NT212 – AN TOÀN DỮ LIỆU VÀ KHÔI PHỤC THÔNG TIN SAU SỰ CỐ

BÁO CÁO THỰC HÀNH BÀI TẬP

Assignment 3.2

***

THÔNG TIN CHUNG

1. Tên nhóm

STT	Họ và tên	MSSV	Email
1	Dương Phạm Huy Thông	22521431	22521431@gm.uit.edu.vn
2	Cao Quí	22521208	22521208@gm.uit.edu.vn

---

2. Nội dung thực hiện

STT	Nội dung	Tự đánh giá	Phụ trách
1	Khôi phục tập tin ảnh	100%	Huy Thông
2	Khôi phục tập tin pdf *(mở rộng)*	100%	Cao Quí

Bên dưới đây là toàn bộ bài báo cáo chi tiết đã được nhóm thực hiện.

BÀI LÀM

I. KHÔI PHỤC TẬP TIN ẢNH

1. Ý tưởng thực hiện

- **Hệ thống tệp FAT32 (File Allocation Table 32)** là một hệ thống tệp được Microsoft phát triển từ năm 1977, được sử dụng rộng rãi nhờ tính tương thích cao với nhiều thiết bị.

- Khi một đĩa ảo được format, bảng phân bổ tệp (FAT) bị ghi đè, nhưng dữ liệu thực tế vẫn có thể còn nguyên cho đến khi bị ghi đè bởi dữ liệu mới. Do đó, việc phục hồi dựa trên signature của tệp là phương pháp khả thi.

- **Phương pháp phục hồi dữ liệu:** Khi đĩa ảo bị format, cấu trúc hệ thống tệp FAT32 bị xóa, nhưng dữ liệu thực tế vẫn có thể tồn tại. Phương pháp phục hồi dựa trên signature (file signature recovery) được sử dụng để tìm và trích xuất các tệp JPG và PNG.

- **Signature của JPG và PNG**:

+ JPG: Bắt đầu bằng ***\xff\xd8*** (2 byte) và kết thúc bằng ***\xff\xd9*** (2 byte).

+ PNG: Bắt đầu bằng ***\x89\x50\x4E\x47\x0D\x0A\x1A\x0A*** (8 byte) và kết thúc bằng chunk IEND với cấu trúc ***\x00\x00\x00\x00IEND*** (8 byte) theo sau là ***4 byte CRC***.

- **Quy trình phục hồi**:

+ Đọc toàn bộ file nhị phân.

+ Tìm tất cả vùng có dấu hiệu là ảnh .jpg hoặc .png.

+ Trích xuất các vùng hợp lệ thành file mới.

+ Ghi ra ổ đĩa để phục hồi dữ liệu.

2. Mã nguồn thực hiện

- Chuẩn bị đầu vào, cần có một file ảnh đĩa hoặc file nhị phân có chứa dữ liệu các ảnh .jpg hoặc .png bị mất/lẫn trong đó. Đọc toàn bộ nội dung nhị phân của file vào biến data.

```
# Đọc toàn bộ file image vào bộ nhớ (cẩn thận với file lớn)
with open(image_path, 'rb') as f:
    data = f.read()
```

- Xác định dấu hiệu nhận dạng ảnh, những cặp header và footer sẽ được dùng để xác định vùng dữ liệu có thể là ảnh.

```
# Dấu hiệu nhận diện JPG và PNG
jpg_header = b'\xFF\xD8\xFF'
jpg_footer = b'\xFF\xD9'
png_header = b'\x89\x50\x4E\x47\x0D\x0A\x1A\x0A'
png_footer = b'\x49\x45\x4E\x44\xAE\x42\x60\x82'
```

- Với mỗi cặp header-footer:

+ Tìm ***start = vị trí header***.

+ Tìm ***end = vị trí footer sau đó***.

+ Ghi đoạn ***data[start:end]*** ra file mới.

```
# Hàm tìm và phục hồi file
def extract_files(header, footer, ext):
    nonlocal found, offset
    pos = 0
    while pos < length:
        # Tìm vị trí header
        start = data.find(header, pos)
        if start == -1:
            break
        # Tìm vị trí footer sau header
        end = data.find(footer, start + len(header))
        if end == -1:
            break
        end += len(footer)
        found += 1
        filename = f"{output_dir}/recovered_{found:03d}.{ext}"
        with open(filename, 'wb') as out:
            out.write(data[start:end])
        print(f"Recovered: {filename} (offset {start}-{end})")
        # Tiếp tục tìm sau footer
        pos = end
```

- Lưu file ảnh đã phục hồi, ghi từng file phục hồi ra thư mục output_dir

```
filename = f"{output_dir}/recovered_{found:03d}.{ext}"
with open(filename, 'wb') as out:
    out.write(data[start:end])
print(f"Recovered: {filename} (offset {start}-{end})")
```

3. Kết quả thực nghiệm

- Được minh họa trong video đính kèm

II. KHÔI PHỤC TẬP TIN PDF

1. Ý tưởng thực hiện

- Tương tự như việc khôi phục tập tin ảnh, pdf cũng có những dấu hiệu đặc trưng để phát hiện

+ Bắt đầu bằng Header **%PDF (25 50 44 46)**

+ Kết thúc bằng **%%EOF**

2. Mã nguồn thực hiện

- Chuẩn bị file đĩa hoặc nhị phân chứa các đoạn PDF, đảm bảo đoạn dữ liệu PDF không bị thiếu header hoặc footer.

- Chạy hàm **recover_pdf_files()** để đọc toàn bộ file vào biến data và tìm các đoạn bắt đầu bằng **%PDF** và kết thúc bằng **%%EOF**.

```
def recover_pdf_files(image_path, output_dir):
    # Định nghĩa header và footer cho file PDF
    pdf_header = b'%PDF'
    pdf_footer = b'%%EOF'

    # Đọc toàn bộ file image vào memory (chỉ nên dùng cho image nhỏ, nếu lớn thì nên đọc từng phần)
    with open(image_path, 'rb') as f:
        data = f.read()
    length = len(data)
    found = 0

    if not os.path.exists(output_dir):
        os.makedirs(output_dir)

    start = 0
    while True:
        start = data.find(pdf_header, start)
        if start == -1:
            break
        end = data.find(pdf_footer, start)
        if end == -1:
            break
        end += len(pdf_footer)
        found += 1
```

- Mỗi đoạn hợp lệ được ghi ra file trong thư mục **recovered_pdf**.

```
        filename = f"{output_dir}/recovered_{found:03d}.pdf"
        with open(filename, 'wb') as out:
            out.write(data[start:end])
        print(f"Recovered: {filename} (offset {start}-{end})")
        start = end

    print(f"Đã phục hồi tổng cộng {found} file PDF vào thư mục '{output_dir}'.")
```

3. Kết quả thực nghiệm

- Được minh họa trong video đính kèm

--Hết--