

Thiết kế kỹ thuật và Hướng dẫn sử dụng của phần mềm Đếm trang A4

(Link phần mềm:

https://drive.google.com/file/d/19_n8hDbmP3Nc65_ho05yiXHNQV88Es1a/view?usp=sharing)

I. Thiết kế kỹ thuật phần mềm:

1. Kiến trúc tổng quan:

- Ứng dụng được phát triển bằng Python, sử dụng PyQt5 cho giao diện đồ họa.
- Sử dụng thư viện fitz (PyMuPDF) để xử lý file PDF.
- Sử dụng pandas để xử lý file Excel.

2. Các module chính:

- Giao diện người dùng (GUI):
- Sử dụng PyQt5 để tạo cửa sổ chính và các widget.
- Giao diện chính bao gồm hai tab: "Đếm trang A4" và "Thống kê file Metadata".
- Có thanh menu với các chức năng và Cài đặt.

3. Xử lý PDF:

- Module DemA4 xử lý việc đếm trang và phân tích kích thước trang PDF.
- Sử dụng đa luồng để tránh treo giao diện trong quá trình xử lý.

4. Xử lý Excel:

- Sử dụng pandas để đọc và phân tích file Excel metadata.
- Tính toán các thống kê như số hồ sơ, số văn bản, tổng số trang quy đổi.

5. Cài đặt:

- Module CaiDatDialog cho phép người dùng tùy chỉnh các thông số.

6. Luồng xử lý chính:

- Đếm trang A4:
 1. Chọn thư mục chứa file PDF
 2. Quét tất cả file PDF trong thư mục và thư mục con
 3. Phân tích từng file PDF, xác định kích thước trang
 4. Tính toán số trang quy đổi sang A4
 5. Hiển thị kết quả và cho phép lưu
- Thống kê file Metadata:
 1. Chọn file Excel metadata
 2. Đọc dữ liệu từ các sheet cụ thể
 3. Tính toán các thống kê yêu cầu
 4. Hiển thị kết quả

7. Nguyên lý hoạt động:

Chức năng Đếm trang A4:

- Quét thư mục: Sử dụng `os.walk()` để duyệt qua tất cả các file PDF trong thư mục được chọn và các thư mục con.
- Đọc từng file PDF: Sử dụng thư viện `fitz` (PyMuPDF) để mở và đọc thông tin của từng file PDF.
- Xác định kích thước trang: Với mỗi trang trong file PDF, lấy kích thước thực tế (chiều rộng và chiều cao).
- Phân loại kích thước trang: So sánh kích thước thực tế với kích thước chuẩn của các khổ giấy A0 đến A7. Xác định khổ giấy gần nhất dựa trên cạnh dài của trang.
- Xử lý trường hợp đặc biệt: Có xử lý đặc biệt cho trường hợp A4 ngang, nếu sự chênh lệch giữa chiều rộng và chiều cao nhỏ hơn một ngưỡng nhất định.
- Đếm số lượng trang theo từng khổ giấy: Tính tổng số trang cho mỗi khổ giấy (A0, A1, A2, ..., A7).
- Quy đổi sang trang A4: Sử dụng hệ số quy đổi để chuyển đổi số trang của các khổ giấy khác sang số trang A4 tương đương. Hệ số quy đổi mặc định: A0:16, A1:8, A2:4, A3:2, A4:1, A5:0.5, A6:0.25, A7:0.125
- Tổng hợp kết quả: Tính tổng số trang A4 quy đổi cho toàn bộ thư mục. Hiển thị thông tin chi tiết về số lượng trang của từng khổ giấy.
- Xử lý đa luồng: Sử dụng `QThread` để thực hiện việc đếm trang trong một luồng riêng biệt, tránh treo giao diện người dùng.
- Cập nhật tiến trình: Sử dụng tín hiệu (signals) để cập nhật tiến trình xử lý trên giao diện người dùng.
- Lưu kết quả: Cho phép người dùng lưu kết quả chi tiết vào file text sau khi hoàn thành.
- Lưu ý: Độ chính xác phụ thuộc vào việc xác định đúng khổ giấy, có thể cần điều chỉnh logic cho các trường hợp đặc biệt. Hệ số quy đổi có thể cần được điều chỉnh tùy theo yêu cầu cụ thể của người dùng hoặc tiêu chuẩn áp dụng.

