

Th.S LÊ NGỌC THẠCH

**ỨNG DỤNG
PHÂN TÍCH DỮ LIỆU
VÀ
TRÍ TUỆ NHÂN TẠO
VỚI PYTHON**

2021

Đây là eBook của riêng bạn – đề nghị không chia sẻ cho ai khác nhé!

Lời nhắn

eBook "ỨNG DỤNG PHÂN TÍCH DỮ LIỆU VÀ TRÍ TUỆ NHÂN TẠO VỚI PYTHON" này dự kiến phát hành vào tháng 12/2021. Bạn có thể đặt hàng ngay bây giờ với ưu đãi giảm 50% bằng 2 cách sau:

Cài **App MinePI** cho điện thoại tại theo link:

<https://minepi.com/thachln>

Sử dụng invitation code: **thachln**

Thử dùng điện thoại để đào Pi Coin. eBook được chấp nhận thanh toán Pi Coin với giá tương đương 399K đồng (Xem hình thức thanh toán tiền mặt).

Phiên bản bạn đang nhận là bản nháp trong quá trình hoàn thiện.

Bạn được gửi riêng để tham khảo hoặc để góp ý. Vì thế bạn được toàn quyền sử dụng và **KHÔNG** chia sẻ với bất kỳ ai khác nhé, **KHÔNG** lưu trữ trên internet nói chung để hạn chế đến tay người không thật sự cần nó!

Về nội dung bạn thu lượm được từ eBook dưới dạng các bài tóm tắt, đánh giá, hoặc đề nghị bổ sung thì rất được **KHUYẾN KHÍCH** chia sẻ công khai.



Đặc biệt khuyến khích bạn chia sẻ link:

<https://ThachLN.github.io>

Lê Ngọc Thạch

Hãy cài app [MinePI](https://minepi.com/thachln) ngay với Invitation Code là **thachln** để nhận ngay bản nháp (hơn 600 trang) nhé!

Hình thức thanh toán tiền mặt – Đặt hàng ngay bây với 199K, tiết kiệm 200K qua:

① MoMo	② Chuyển khoản
<div><p>Thanh toán qua MoMo</p><p>0908550642 Lê Ngọc Thạch</p><p>Nội dung tin nhắn: email sdt Ví dụ: abc@gmail.com 0908550642 AIPYTHON Email và sdt của người nhận eBook.</p><p>Trường hợp tặng bạn bè thì ghi thông tin email và sdt của bạn.</p><p>Quét mã QR thanh toán 199K.</p><p>199.000đ</p></div>	<div><p>Thanh toán qua NH Tiên Phong</p><p>Lê Ngọc Thạch, Ngân Hàng Tiên Phong, CN HCM Số tài khoản: 00002888001 Nội dung tin nhắn: email sdt AIPYTHON Vd tin nhắn: abc@gmail.com 0908456321 AIPYTHON</p><p>Quét mã QR để thanh toán cho:</p><p>Quét mã vạch này để giao dịch</p></div>

Mục lục

Quy ước	7
Ngày 1 – Chủ đề: Giới thiệu về ngôn ngữ thống kê, ngôn ngữ lập trình	10
Bài 1: Tóm tắt về thống kê (Statistics)	12
Bài 2: Ngôn ngữ lập trình Python	19
Bài 3: Ngôn ngữ Python và phần mềm Anaconda	27
Bài 4: Cài đặt thêm phần mềm	48
Bài 5: Nhập liệu, biên tập, lưu trữ dữ liệu với Python	53
Bài tập ngày 1	70
Thử thách cho bạn!	72
Ngày 2 – Chủ đề: Biểu đồ	73
Bài 6: Các loại biểu đồ	75
Bài 7: Vẽ biểu đồ trong Python	81
Bài 8: Nguyên tắc soạn biểu đồ	96
Bài 9: Giới thiệu Matplotlib	98
Bài 10: Giới thiệu Bokeh	112
Bài 11: Khai phá Bokeh	120
Ngày 3 – Phân tích mô tả	146
Bài 12: Phân tích mô tả dữ liệu Bank Marketing	148
Bài 13: Phân tích dữ liệu Marketing #2	159
Bài 14: So sánh 2 tỉ lệ	166
Bài 15: Mô hình kiểm định giả thuyết	177
Bài 16: Ứng dụng minh họa kiểm định giả thuyết	178
Bài 17: Phân tích mối tương quan	188
Ngày 4 – Chủ đề: Dữ liệu lớn	196
Bài 18: Cách xử lý tập hợp dữ liệu lớn	197
Bài 19: Sử dụng Ubuntu	232
Bài 20: Cài đặt Hadoop 3.2	241
Bài 21: Trải nghiệm Hadoop với Python	249
Ngày 5 – Chủ đề: Dự báo bằng mô hình hồi qui tuyến tính	255

Đây là eBook của riêng bạn – đề nghị không chia sẻ cho ai khác nhé!

Bài 22: Giới thiệu mô hình hồi qui tuyến tính	256
Bài 23: Diễn giải mô hình hồi qui tuyến tính.....	260
Bài 24: Mô hình hồi qui tuyến tính đa biến.....	273
Bài 25: Dự báo bằng mô hình hồi qui tuyến tính	277
Ngày 6 – Chủ đề: Dự báo bằng mô hình hồi qui logistic	281
Bài 26: Giới thiệu mô hình hồi qui logistic.....	282
Bài 27: Mô hình hồi qui logistic đa biến (Multiple logistic regression model).....	286
Bài 28: So sánh mô hình.....	290
Bài 29: Dự báo bằng mô hình hồi qui logistic	296
Ngày 7 – Chủ đề: Phân tích đa biến	303
Bài 30: Xử lý giá trị trống	304
Bài 31: Mô hình phân tích phân định (Linear discriminant analysis) ..	308
Bài 32: Mô hình thành phần (Principal Component Analysis)	316
Bài 33: Mô hình phân tích cụm/nhóm (cluster analysis)	324
Ngày 8 – Chủ đề: Machine Learning	332
Bài 34: Giới thiệu Machine learning	333
Bài 35: Mô hình SVM.....	335
Bài 36: Mô hình Random Forest	343
Bài 37: Mô hình Artificial Neural Network	347
Bài 38: Machine Learning với Python Tensorflow.....	353
Ngày 9 – Chủ đề: Recommendation.....	382
Bài 39: Giới thiệu phương pháp gợi ý Collaborative filtering	383
Bài 40: Triển phương pháp gợi ý Collaborative filtering bằng R	393
Ngày 10 – Chủ đề: Natural Language Processing.....	399
Bài 41: Các kỹ thuật cơ bản	400
Bài 42: Trích đặc trưng (Feature extraction).....	405
Bài 43: Giới thiệu ứng dụng phân tích cảm xúc (Sentiment Analysis) ..	415
Bài 44: Giới thiệu ứng dụng phân tích từ vựng (Word Embedding) ...	426
Bài 45: Giới thiệu ứng dụng xác định chủ đề (Topic Modeling)	438
Ngày 11 – Chủ đề: Computer Vision	449
Bài 46: Giới thiệu Face recognition	450

Đây là eBook của riêng bạn – đề nghị không chia sẻ cho ai khác nhé!

