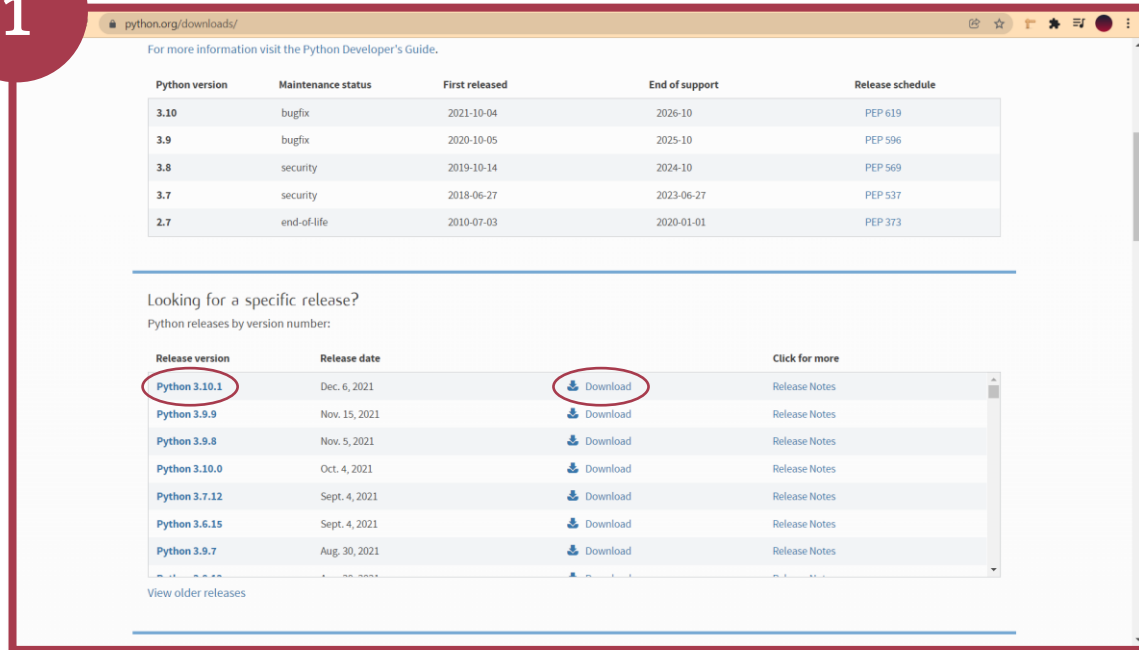


HƯỚNG DẪN CÀI ĐẶT PDF OPTIMIZER

1 – CÀI ĐẶT PYTHON

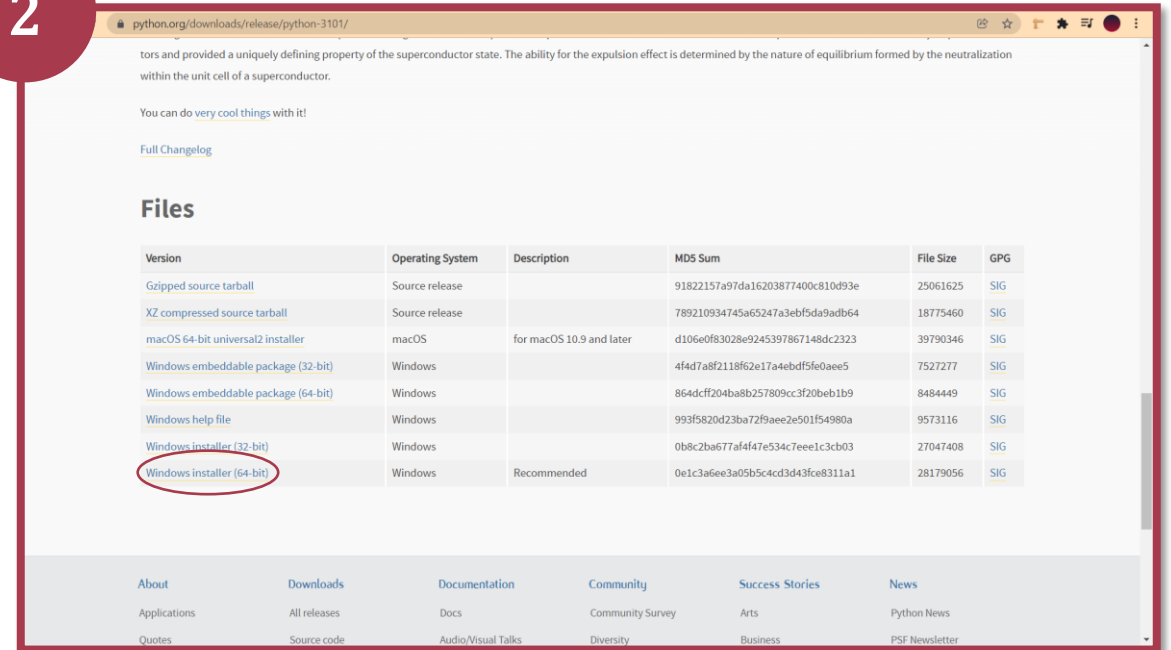
Download Python 3.10.1 tại www.python.org/downloads