Chức năng Thống kê file Metadata:

Đọc dữ liệu từ file Excel:

- Sử dụng thư viện `pandas` để đọc file Excel (.xlsx)
- Đọc dữ liệu từ hai sheet cụ thể: 'Files' và 'Docs'

Xác định vị trí cột dữ liệu:

- Sử dụng cài đặt do người dùng định nghĩa để xác định các cột chứa thông tin cần thống kê
- Chuyển đổi chữ cái cột (ví dụ: 'A', 'B', 'C') thành chỉ số cột số học

Tính toán thống kê:

- Tổng số hồ sơ Files: Đếm số giá trị duy nhất trong cột "Mã hồ sơ", bỏ qua hàng đầu tiên (thường là tiêu đề)
- Tổng số văn bản Docs: Đếm số giá trị duy nhất trong cột "Mã văn bản", bỏ qua hàng đầu tiên
- Tổng số trang quy đổi: Chuyển đổi dữ liệu trong cột "Số lượng trang quy đổi" sang dạng số. Tính tổng các giá trị trong cột này, bỏ qua hàng đầu tiên
- Tổng số trường trong hồ sơ và văn bản: Đếm tổng số ô không rỗng trong toàn bộ DataFrame

Xử lý lỗi và ngoại lệ:

- Sử dụng try-except để bắt và xử lý các lỗi có thể xảy ra trong quá trình đọc và xử lý file Excel
- Hiển thị kết quả: Kết quả thống kê được định dạng và hiển thị trong giao diện người dùng

Lưu ý:

Độ chính xác của kết quả phụ thuộc vào cấu trúc và tính nhất quán của dữ liệu trong file Excel đầu vào. Người dùng cần đảm bảo file metadata tuân thủ đúng định dạng yêu cầu để có kết quả chính xác.

II. Hướng dẫn sử dụng phần mềm:

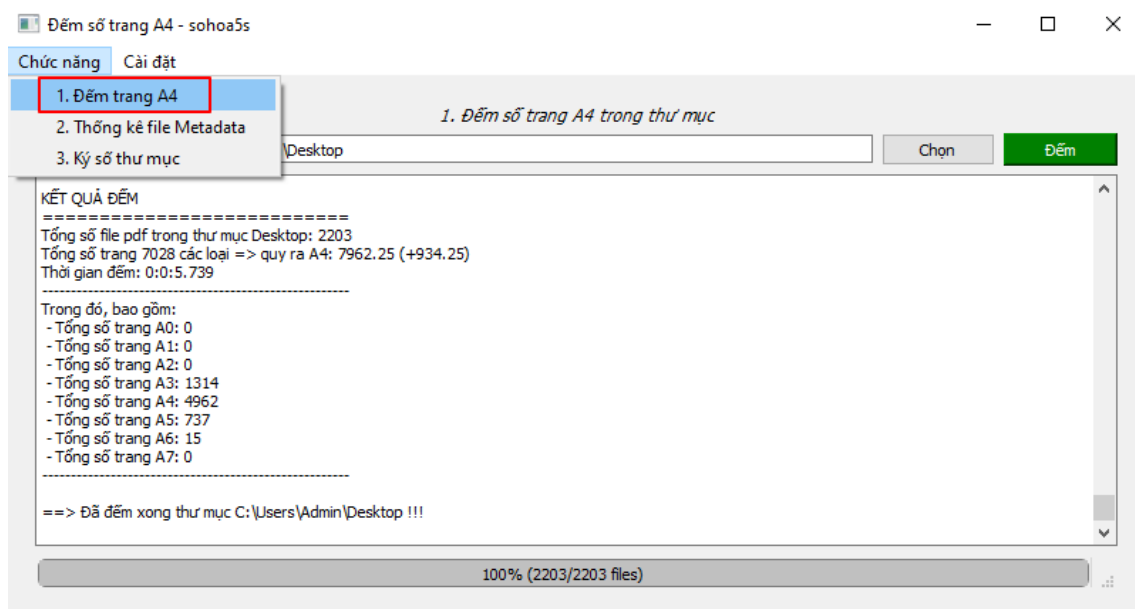
1. Khởi động ứng dụng:

- Chạy file chính của ứng dụng (ví dụ: python dema4_gom.py)