Bài 47: Giới thiệu mô hình CNN	465
Ngày 12 – Chủ đề: Nhận diện tiếng nói (Speech Recognition)	484
Bài 48: Giới thiệu đặc trưng của âm thanh	485
Bài 49: Các thao tác cơ bản với file âm thanh	491
Bài 50: Mô hình Chuyển giọng nói thành văn bản	495
Ngày 13 – Chủ đề: Phân tích dữ liệu theo trường phái Bayes	498
Bài 51: Nhập môn.....	499
Tạm kết thúc	499
Phụ lục	500
Quan sát giao dịch cổ phiếu VNM (Vinamilk)	501
Đọc và vẽ tín hiệu âm thanh.....	511
Tải sách nói “Từ tốt đến vĩ đại”	514
Đọc ảnh y khoa DiCOM.....	517
Áp dụng biến đổi Fourier cho ảnh.....	520
Sử dụng Git.....	524
Khảo sát ảnh và ma trận	553
Phát triển ứng dụng với Python.....	555
Xử lý file pdf	562
Khảo sát file âm thanh.....	570
Phân tích âm thanh với thư viện mutagen	573
Khám phá Python trong WSL2	574
Crawl dữ liệu bằng Selenium	576
Sử dụng OpenCV để phân tích dữ liệu ảnh.....	577
Cài đặt OpenCV	578
Đóng gói chương trình Python	579
Tải file video từ Youtube	582
Sinh code Restful API từ database	583
Trải nghiệm Restful API với Flask	585
Trải nghiệm Kafka.....	586
Trải nghiệm Apache NiFi.....	589
Giới thiệu superset.....	594
Cài đặt.....	595

Đây là eBook của riêng bạn – đề nghị không chia sẻ cho ai khác nhé!

Khởi động lại superset.....	600
Cài đặt và cấu hình Nginx	601
Truy cập nginx trên Ubuntu từ máy Windows.....	602
Khai thác superset	604
Cài đặt Ubuntu Server 20.04.2	608
Superset introduction.....	609
Giới thiệu PowerBI.....	609
Import data	612
Tạo biểu đồ	615
Bài bổ sung	617
Bài 101: Trải nghiệm ứng dụng với Flask	618
Bài 102: Xử lý dữ liệu	624

Quy ước

Một số nội dung trong tài liệu được trình bày với các định dạng khác nhau thì có ý nghĩa của nó, bạn đọc nên nắm thông tin này để tiện theo dõi.

Mã nguồn

Mã lệnh được viết và đóng khung với font chữ Courier New như sau:

```
print('Xin chào!')
print('Welcome!')
print('{} + {} = {}'.format(1, 2, (1 + 2)))
print('%d + %d = %d' % (1, 1, 4))
```

Bạn có thể sao chép và dán (đôi khi trong tài liệu viết luôn tiếng Anh: copy & paste) vào phần mềm để chạy.

Kết quả của lệnh, tùy theo phần mềm bạn sử dụng để chạy mã nguồn thì kết quả sẽ hiển thị ở các vị trí khác nhau. Phần văn bản kết xuất của phần mềm sẽ được trình bày theo khung màu đỏ gạch bên dưới:

```
Xin chào độc giả của ebook Chạm tới AI trong 10 ngày.
welcome to ebook Touch on AI in ten days.
1 + 1 = 4
```

Lệnh thực thi trong hệ điều hành

Trường hợp các lệnh thực thi trong môi trường hệ điều hành (phân biệt với các lệnh, hoặc mã nguồn của chương trình thực thi trong môi trường của R hoặc Python như RStudio hoặc Spyder như đã qui ước ở mục Mã nguồn) thì dấu hiệu như sau:

Đối với lệnh thực thi trong dấu nhắc lệnh của Anaconda hoặc trong cửa sổ lệnh CMD của Windows, hoặc trong Terminal của Linux/MacOS thì khung màu vàng có 2 vạch đậm ở cạnh trái và phải như sau:

```
pip install python-docx
```

Cặp dấu nháy

Các dữ liệu dạng chuỗi (string, text, char nói chung là có nghĩa giống nhau trong Python) được bao đóng trong **dấu nháy đơn** hoặc **dấu nháy đôi**. Trên bàn phím máy tính thì dấu **nháy trái** và **phải** là giống nhau. Tuy nhiên trong phần mềm soạn thảo văn bản như Microsoft Word thì gập dấu nháy đơn và đôi được thay thế bằng ‘, ’’ để tăng tính thẩm mỹ. Các dấu nháy thẩm mỹ này khác với kí tự ' và " trên bàn phím (phím bên trái phím Enter).

Đôi khi bạn copy & paste mã nguồn vào các phần mềm như Microsoft Word thì các dấu nháy có thể bị “trang trí” lại như trên. Vì vậy khi copy mã nguồn từ Microsoft vào các phần mềm chạy R hoặc Python thì hãy thay thế lại cho đúng.

Ứng dụng Phân tích dữ liệu và Trí tuệ nhân tạo với Python

Một qui ước khác liên quan đến dấu nhảy đôi là khi dùng trong văn bản để bao đóng danh từ riêng, hoặc lệnh như: *Bạn hãy thử gõ lệnh “exit()” trong cửa sổ console để thoát chương trình Python.* Trong câu hướng dẫn này thì lệnh `exit()` được gõ vào console **KHÔNG** bao gồm cặp dấu nhảy.

Cách viết thông tin lặp lại với dấu ba chấm

Khi cần mô tả một lệnh có nhiều thông tin lặp lại thì dùng dấu ba chấm như ví dụ sau.

Khi cần mô tả hàm xóa cột dữ liệu trong tham số thứ nhất của hàm `drop` như:

```
df.drop(['cột 1', 'cột 2', ...], axis =1)
```

thì phần in đậm có nghĩa là có thể gồm 1 hoặc nhiều tên cột dữ liệu. Ví dụ lệnh sau có nghĩa là xóa cột `Fullname` khỏi DataFrame `df`.

```
df.drop(['Fullname'], axis =1)
```

Hoặc lệnh sau sẽ xóa 2 cột `Fullname` và `Year` khỏi DataFrame `df`:

```
df.drop(['Fullname', 'Year'], axis =1)
```

Kí hiệu optional (không bắt buộc)

Khi sử dụng hàm số thì có nhiều tham số (argument, parameter) không bắt buộc (optional) thì sử dụng cặp dấu ngoặc vuông `[]`. Ví dụ hàm `plot` bên dưới không bắt buộc tham số `x` và `format`:

```
plot([x], y, [format])
```

Cách viết in nghiêng cho các biến

Thông thường các biến được mô tả trong các câu lệnh sẽ để trong cặp dấu ngoặc nhọn `<>`. Ví dụ lệnh sau có nghĩa là khi gõ lệnh bạn phải thay nội dung *<tên cột>* thành tên cột cụ thể trong data frame của bạn:

```
df[df['<tên cột>'].notnull()]
```

Trong tài liệu này đôi lúc sẽ không dùng cặp dấu ngoặc nhọn để mô tả lệnh chung như sau:

```
df[df['tên cột'].notnull()]
```

Cách viết trình tự bấm chọn menu

Khi cần trình bày thứ tự các nút bấm, hoặc các mục cần bấm trong các thao tác thì sẽ dùng dấu lớn hơn `>`. Ví dụ khi hướng dẫn bạn vào trang web

“<https://github.com/vncorenlp/VnCoreNLP>”, bấm vào nút “Clone”, sau đó bấm tiếp vào nút hoặc link “Download Zip” thì sẽ viết gọn như sau:

Bấm vào nút Clone > nút Download Zip, hoặc nút Clone > Download Zip.

Các từ tiếng Anh viết tắt thường xuyên được sử dụng trong sách

AI: Artificial Intelligent - **Trí thông minh nhân tạo**. Nhiều người dịch là Trí Tuệ Nhân Tạo. Trong sách này tôi muốn dùng đúng nghĩa Intelligent có nghĩa là Trí thông minh thôi vì khoảng cách từ Thông Minh đến Tuệ thì rất rất là xa. Trí thông minh nhân tạo tôi cho là phù hợp nhất trong bối cảnh hiện nay. Có thể bạn và cả tôi quen với cách đọc Trí Tuệ Nhân Tạo vừa gọn và vừa sang. Tuy nhiên nếu khi cần nói thì vẫn nên dùng từ “Thông minh” để phản ánh đúng mức độ của nó để mà còn phấn đấu đến mức “Tuệ”. Đẳng nào thì tôi cũng viết là AI thay vì viết tiếng Việt nên chắc không nhầm lẫn.

Đường dẫn thư mục (Path)

Trong Windows thì dấu cách thư mục là dấu xuyệt trái (back slash). Ví dụ: D\ai2021\data.