1



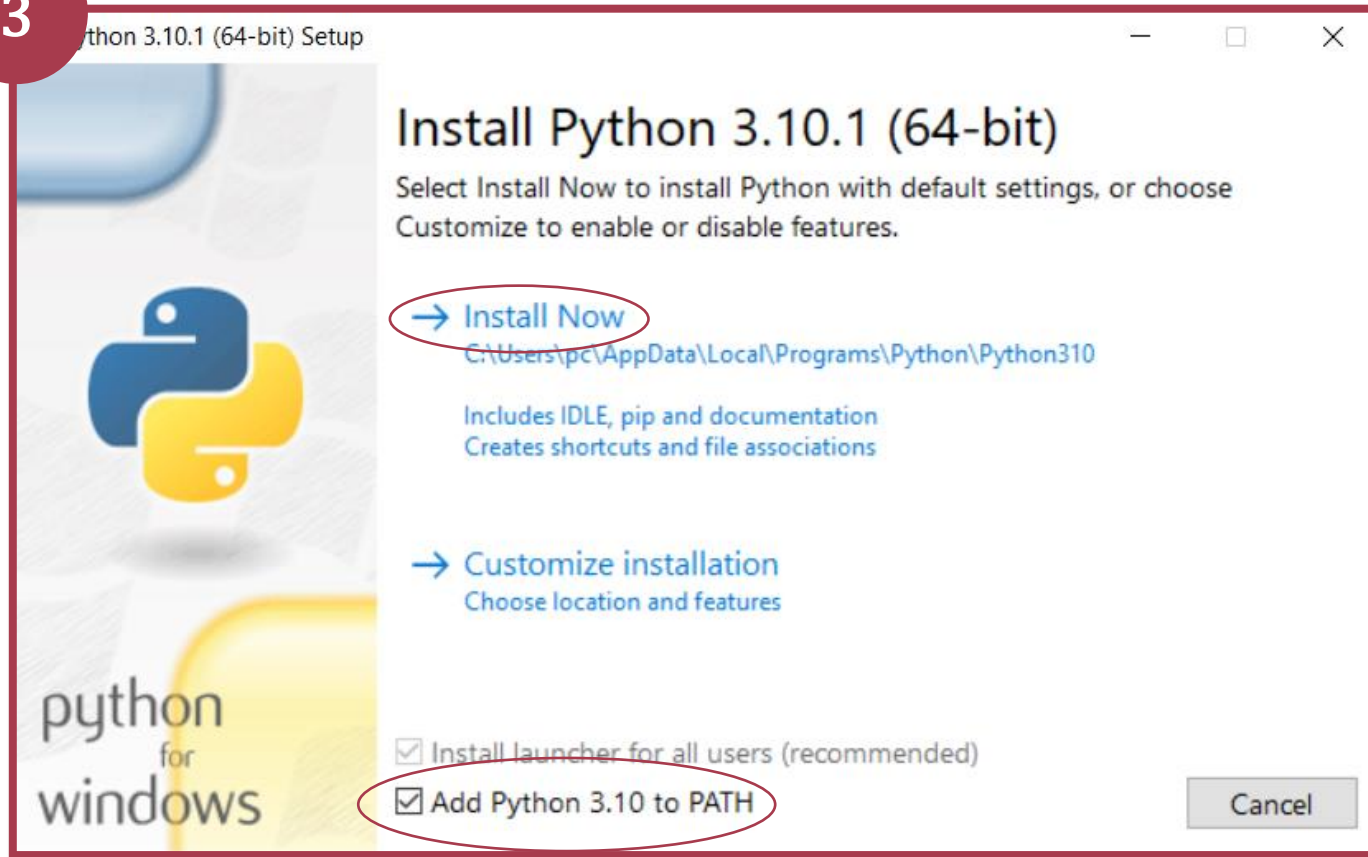
Tìm phiên bản 3.10.1 và nhấn vào Download.

2



Tiếp tục cuộn đến cuối trang và nhấn vào Windows Installer (64-bit) để tiến hành download.

3



Tại giao diện cài đặt python, đánh dấu vào ô **“Add Python 3.10 to PATH”**, sau đó nhấn **“Install Now”**.

Chờ đến khi cài đặt xong và nhấn **“Close”**. (Không nhấn thêm bất cứ gì khác)

2 – CÀI ĐẶT THƯ VIỆN (Chỉ thực hiện bước này khi máy đã được cài đặt Python)

```
Microsoft Windows [Version 10.0.19042.1415]
(c) Microsoft Corporation. All rights reserved.

C:\Users\pc>pip install PyMuPDF
Collecting PyMuPDF
  Using cached PyMuPDF-1.19.4-cp310-cp310-win_amd64.whl (6.4 MB)
Installing collected packages: PyMuPDF
Successfully installed PyMuPDF-1.19.4

C:\Users\pc>
```

Có 3 gói thư viện cần cài đặt bao gồm:
fitz, PyMuPDF, Pillow

Nhấn nút Windows  và gõ “cmd” để mở giao diện Command line.

Lần lượt thực thi các lệnh sau:

- 1) pip install --upgrade pip
- 2) pip install fitz
- 3) pip install PyMuPDF
- 4) pip install Pillow

(Trong quá trình cài đặt, sẽ có lúc bị “đơ”. Dấu hiệu là chạy lâu, nên lúc này cần nhấn phím bất kỳ trên bàn phím để tiếp tục).

- Để kiểm tra đã cài đặt thành công hay chưa thì gõ “pip list” và nhấn “enter”.
- Sau đó kiểm tra các gói thư viện có xuất hiện trong danh sách hay không. Nếu không xuất hiện nghĩa là gói thư viện chưa được cài đặt và cần cài đặt lại.

```
Select Command Prompt

C:\Users\pc>pip list
Package            Version
-----
certifi             2021.10.8
charset-normalizer  2.0.9
ci-info             0.2.0
click               8.0.3
colorama            0.4.4
configobj           5.0.6
configparser        5.2.0
etelemetry          0.2.2
filelock            3.4.2
fitz                 0.0.1.dev2
httplib2            0.20.2
idna                 3.3
isodate              0.6.1
lxml                 4.7.1
networkx            2.6.3
nibabel              3.2.1
nipy                1.7.0
numpy               1.22.0
packaging            21.3
pandas              1.3.5
Pillow              8.4.0
pip                 21.3.1
prov                2.0.0
pydot               1.4.2
PyMuPDF             1.19.4
pyparsing            3.0.6
PyPDF3              1.0.5
```

HƯỚNG DẪN SỬ DỤNG PDF OPTIMIZER

ĐIỀU KIỆN FILE ĐỂ CÓ THỂ SỬ DỤNG CHƯƠNG TRÌNH

Đầu vào (input) phải là:

- 1) File PDF (1 lớp hoặc 2 lớp) của **ảnh**. Ví dụ như file scan, file thuần ảnh không bao gồm text. Nếu có text thì là lớp text nằm phía trên ảnh được tự động thêm vào qua các phần mềm nhận diện text trong ảnh (dạng PDF 2 lớp).
- 2) Giới hạn số trang của file PDF để chương trình xử lý chính xác là 9999 trang, nếu vượt quá số trang đó thì file sau quá trình nén (output) sẽ không còn đúng thứ tự trang.