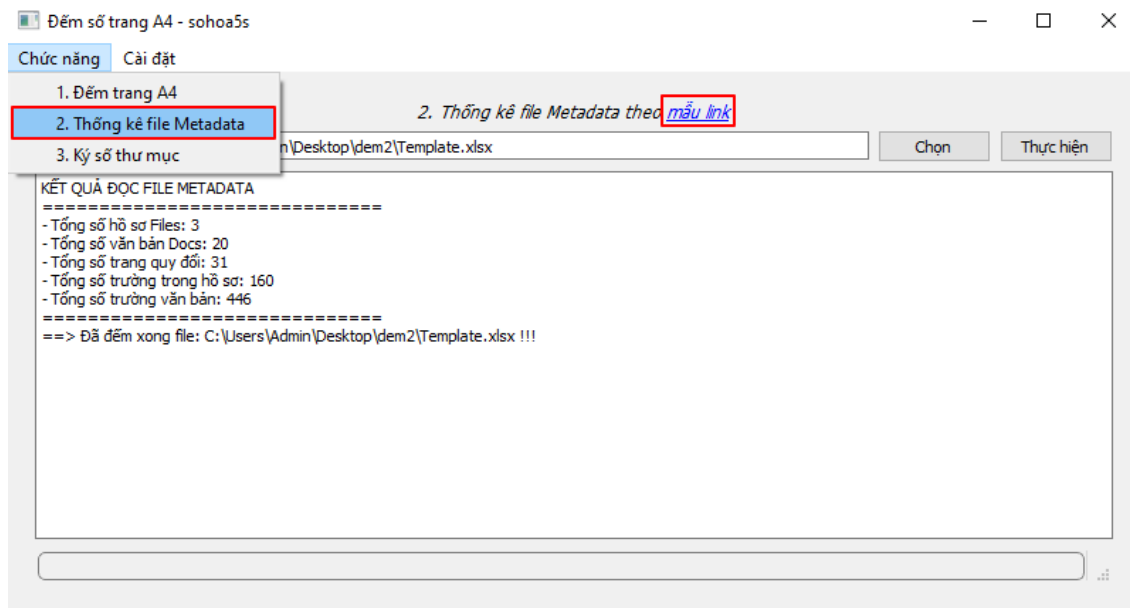
2. Đếm trang A4 PDF:

- Chọn tab "Đếm trang A4" trên giao diện chính.
- Nhấn nút "Chọn" để chọn thư mục chứa file PDF cần đếm.
- Nhấn nút "Đếm" để bắt đầu quá trình đếm.
- Theo dõi tiến trình trên thanh tiến độ.
- Kết quả sẽ hiển thị trong ô văn bản bên dưới.
- Sau khi hoàn thành, bạn có thể chọn lưu kết quả chi tiết vào file text.



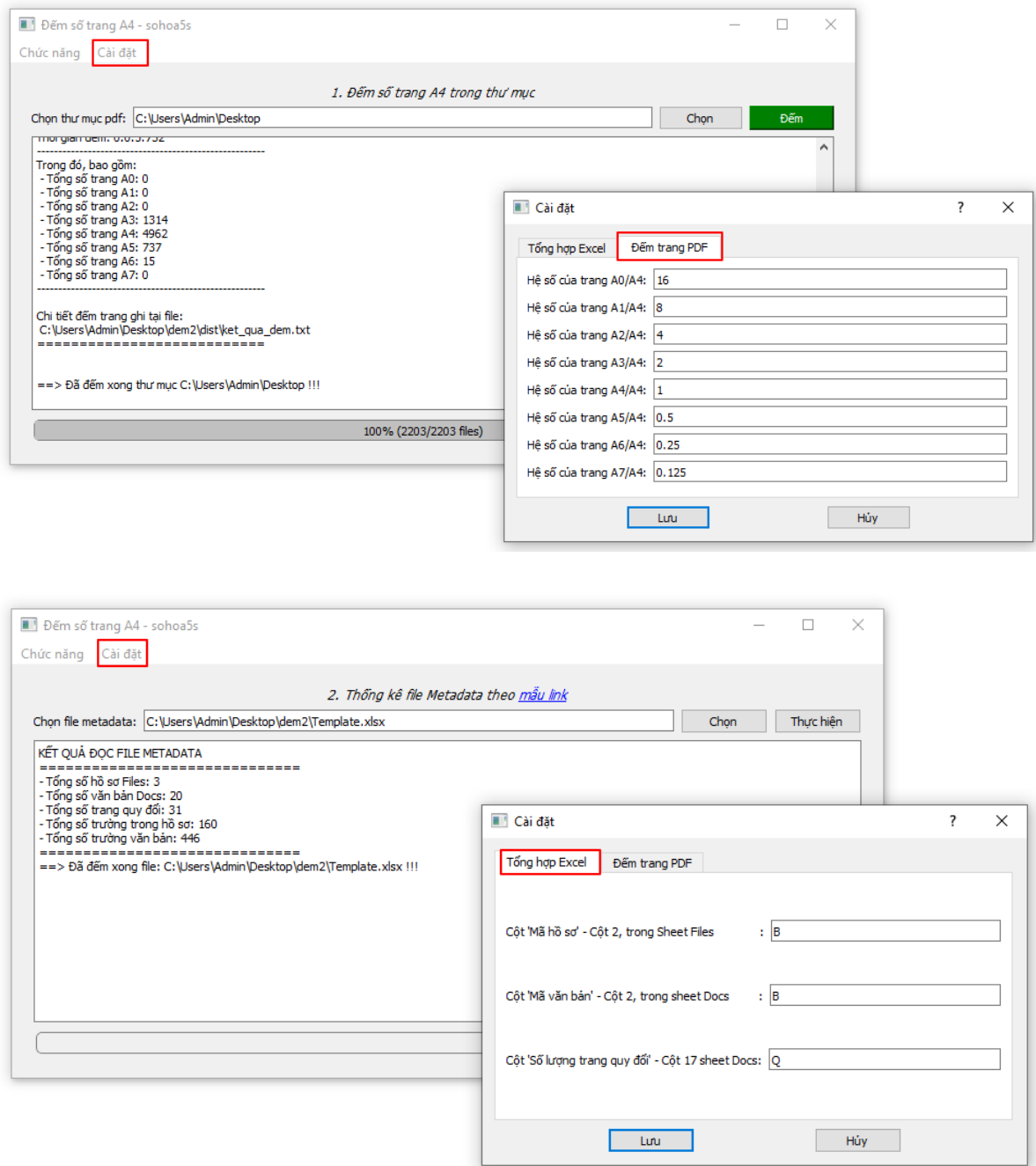
3. Thống kê file Metadata:

