo cáo. Yêu cầu mã nguồn cần trình bày logic, có chú thích rõ ràng.