MÔ TẢ CHƯƠNG TRÌNH

Chương trình bao gồm 4 thông số chính:

- 1) **Quality:** Chất lượng ảnh (hay chất lượng file pdf) được giữ lại sau khi nén. Giá trị quality dao động từ 0 đến 95,. Quality càng lớn thì chất lượng ảnh càng cao và dung lượng cũng càng cao.
- 2) **DPI:** Giá trị ảnh hưởng tới chất lượng bản in.
- 3) **Image size (%):** Tỷ lệ kích thước ảnh được giữ lại.
- 4) **Image directory:** Thư mục chứa hình ảnh được trích xuất từ file pdf, thư mục tự động được tạo nếu không tồn tại và tự động xóa sau khi hoàn thành việc nén file.

Giá trị mặc định của 4 thông số trên:

Quality: 6

DPI: 300, 300


Image_size: 100

Image directory: image_page

CÁCH SỬ DỤNG

Hướng dẫn cài đặt mặc định UTF-8 khi mở notepad:
<https://www.youtube.com/watch?v=cxMrIW2EXfo>

1) Chuẩn bị file .txt chứa các file PDF cần nén theo định dạng sau:



File/Folder 1 Xuống hàng

Dòng 1: Đường dẫn của file .pdf HOẶC của THƯ MỤC chứa các file .pdf
Dòng 2: Đường dẫn của nơi chứa file kết quả sau khi nén
Dòng 3: Thông số 1
Dòng 4: Thông số 2
Dòng 5: Thông số 3

File/Folder 2 Xuống hàng

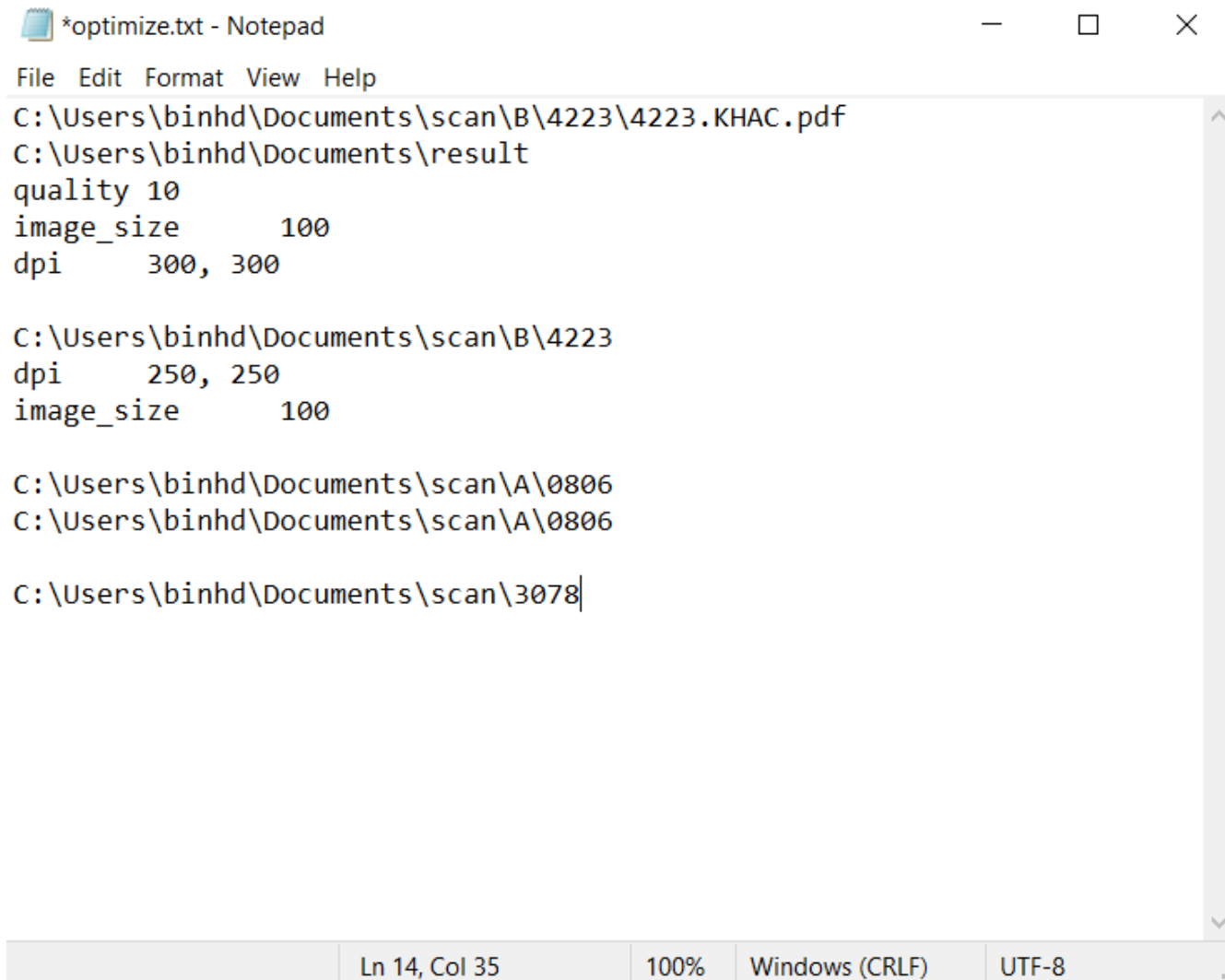
Dòng 1: Đường dẫn của file .pdf hoặc của THƯ MỤC chứa các file .pdf
Dòng 2: Đường dẫn của nơi chứa file kết quả sau khi nén
Dòng 3: Thông số 1
Dòng 4: Thông số 2
Dòng 5: Thông số 3