Tuy nhiên ngôn ngữ R hoặc Python được thiết kế tương thích với các hệ điều hành khác như Macintosh, Linux. Các hệ điều hành thì dùng dấu xuyệt phải (right slash) để phân cách thư mục. Ví dụ: /mnt/d/ai2021.

Vì vậy khi trình bày đường dẫn thư mục trong câu văn thì đôi lúc dùng \, hoặc đôi lúc dùng / do dữ liệu được minh họa trên Windows hoặc Linux.

Nhưng trong mã nguồn (R hoặc Python) thì đều thống nhất là dùng dấu xuyệt phải / như sau:

```
read.csv("D:/ai2021/data/test.csv")
```

Trong Windows, code R hoặc Python có một cách khác là dùng hai (double) dấu \. Ví dụ:

```
read.csv("D:\\ai2021\\data\\test.csv")
```

Tuy nhiên code này không tương thích trong Python trên Linux và cả MacOS nên **không** khuyến khích dùng.

Bài 3: Ngôn ngữ Python và phần mềm Anaconda

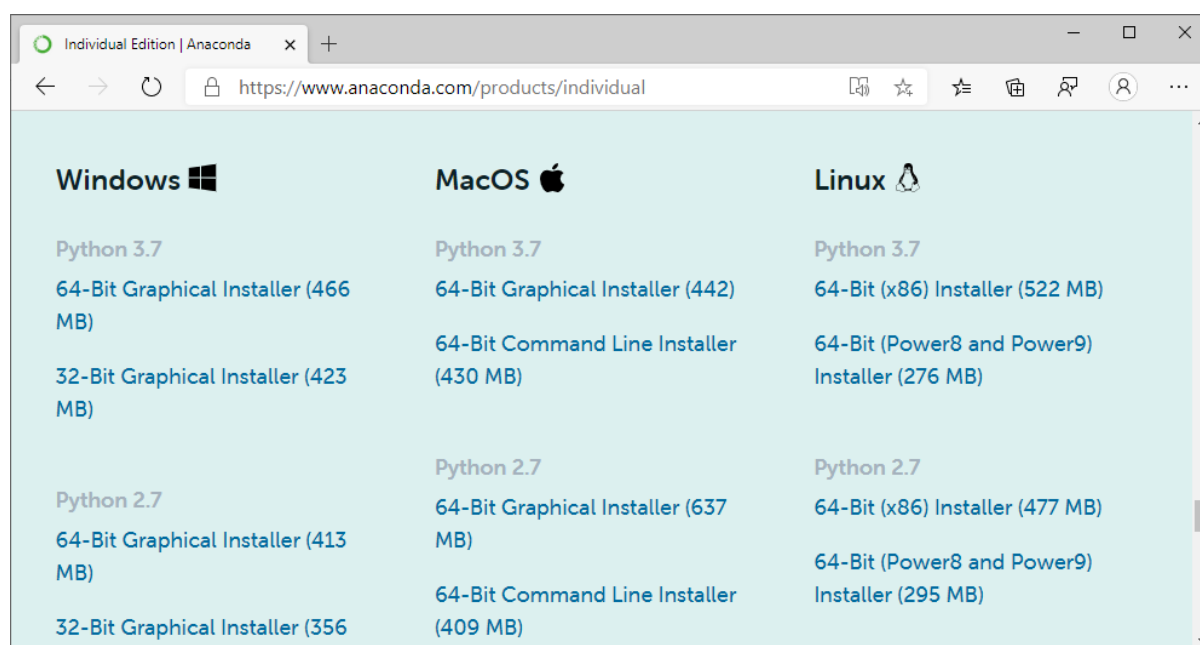
Anaconda

Đối với người mới bắt đầu làm quen với phân tích dữ liệu thì nên cài đặt phần mềm Anaconda tại địa chỉ “<https://anaconda.com>”. Anaconda là bộ quản lý các gói phần mềm (package manager). Trong đó tập trung chủ yếu các gói phần mềm về R và Python. Anaconda miễn phí, dễ sử dụng, có thể chạy được trên các hệ điều hành phổ biến như Windows, Mac, Linux.

Anaconda phù hợp với mọi người để học, thực hiện phân tích dữ liệu, Máy học (Machine learning) bằng ngôn ngữ R và Python.

Cài đặt Anaconda

Vào trang “<https://www.anaconda.com/products/individual>”, bấm vào nút Download:

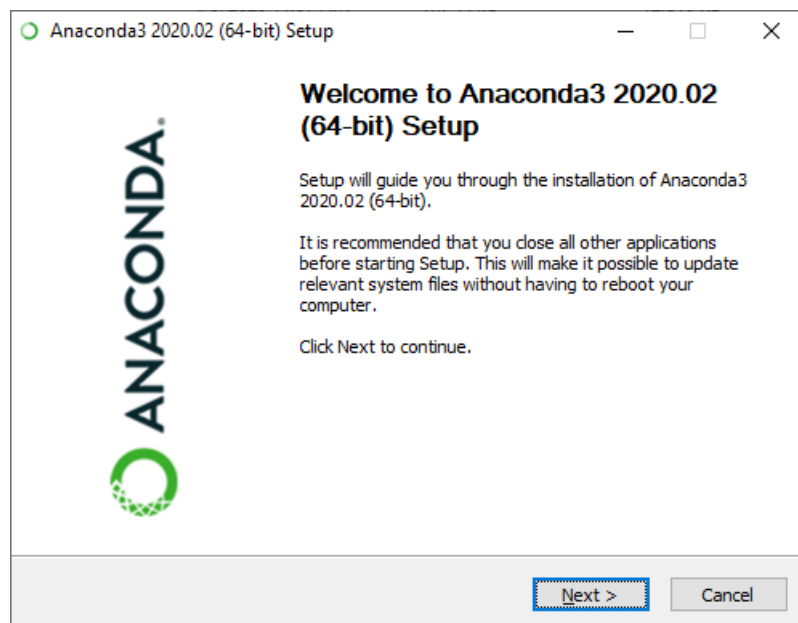
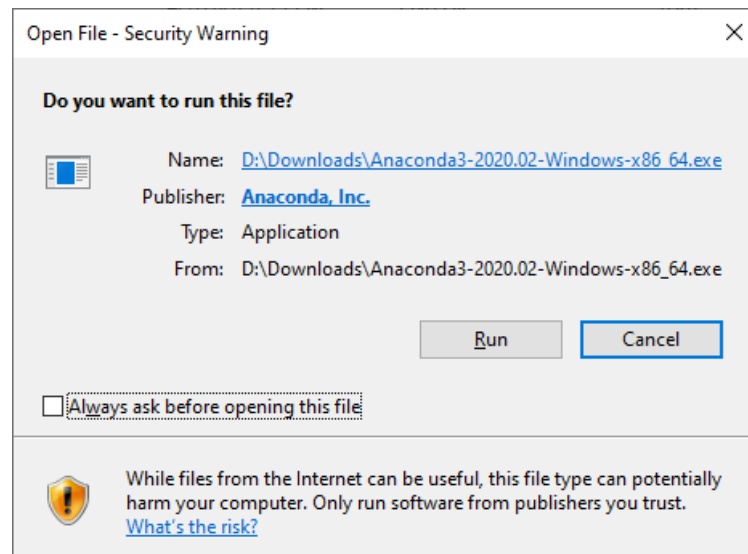