- Chọn tab "Thống kê file Metadata" trên giao diện chính.
- Nhấn nút "Chọn" để chọn file Excel metadata cần phân tích. File template mẫu: [link](#), chi tiết các trường trong bảng dưới.
- Nhấn nút "Thực hiện" để bắt đầu quá trình phân tích.
- Kết quả sẽ hiển thị trong ô văn bản bên dưới.



4. Cài đặt:

- Chọn "Cài đặt" từ thanh menu.
- Trong hộp thoại cài đặt, bạn có thể điều chỉnh:
 - Các cột dữ liệu trong file Excel metadata
 - Hệ số quy đổi cho các kích thước trang khác nhau
 - Nhấn "Lưu" để áp dụng cài đặt mới.



5. Lưu ý:

- Đảm bảo file Excel metadata có cấu trúc đúng với các sheet và cột dữ liệu phù hợp.
- Quá trình đếm trang PDF có thể mất thời gian với số lượng file lớn.
- Có thể dừng quá trình đếm bất cứ lúc nào bằng cách nhấn nút "Dừng".

6. Xử lý lỗi:

- Nếu gặp lỗi khi mở file hoặc xử lý dữ liệu, ứng dụng sẽ hiển thị thông báo lỗi.
- Kiểm tra đường dẫn file và định dạng dữ liệu nếu gặp vấn đề.

Cấu trúc dữ liệu Sheet Files trong file Metadata Export từ phần mềm (Dữ liệu đặc tả của Hồ sơ)

STT	Trường thông tin	Tên (viết tắt tiếng Anh)	Kiểu dữ liệu	Độ dài	Ghi chú
1	Mã hồ sơ	FileCode	String		
1.1	Mã định danh của cơ quan, tổ chức lập danh mục hồ sơ	OrganId	String	35	Bổ sung theo Nghị định 30/2020/NĐ-CP
1.2	Mã cơ quan lưu trữ lịch sử	Identifier	String	35	
1.3	Mã phong/công trình/sưu tập lưu trữ	OrganId	String	13	
1.4	Mục lục số hoặc năm hình thành hồ sơ	FileCatalog	Number	10	
1.4	Số và ký hiệu hồ sơ	FileNotation	String	20	
2	Tiêu đề hồ sơ	Title	String	1000	
3	Thời hạn bảo quản	Maintenance	String	100	
4	Chế độ sử dụng	Rights	String	30	
5	Người lập hồ sơ	Creator	String	30	Bổ sung theo Nghị định 0/2020/NĐ-CP
6	Ngôn ngữ	Language	String	100	
7	Thời gian bắt đầu	StartDate	Date	DD/MM/YYYY	
8	Thời gian kết thúc	EndDate	Date	DD/MM/YYYY	
9	Tổng số văn bản trong hồ sơ	TotalDoc	Number	10	
10	Chú giải	Description	String	2000	
11	Ký hiệu thông tin	InforSign	String	30	
12	Từ khóa	Keyword	String	100	
13	Số lượng tờ	Maintenance	Number	10	
14	Số lượng trang	PageNumber	Number	10	
15	Tình trạng vật lý	Format	String	50	