File/Folder 3

Dòng 1: Đường dẫn của file .pdf hoặc của THƯ MỤC chứa các file .pdf
Dòng 2: Đường dẫn của nơi chứa file kết quả sau khi nén
Dòng 3: Thông số 1
Dòng 4: Thông số 2
Dòng 5: Thông số 3

File .txt **PHẢI** được save ở dạng UTF-8

Một số mẫu hợp lệ



```
*optimize.txt - Notepad
File Edit Format View Help
C:\Users\binhd\Documents\scan\B\4223\4223.KHAC.pdf
C:\Users\binhd\Documents\result
quality 10
image_size      100
dpi      300, 300

C:\Users\binhd\Documents\scan\B\4223
dpi      250, 250
image_size      100

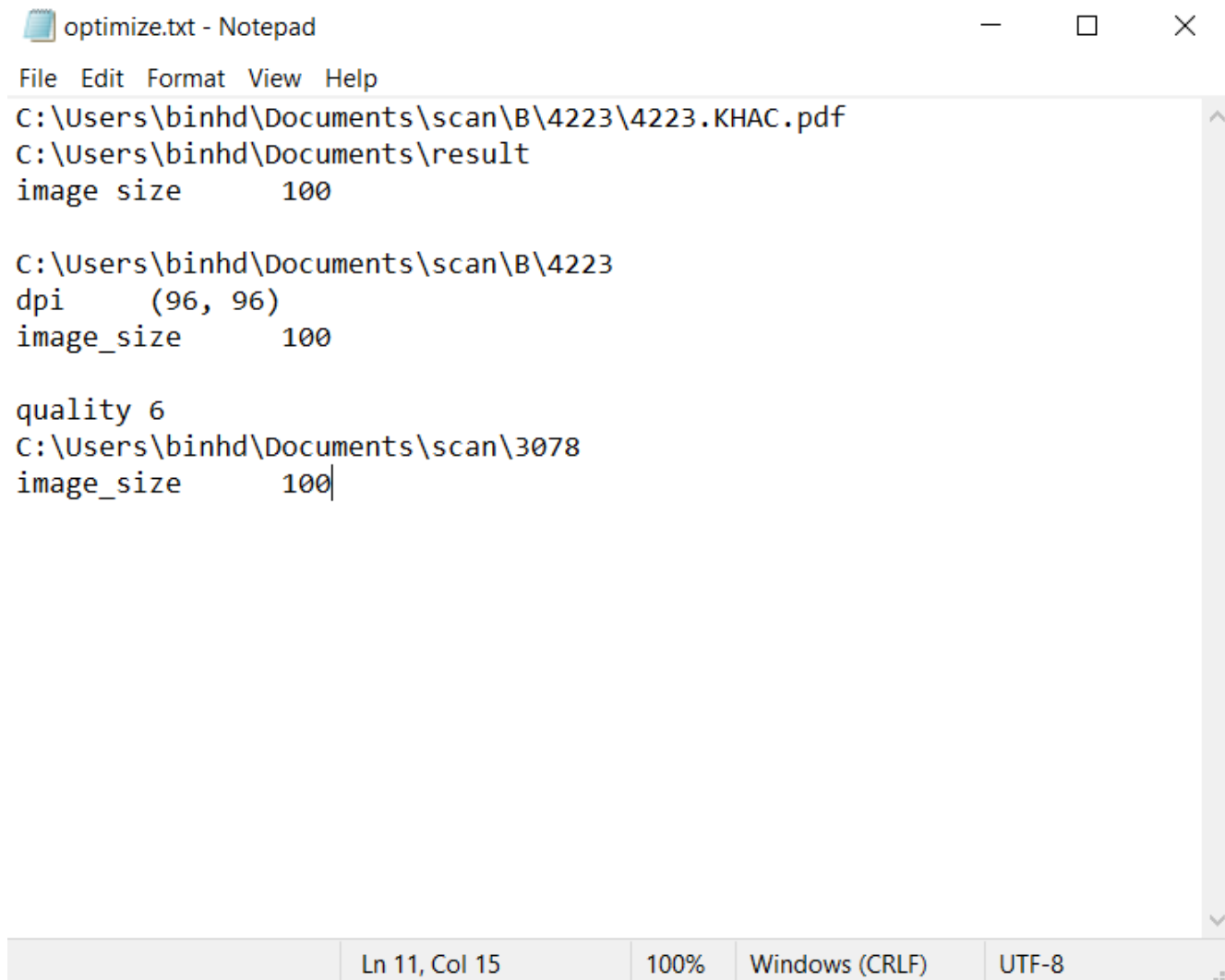
C:\Users\binhd\Documents\scan\A\0806
C:\Users\binhd\Documents\scan\A\0806

C:\Users\binhd\Documents\scan\3078|

Ln 14, Col 35    100%    Windows (CRLF)    UTF-8
```

- Dòng 1 **bắt buộc** phải là đường dẫn file hoặc thư mục đầu vào.
- Dòng 2 có thể có hoặc không có đường dẫn chứa file kết quả. Nếu không có thì đường dẫn sẽ giống đường dẫn dòng 1.
- Các thông số có thể đảo thứ tự, viết hoa, thường tùy ý nhưng **PHẢI ĐÚNG CHÍNH TẢ**.
- Các giá trị phải cách tên thông số bằng **1 TAB**.
- Nếu các thông số không được nhập sẽ có giá trị mặc định như đã đề cập ở phần Mô tả.