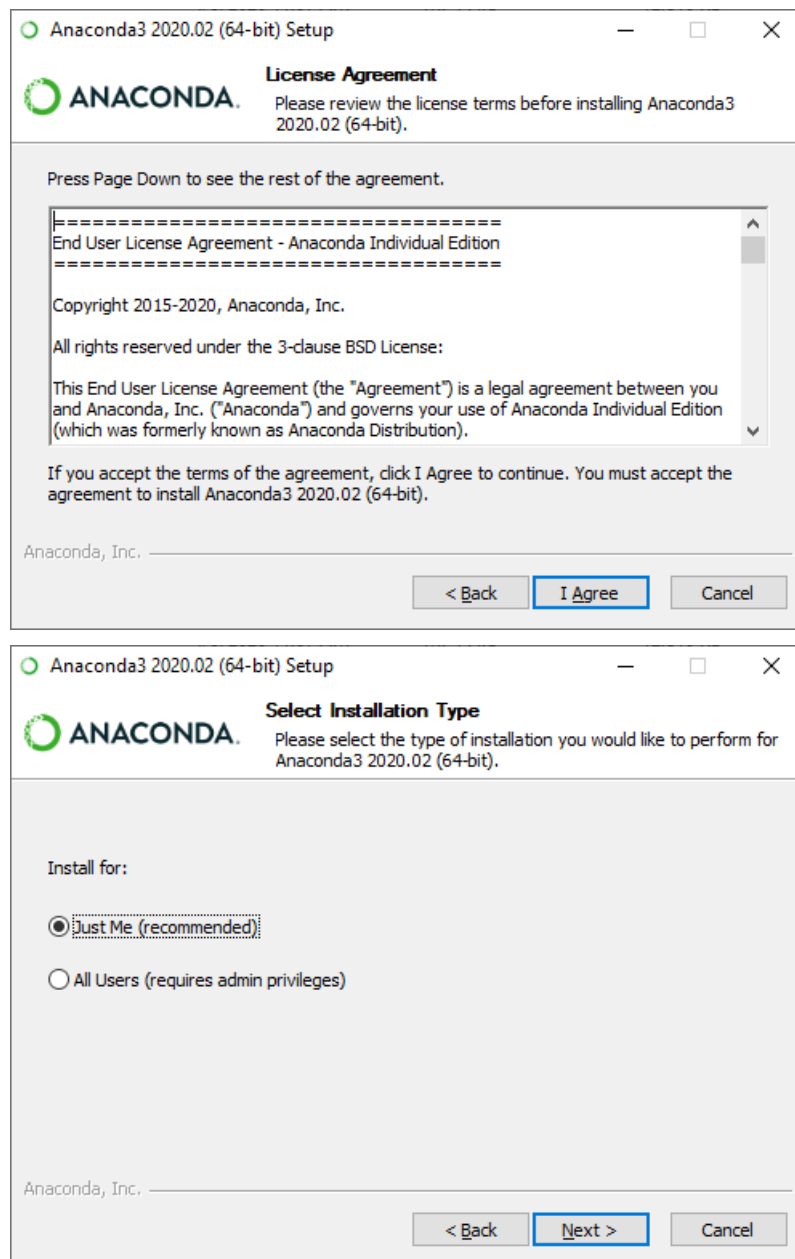
Sau đó bấm mục của gói cài đặt tùy theo máy của bạn. Tại thời điểm viết phần này thì Anaconda cung cấp 2 phiên bản phổ biến cho Python 3.7 và Python 2.7. Có nhiều sự khác biệt lớn giữa hai phiên bản Python 3 (gọi chung là 3.x) và Python 2 (gọi chung là 2.x); cũng có nhiều lý do vì sao mọi người đang dùng cả hai phiên bản. Tuy nhiên chi tiết về sự khác biệt này không nằm trong phạm vi của cuốn sách. Và không để bạn mất tập trung thì tạm thời cứ cài đặt phiên bản mới nhất để thực hành. Khi nào gặp vấn đề và cần dùng đến phiên bản cũ (Python 2.7 hoặc 2.x nói chung thì chúng ta tính sau).

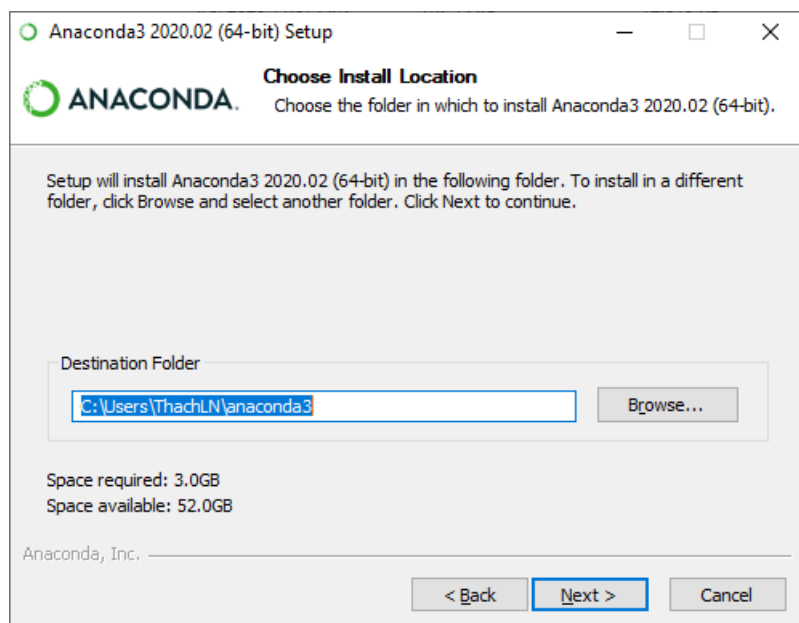
Tôi sẽ dùng phiên bản 3.7 và file tải về là “[64-Bit Graphical Installer \(466 MB\)](#)” do máy tôi dùng Windows 64 bit. Nếu máy bạn đang dùng Windows 32 bit thì tải link “[32-Bit Graphical Installer \(423 MB\)](#)”. Tương tự nếu bạn dùng MacOS hoặc Linux thì bấm vào link tương ứng phía trên màn hình.

Ứng dụng Phân tích dữ liệu và Trí tuệ nhân tạo với Python

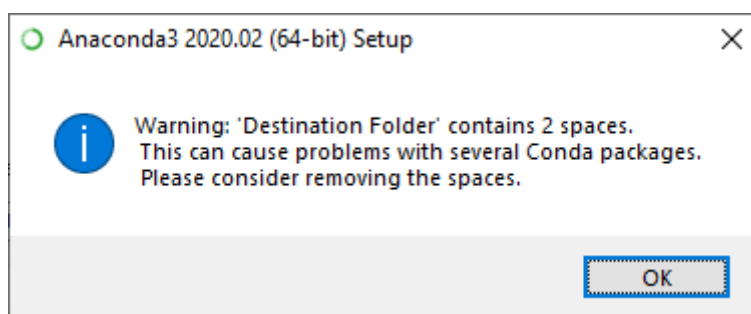
Quá trình cài đặt khá đơn giản. Cơ bản là cứ bấm “Next” và “Agree” rồi làm theo hướng dẫn.