Cấu trúc dữ liệu Sheet Docs trong file Metadata Export từ phần mềm (Dữ liệu đặc tả của văn bản)

STT	Trường thông tin	Tên (viết tắt tiếng Anh)	Kiểu dữ liệu	Độ dài	Ghi chú
1	Mã định danh văn bản	DocCode	String	25	
2	Mã hồ sơ	FileCode			
2.1	Mã định danh của cơ quan, tổ chức lập danh mục hồ sơ	OrganId	String	35	Bổ sung theo Nghị định 30/2020/NĐ-CP
2.2	Mã cơ quan lưu trữ lịch sử ¹	Identifier	String	35	
2.3	Mã phong/công trình/sưu tập lưu trữ	OrganId	String	13	
2.4	Mục lục số hoặc năm hình thành hồ sơ	FileCatalog	Number	4	
2.5	Số và ký hiệu hồ sơ	FileNotation	String	20	
3	Số thứ tự văn bản trong hồ sơ	DocOrdinal	Number	4	
4	Tên loại văn bản	TypeName	String	100	
5	Số của văn bản	CodeNumber	String	11	
6	Ký hiệu của văn bản	CodeNotation	String	30	
7	Ngày, tháng, năm văn bản	IssuedDate	Date	DD/MM/YYYY	
8	Tên cơ quan, tổ chức ban hành văn bản	OrganName	String	200	
9	Trích yếu nội dung	Subject	String	500	
10	Ngôn ngữ	Language	String	100	
11	Số lượng trang của văn bản	PageAmount	Number	4	
12	Ghi chú	Description	String	500	
13	Ngày, tháng, năm đến (nếu là văn bản đến)	ArrivalDate	Date	DD/MM/YYYY	Bổ sung theo Nghị định 30/2020/NĐ-CP
14	Số đến (nếu là văn bản đến)	ArrivalNumber	Number	10	Bổ sung theo Nghị định 30/2020/NĐ-CP
15	Chức vụ, họ tên người ký	SignerInfo			Bổ sung theo Nghị định 30/2020/NĐ-CP
15.1	Chức vụ của người ký văn bản	Position	String	100	Bổ sung theo Nghị định 30/2020/NĐ-CP
15.2	Họ và tên người ký văn bản	FullName	String	50	Bổ sung theo Nghị định 30/2020/NĐ-CP

16	Nơi nhận (nếu là văn bản đi)	To			Bổ sung theo Nghị định 30/2020/NĐ-CP
16.1	Mã định danh của cơ quan, tổ chức nhận văn bản	OrganId	String	35	Bổ sung theo Nghị định 30/2020/NĐ-CP
16.2	Tên cơ quan, tổ chức nhận văn bản	OrganName	String	200	Bổ sung theo Nghị định 30/2020/NĐ-CP
17	Mức độ khẩn, độ mật	Priority	Number	1	Bổ sung theo Nghị định 30/2020/NĐ-CP
18	Đơn vị hoặc người nhận (nếu là văn bản đến)	ToPlaces	String	1000	Bổ sung theo Nghị định 30/2020/NĐ-CP
19	Ý kiến phân phối, chỉ đạo, trạng thái xử lý văn bản (nếu là văn bản đến)	TraceHeaderList	LongText		Bổ sung theo Nghị định 30/2020/NĐ-CP
20	Thời hạn giải quyết (nếu là văn bản đến)	DueDate	Date	DD/MM/YYYY	Bổ sung theo Nghị định 30/2020/NĐ-CP
21	Số lượng bản phát hành (nếu là văn bản đi)	IssuedAmount	Number	3	Bổ sung theo Nghị định 30/2020/NĐ-CP
22	Hạn trả lời văn bản (nếu là văn bản đi)	DueDate	Date	DD/MM/YYYY	Bổ sung theo Nghị định 30/2020/NĐ-CP
23	Ký hiệu thông tin	InforSign	String	30	
24	Từ khóa	Keyword	String	100	
25	Chế độ sử dụng	Mode	String	20	
26	Mức độ tin cậy	ConfidenceLevel	String	30	
27	Bút tích	Autograph	String	2000	
28	Tình trạng vật lý	Format	String	50	

-----Hết-----