Một số mẫu không hợp lệ



```
optimize.txt - Notepad
File Edit Format View Help
C:\Users\binhd\Documents\scan\B\4223\4223.KHAC.pdf
C:\Users\binhd\Documents\result
image size      100

C:\Users\binhd\Documents\scan\B\4223
dpi      (96, 96)
image_size      100

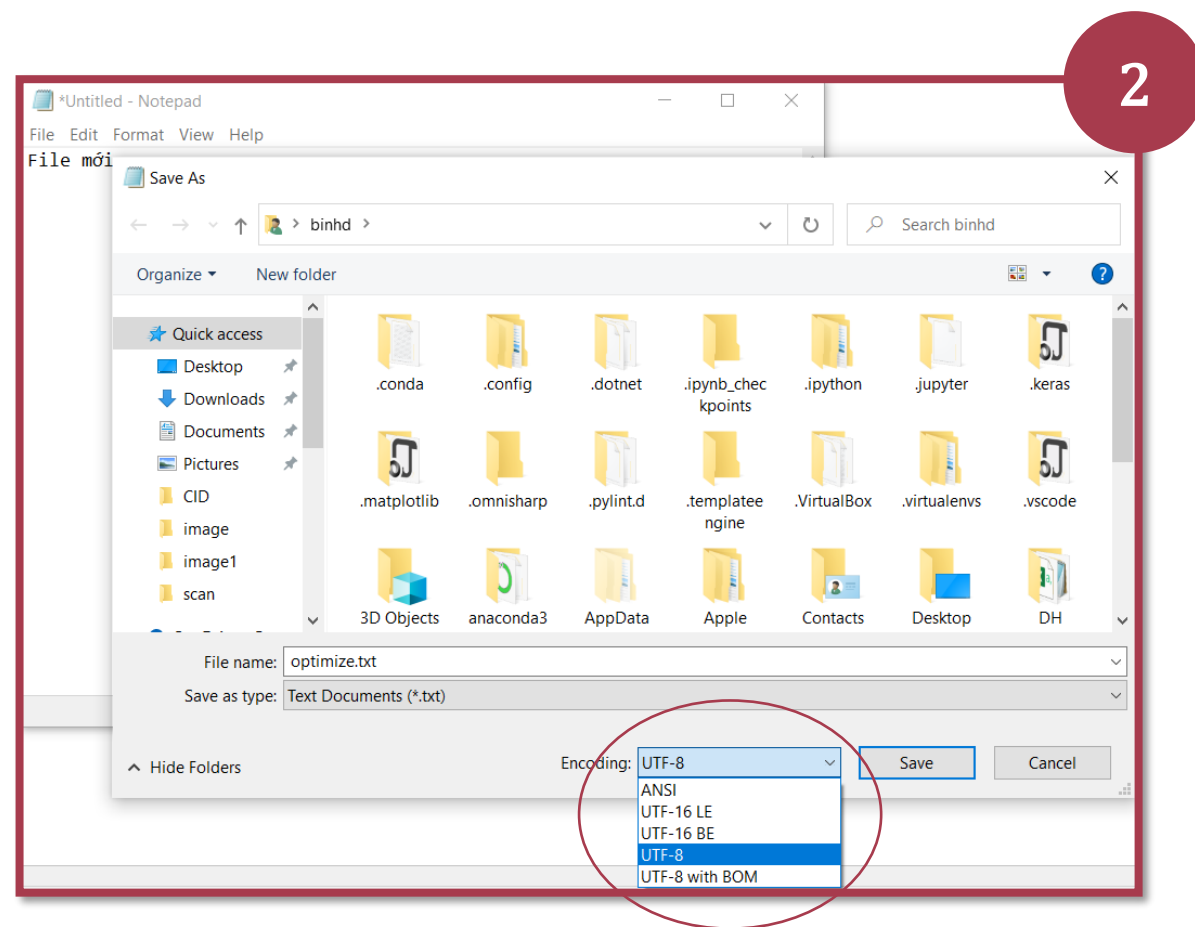