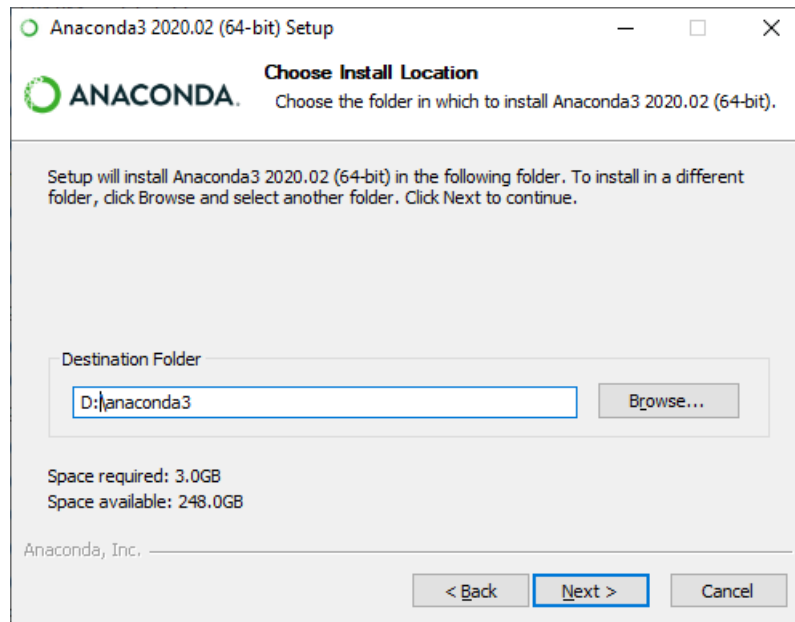




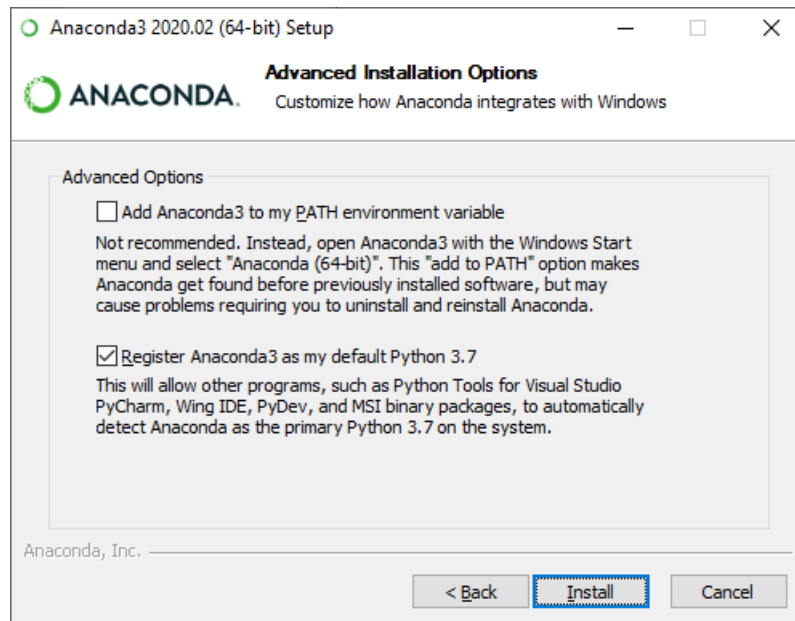
Tới bước chọn thư mục “Destination Folder” thì nếu bạn bấm “Next” mà tên thư mục của bạn có khoảng trắng (ví dụ tôi dùng tên đầy đủ để đăng nhập vào máy nên có khoảng trắng) thì sẽ bị cảnh báo như bên dưới:

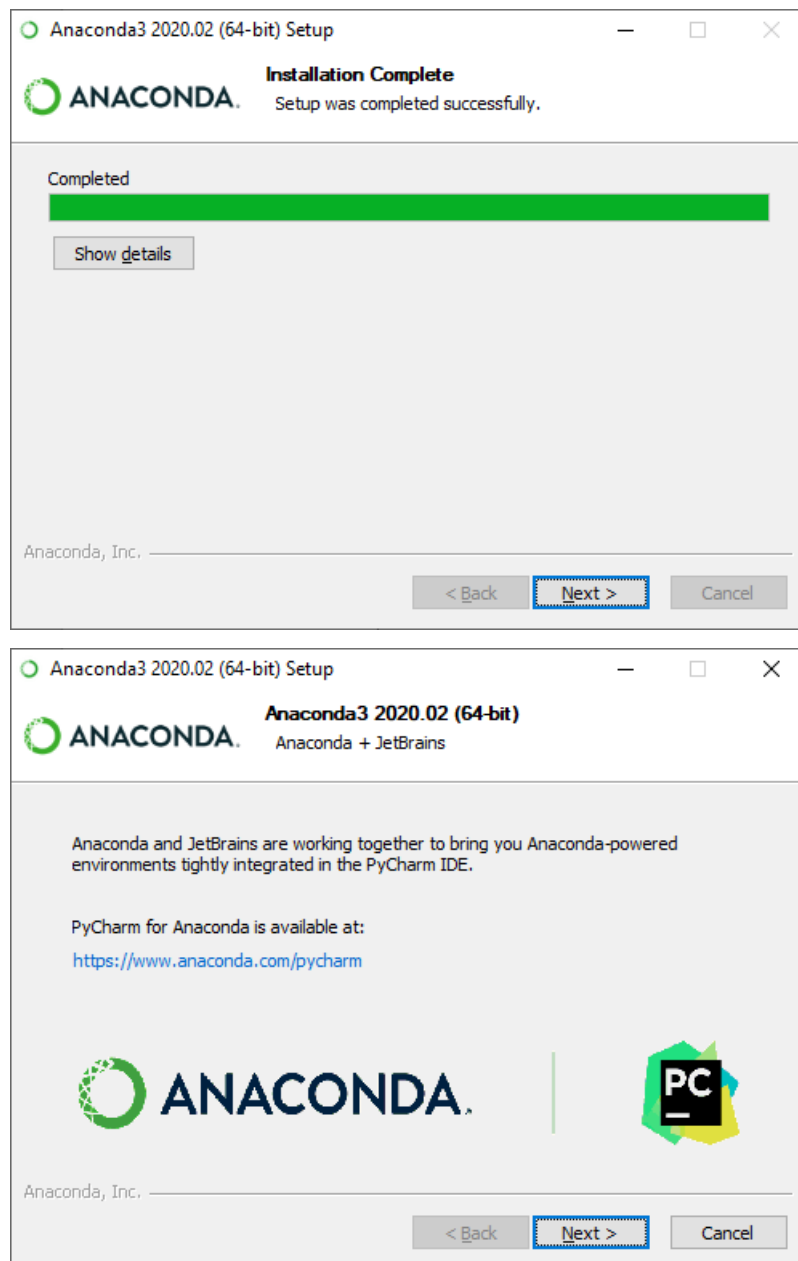


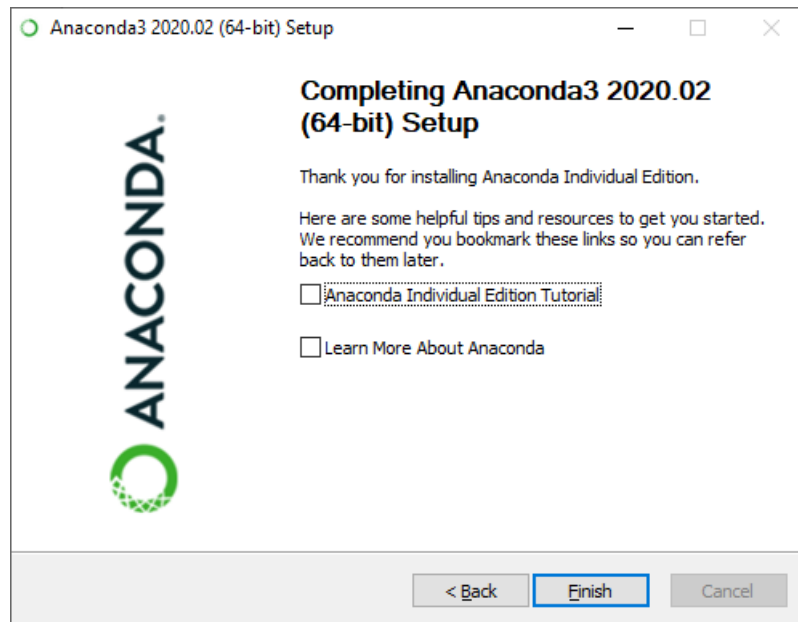
Lúc này bạn nên bấm OK rồi bấm tiếp “Back” trên màn hình tiếp theo để quay lại bước chọn “Destination Folder”. Ví dụ tôi chọn lại thư mục “D:\anaconda3” không có khoảng trắng và để dễ quản lý.





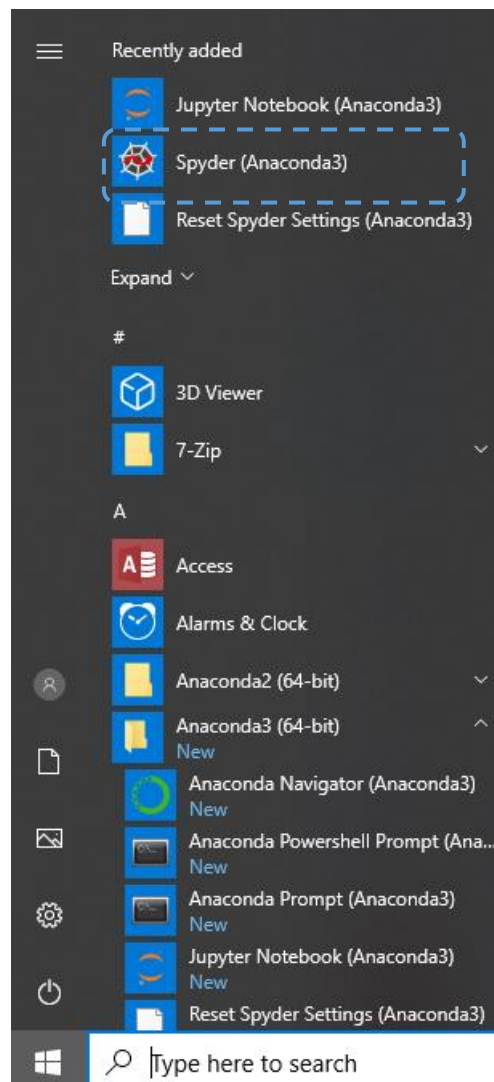
Rồi cứ thế bấm “Next”, “Install” và “Next” cho đến khi “Finish” là xong.







Sau khi cài xong, vào nút Start của Windows ở góc trái dưới màn hình hoặc bấm phím có hình cửa sổ  hoặc  (tùy bàn phím) bạn sẽ thấy biểu chương trình Spyder (Anaconda 3) như sau:



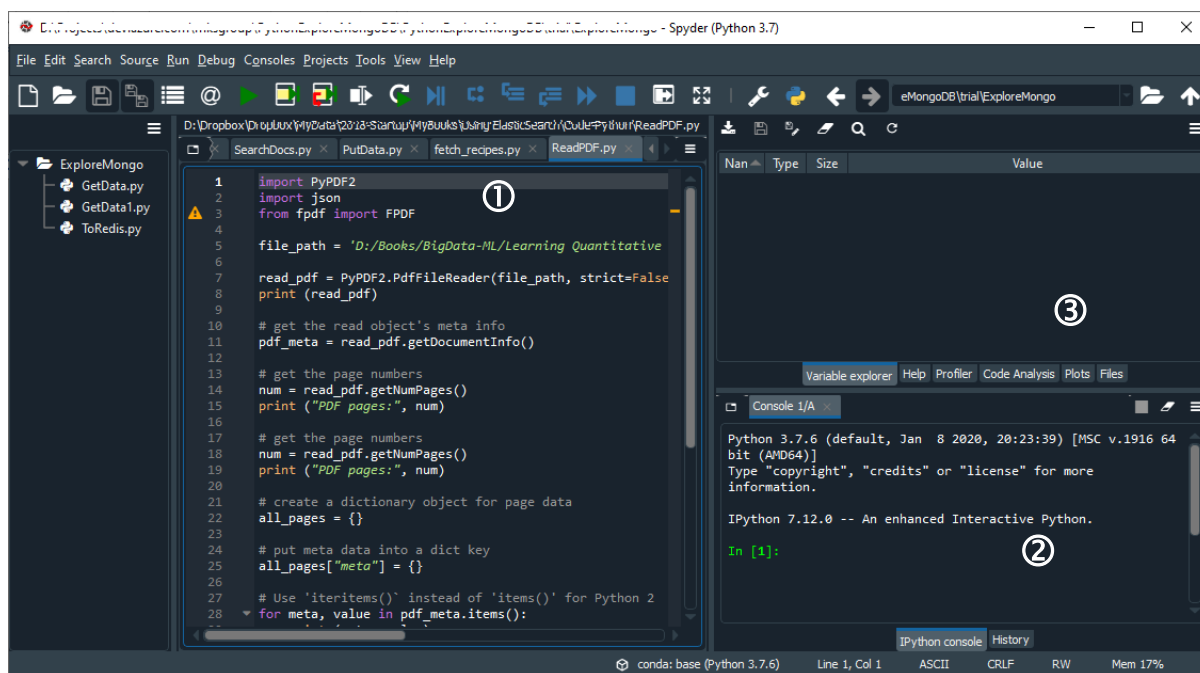
Đến đây bạn đã biết cách tải và cài đặt Anaconda Python 3. Bạn cũng nên thử khởi động Spyder (Anaconda3) và thoát nó, tắt máy tính đi uống một ly café hoặc trà sữa tùy theo sở thích để tự thưởng cho mình.

Ngôn ngữ lập trình Python

Bạn có thể bắt đầu làm quen với ngôn ngữ Python và thực hành với các gói phần mềm trong bộ Anaconda đã cài đặt trong phần trước.

Sử dụng Spyder

Sau khi cài đặt Anaconda Python, hãy khởi động chương trình Spyder sẽ có giao diện như sau:



Spyder là phần mềm để viết mã lệnh Python được thiết kế bởi các nhà khoa học (scientists), các kỹ sư công nghệ (engineers) và các nhà phân tích dữ liệu (data analysts).

① Phần cửa sổ bên trái giúp bạn viết lệnh Python. Các lệnh này sẽ được lưu vào một file tạm trên máy tính của bạn (ví dụ thư mục trên máy tôi là “C:\Users\Le Ngoc Thach”). Tên file untitled0.py có nghĩa là file chưa được đặt tên (untitled) đầu tiên (có thứ tự bắt đầu là 0), phần mở rộng sau dấu chấm là “py” - viết tắt của chữ Python.

② Phần cửa sổ “Console” ở góc phải dưới là nơi trình bày kết quả của lệnh khi các lệnh được thực thi (execute).

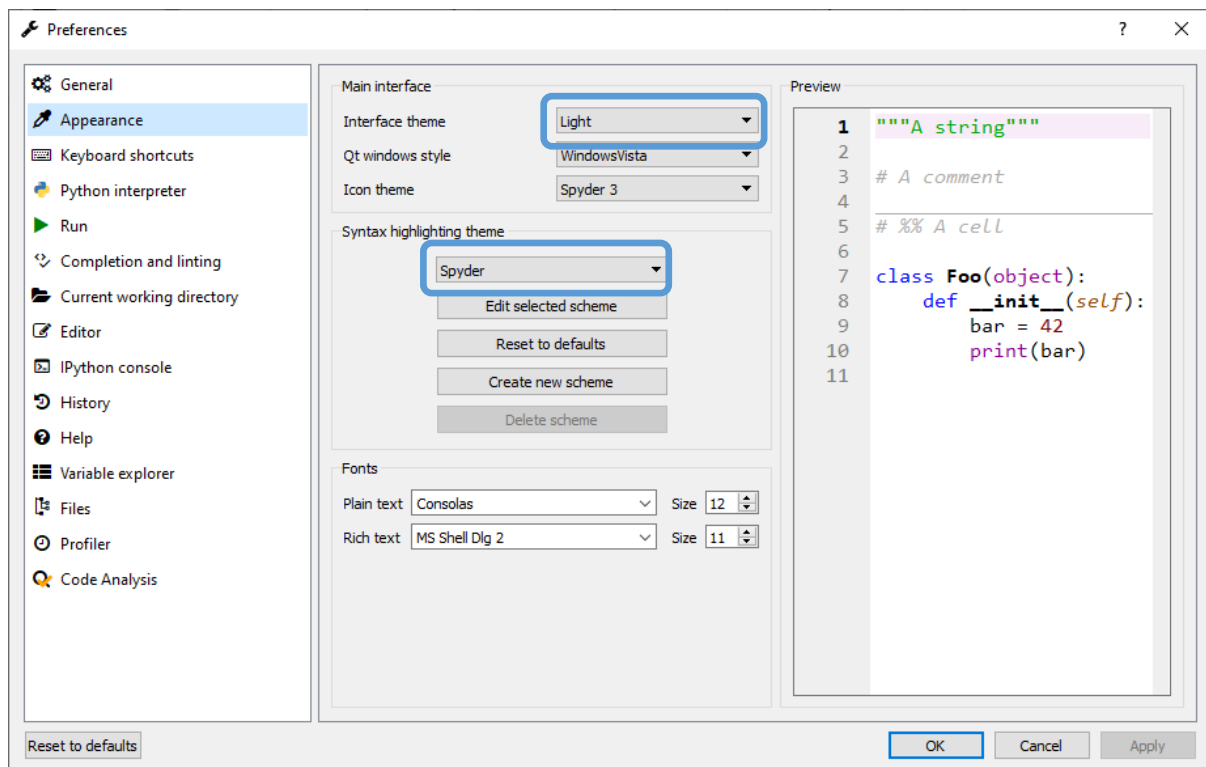
③ Phần cửa sổ ở góc phải trên có nhiều tab, trong đó 2 tab “Variable explorer” và “Plots”. Variable explorer giúp bạn theo dõi các biến mà bạn đã khai báo (declare) trong cửa sổ lệnh bên trái khi các lệnh được thực thi. Plots giúp bạn xem kết quả vẽ biểu đồ.