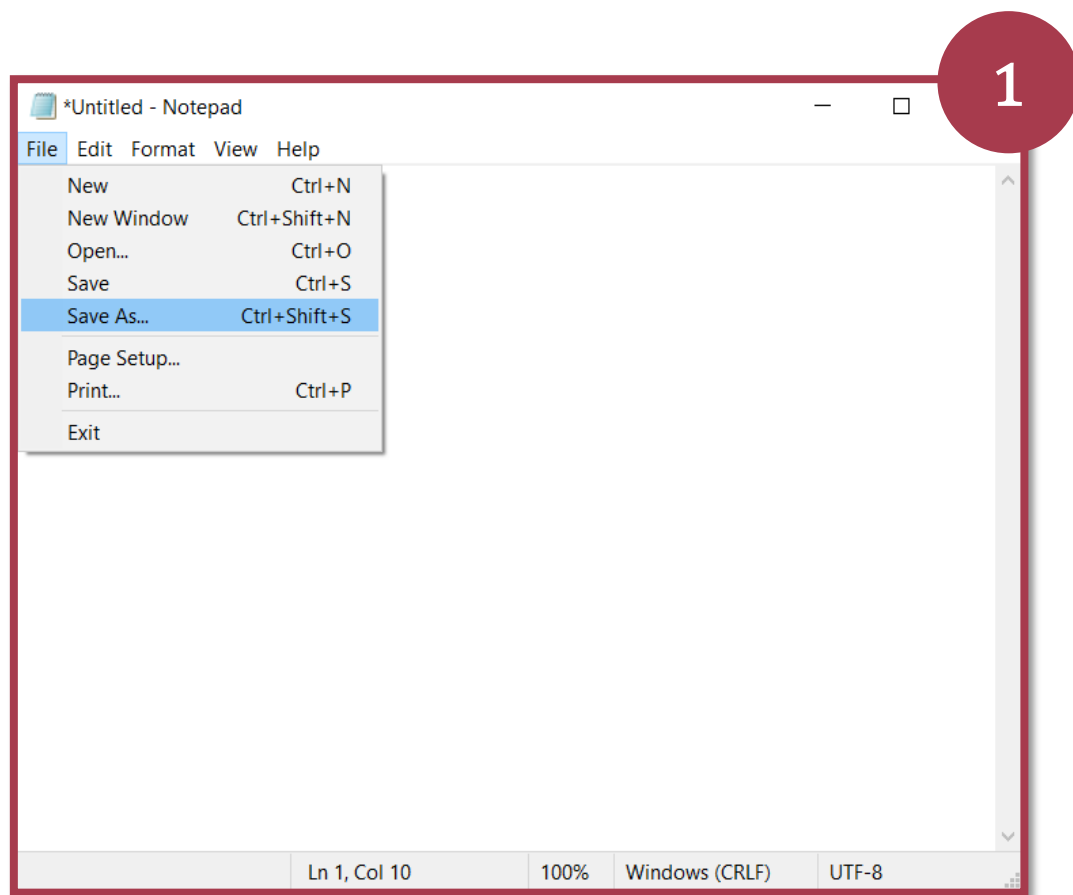
quality 6
C:\Users\binhd\Documents\scan\3078
image_size      100|

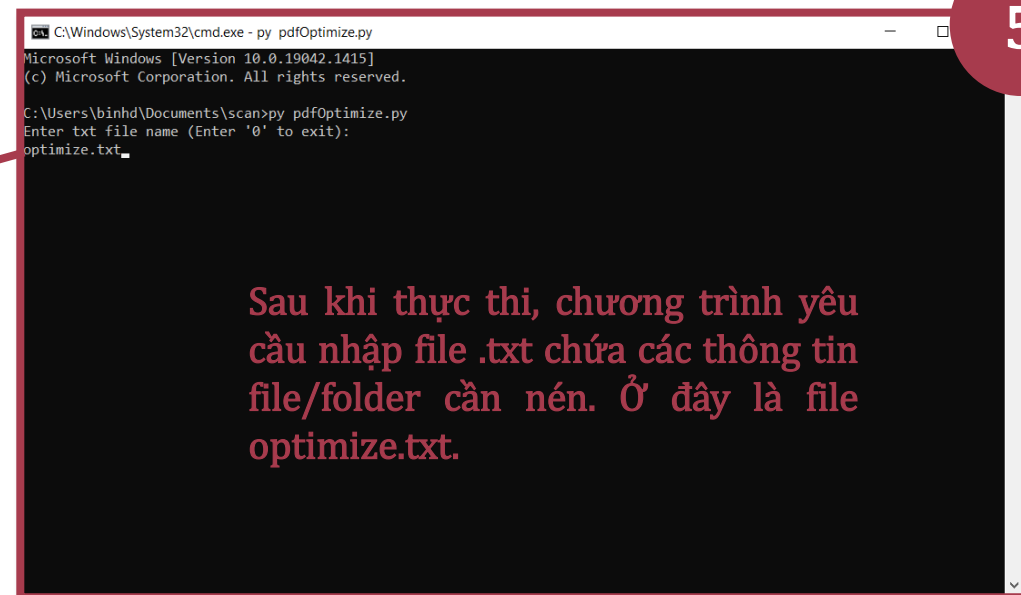
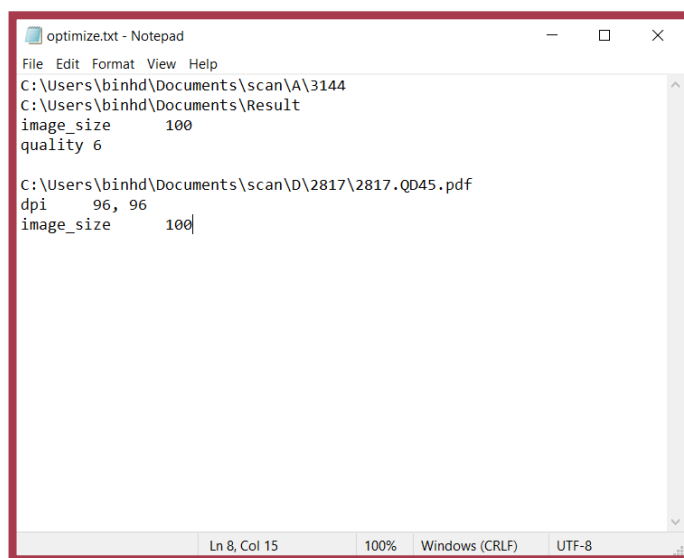
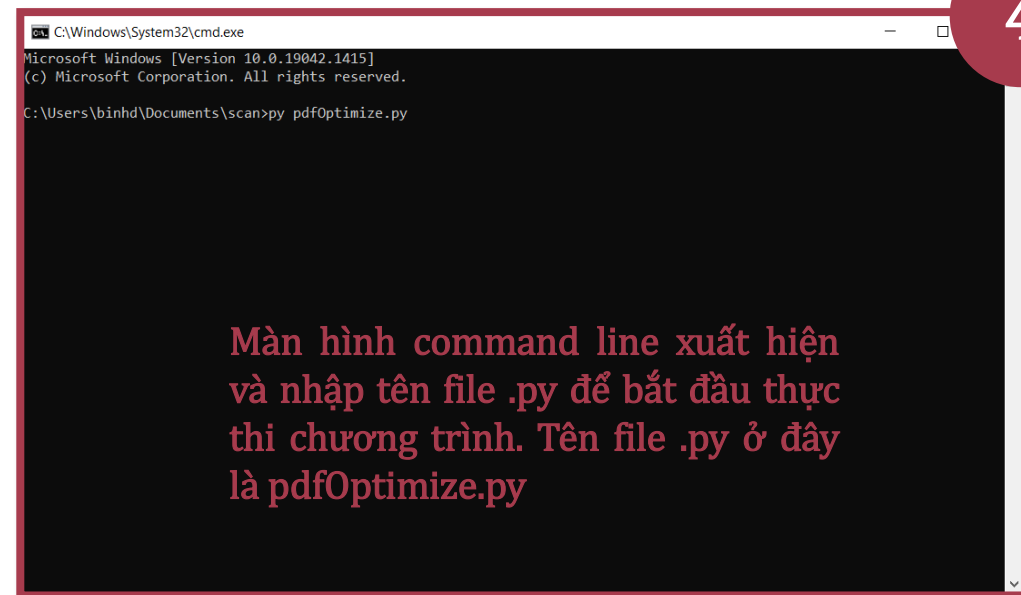
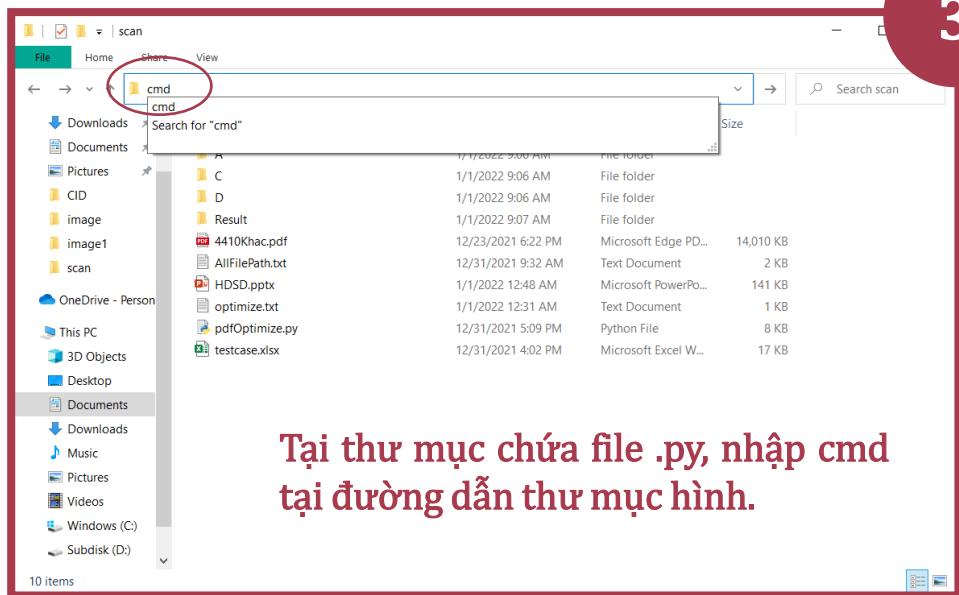
Ln 11, Col 15    100%    Windows (CRLF)    UTF-8
```