Đổi theme

Mặc định thì Spyder phiên bản 4.x có giao diện đen xì như trên. Nếu bạn không quen thì đổi sang giao diện sáng (light) bằng cách vào menu Tools > Preference, chọn lại:

Interface theme: **Light**

Syntax highlighting theme, mục đầu tiên: **Spyder**

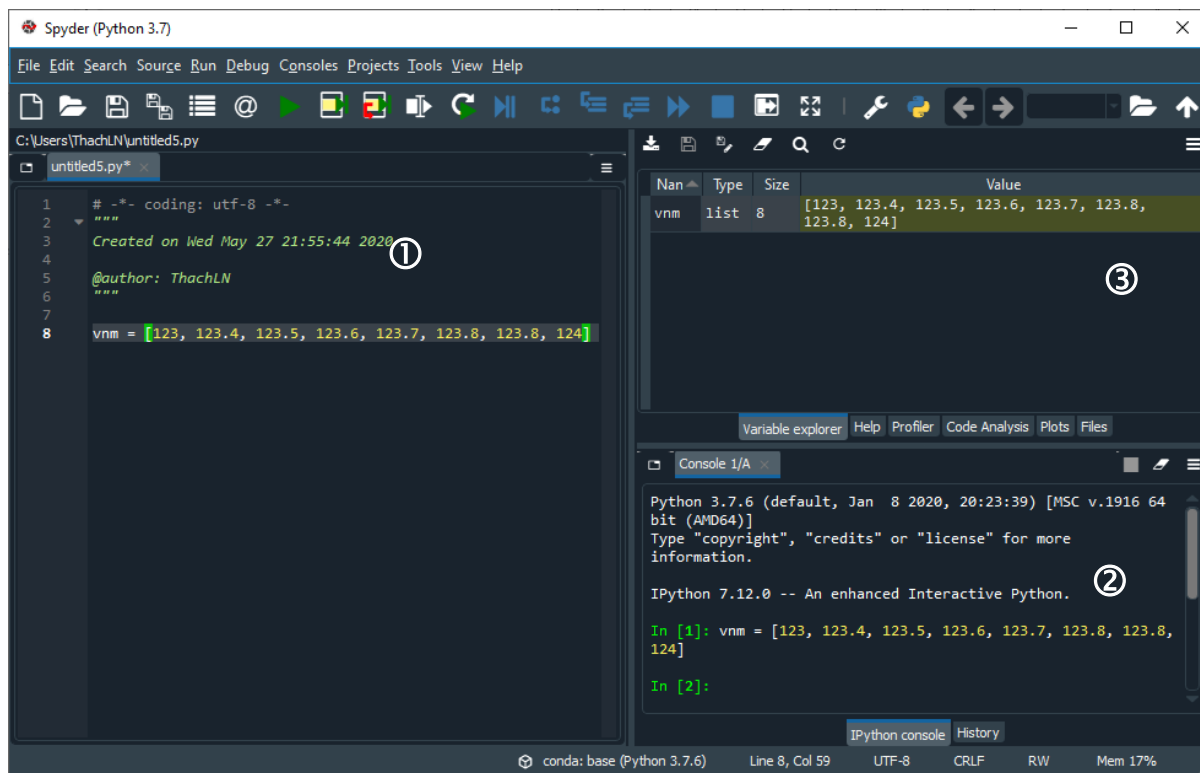


Thực thi lệnh

Chọn dòng lệnh cần thực thi, nhấn phím F9. Spyder phiên bản cũ hơn có thể nhấn phím Ctrl + Enter.

Ví dụ trong hình bên dưới khai báo một biến có tên **vnm** được gán (assign) bằng một mảng (array) gồm nhiều giá trị cách nhau bởi dấu phẩy. Cặp dấu móc vuông [] bao đóng array theo qui ước của Python.

```
vnm = [123, 123.4, 123.5, 123.6, 123.7, 123.8, 123.8, 124]
```



Trong cửa sổ bạn bôi dòng lệnh số 8 bằng các cách sau:

- 1) Dùng chuột bôi từ đầu đến cuối lệnh bằng cách di chuyển con trỏ chuột đến trước biến `vnm`, bấm nút trái chuột giữ nguyên nút trái trong lúc di chuyển con chuột sang phải dòng lệnh – hướng di chuyển chuột theo hàng ngang đảm bảo con trỏ chuột lúc nào cũng nằm trên dòng lệnh. Khi con trỏ chuột đến cuối dòng lệnh bạn sẽ thấy dòng lệnh sẽ được bôi màu nền xanh như hình trên.
- 2) Dùng phím Shift + Home: khi gõ lệnh xong thì con nháy đang ở cuối dòng lệnh. Bạn chỉ cần nhấn tổ hợp phím Shift + Home (tay trái nhấn và giữ nút Shift, sau đó tay phải nhấn phím Home rồi thả cả 2 tay ra khỏi bàn phím cùng lúc).
- 3) Dùng phím Shift + End: khi con nháy đang ở bất kỳ chỗ nào trên dòng lệnh, hãy gõ phím Home để đưa con nháy về vị trí đầu tiên. Sau đó nhấn tổ hợp phím Shift + End (tay trái nhấn và giữ nút Shift, sau đó tay phải nhấn phím End rồi thả cả 2 tay ra khỏi bàn phím cùng lúc).

Thực hành phép gán

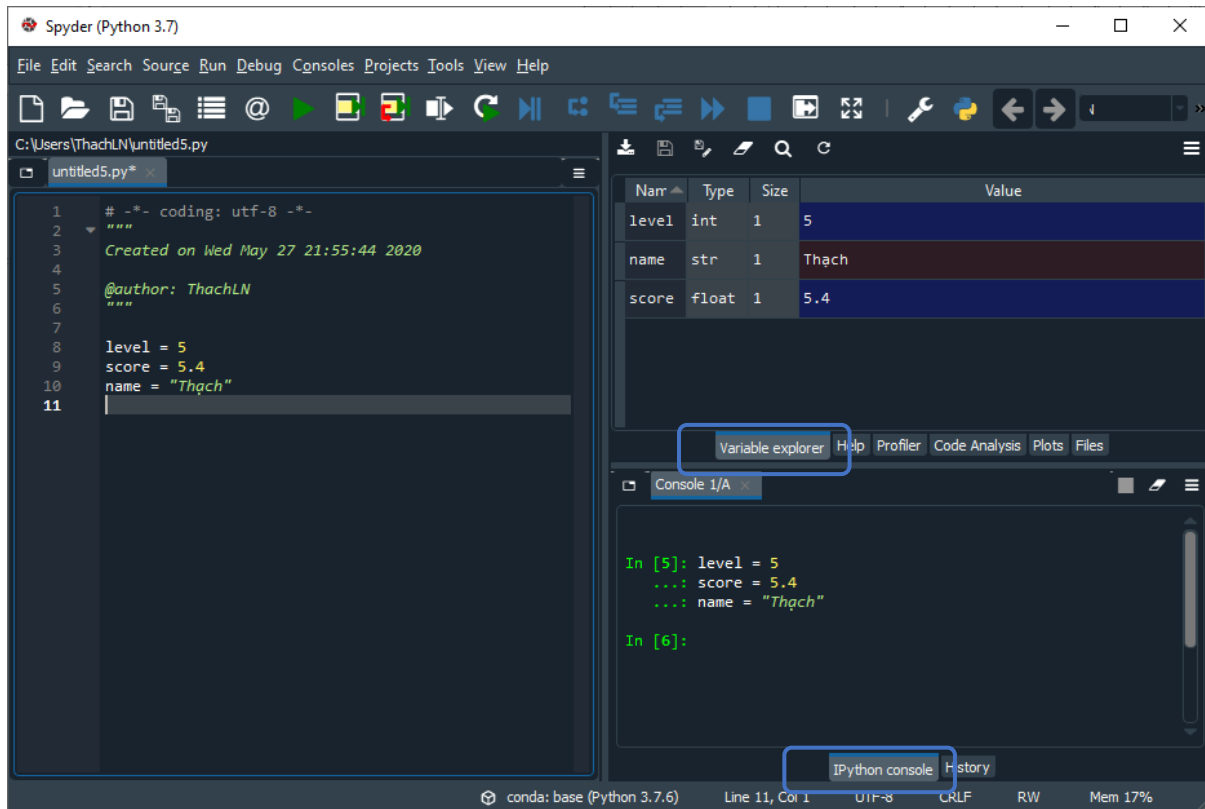
Hãy khởi động chương trình Spyder, mở file mới bằng cách nhấn Ctrl + N. Sau đó gõ 3 lệnh sau:

```
level = 5
```