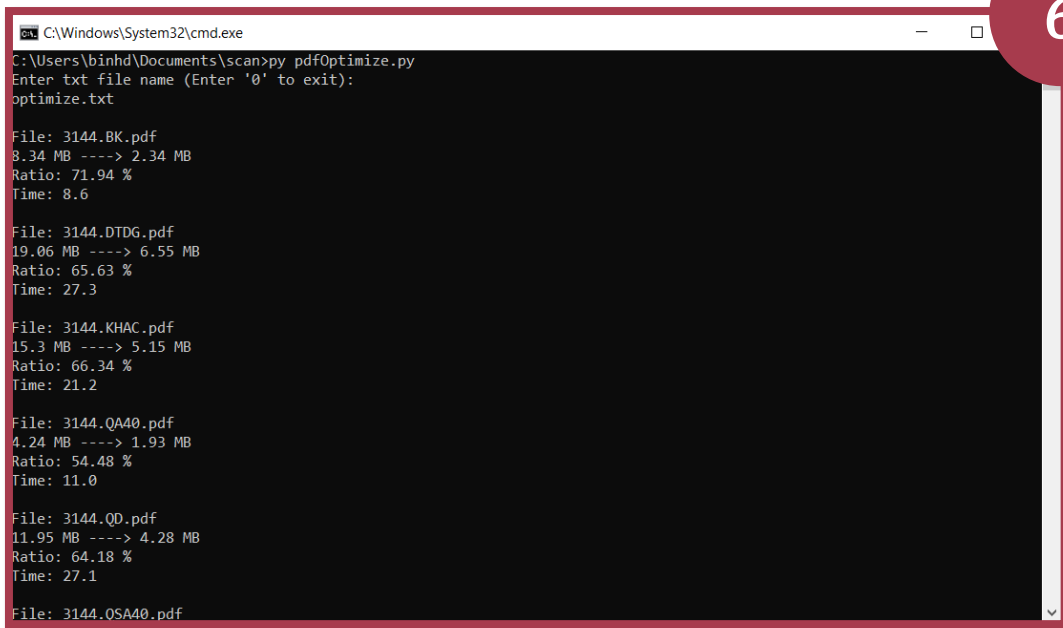
- image size: sai chính tả. Phải là **image_size**.
- dpi: sai cú pháp. Phải là **96, 96**.
- Cuối cùng là sai thứ tự, đường dẫn input phải đặt ở đầu.

2) Thực thi chương trình

Chuẩn bị file thông tin, save ở dạng **UTF-8** thì chương trình mới đọc được dấu.







```
C:\Windows\System32\cmd.exe
C:\Users\binhd\Documents\scan>py pdfOptimize.py
Enter txt file name (Enter '0' to exit):
optimize.txt

File: 3144.BK.pdf
8.34 MB ----> 2.34 MB
Ratio: 71.94 %
Time: 8.6

File: 3144.DTDG.pdf
19.06 MB ----> 6.55 MB
Ratio: 65.63 %
Time: 27.3

File: 3144.KHAC.pdf
15.3 MB ----> 5.15 MB
Ratio: 66.34 %
Time: 21.2

File: 3144.QA40.pdf
4.24 MB ----> 1.93 MB
Ratio: 54.48 %
Time: 11.0

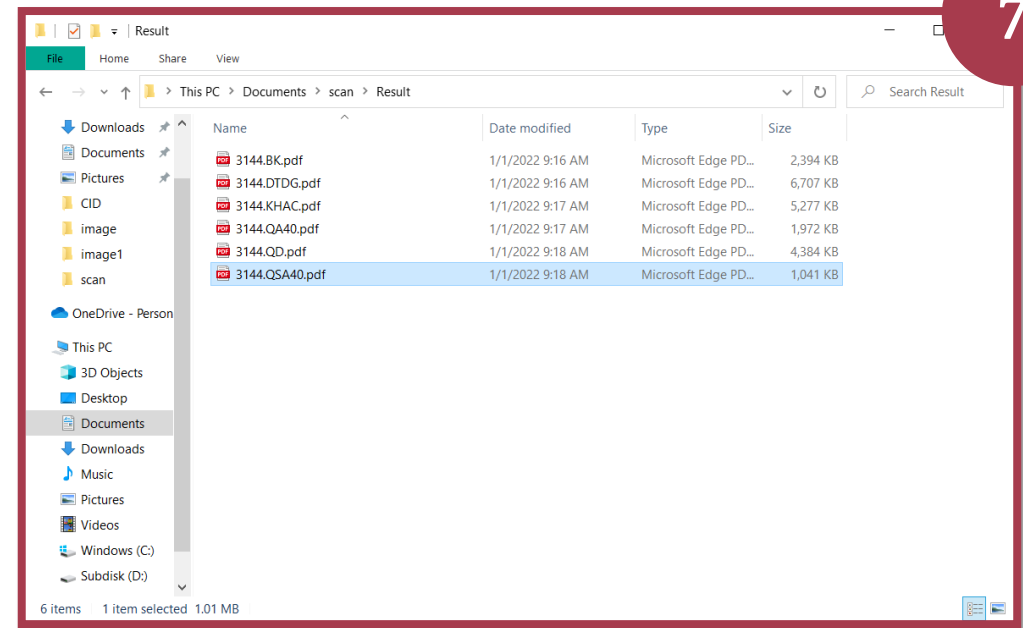
File: 3144.QD.pdf
11.95 MB ----> 4.28 MB
Ratio: 64.18 %
Time: 27.1

File: 3144.QSA40.pdf
```

Trong quá trình nén file, mỗi file sẽ được kiểm soát quá trình nén và cũng được đo đạc các thông tin sau nếu nén thành công:

- Kích thước ban đầu → Kích thước sau khi nén
- Tỷ lệ kích thước đã được nén
- Thời gian thực thi file (giây)

Nếu file nén không thành công, chương trình sẽ thông báo “Having some problem”.



- Kết quả nén tại thư mục được quy định trong file optimize.txt
- Các trường hợp không nhập đường dẫn xuất file thì chương trình sẽ tự động tạo 1 thư mục tên là “**compressed**” chứa các file đã xuất, thư mục này đặt trong thư mục chứa file input.

ĐÁNH GIÁ

File name	Page	Quality	Radius	Filter size	Time (s)	Size (KB)	Final size (KB)	Size (MB)	Final size (MB)	Size/Page (KB)	Ratio (%)
0980.BK.pdf	28	6	2	3	34.8	11412	2443	11.14	2.39	87.3	-78.59
2840.QD.pdf	39	6	2	3	52.9	19573	4874	19.11	4.76	125	-75.1
4075.KHAC.pdf	55	6	2	3	97.3	22936	6155	22.4	6.01	111.9	-73.16
2847.KHAC.pdf	90	6	2	3	127.3	31449	8185	30.71	7.99	90.9	-73.97
2383.QD.pdf	45	6	2	3	81.2	29256	6545	28.57	6.39	145.4	-77.63
3002.KHAC.pdf	203	6	2	3	330.1	58409	17892	57.04	17.47	88.1	-69.37
4051.KHAC.pdf	208	6	2	3	359.6	81519	20347	79.61	19.87	97.8	-75.04
4216.QD.pdf	36	6	2	3	102.1	33607	5468	32.82	5.34	151.9	-83.73
4379.KHAC.pdf	243	6	2	3	408.3	117758	30968	115	30.24	127.4	-73.7
4223.KHAC.pdf	3	6	2	3	5.7	604	386	0.59	0.38	128.7	-36.09
3918.DTDG.pdf	30	6	2	3	51	8630	2374	8.43	2.32	79.1	-72.49
4159.BK.pdf	35	6	2	3	52.2	9521	2446	9.3	2.39	69.9	-74.31
2424.BK.pdf	28	6	2	3	44.6	18392	3247	17.96	3.17	116	-82.35
3268.DTDG.pdf	51	6	2	3	120	18979	5146	18.53	5.03	100.9	-72.89
3711.DTDG.pdf	66	6	2	3	118.2	25748	6806	25.14	6.65	103.1	-73.57
0949		6	2	3	262.9			47.2	12.1		-74.36
3704		6	2	3	328.9			43.9	11.4		-74.03
3886		6	2	3	350.1			72.8	11.2		-84.62
4470		6	2	3	701.8			93.8	24.3		-74.09
2442		6	2	3	388.2			87.7	22.2		-74.69

- Thông tin trên đo đạc được khi tiến hành nén 15 file .pdf và 5 thư mục.
- Tỷ lệ nén trung bình 75%, nghĩa là kích thước sau cùng còn 1/4 so với ban đầu.
- Kích thước mỗi trang trung bình là 108 KB.