Ứng dụng Phân tích dữ liệu và Trí tuệ nhân tạo với Python

```
score = 5.4  
name = "Thạch"
```

Thực thi 3 dòng lệnh bằng cách bôi cả 3 dòng rồi nhấn phím **F9**. Quan sát giá trị các biến trong thẻ “Variable explorer” và quan sát các lệnh được thực thi trong cửa sổ “Console” ở góc phải dưới.



Cài đặt các gói phần mềm

Tương tự như R, Python cũng cung cấp rất nhiều gói thư viện.

```
import <tên thư viện> as <tên viết tắt>
```

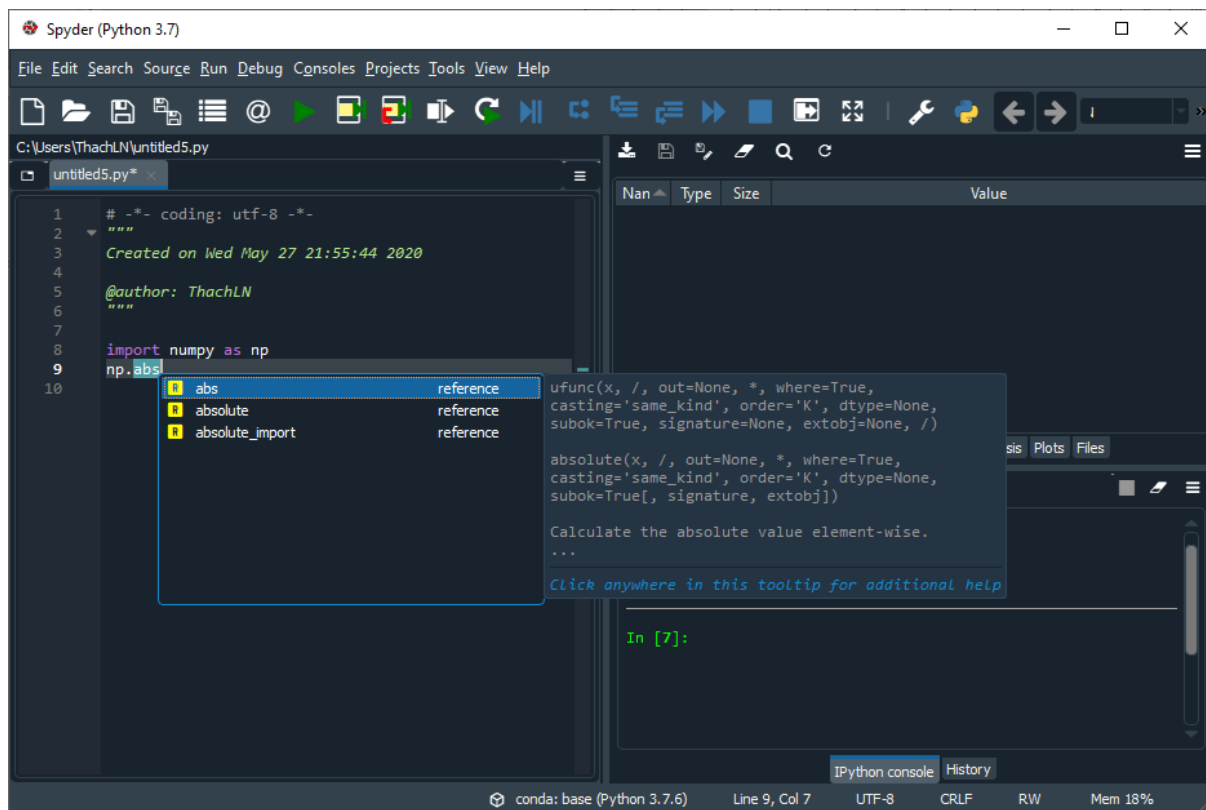
Ví dụ để sử dụng thư viện **numpy** thì sử dụng lệnh

```
import numpy as np
```

Tên viết tắt là do bạn quy định để thuận tiện khi viết lệnh. Dùng tên viết tắt này để cho mã nguồn gọn hơn.

Trong chương trình Spyder khi gõ lệnh **np**, sau đó nhấn **Ctrl + Space** thì bạn sẽ thấy các hàm của numpy hiển thị ra cho bạn để chọn hoặc để gõ tiếp.

Ứng dụng Phân tích dữ liệu và Trí tuệ nhân tạo với Python



Gọi hàm

Tương tự như minh họa gọi hàm trong R, phần này cũng sẽ giới thiệu lại ví dụ về điểm số để bạn làm quen trong Python.

Trong R thì các hàm để tính toán các khái niệm thống kê cơ bản đã có sẵn, bạn chỉ cần gõ lệnh là thực thi được.

Tuy nhiên, trong Python thì một số hàm được cung cấp trong thư viện **NumPy**. Vì vậy bạn cần phải thực thi lệnh `import` như sau để bắt đầu sử dụng NumPy:

```
import numpy as np
```

Dùng cú pháp `[]` để khai báo danh sách điểm. Sau đó gán danh sách cho biến `scores` như sau:

```
scores = [6, 7, 9, 4, 5, 7, 8, 6, 5, 7]
```

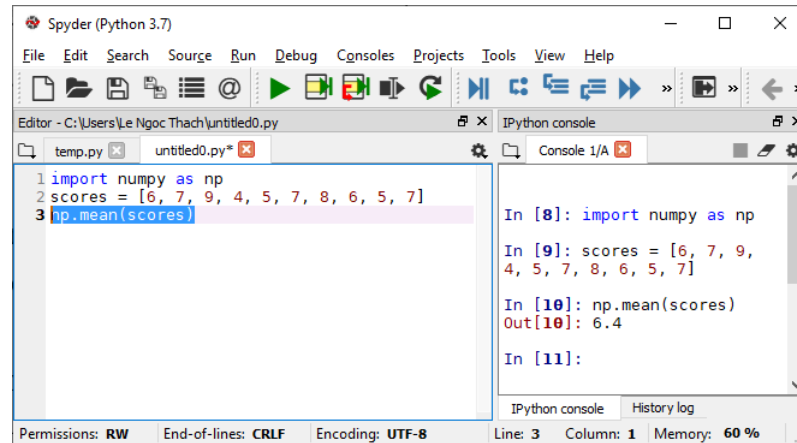
- Gọi hàm `mean` của thư viện `numpy` thông qua kí hiệu `np`:

```
np.mean(scores)
```

sẽ cho kết quả: 6.4

Ứng dụng Phân tích dữ liệu và Trí tuệ nhân tạo với Python

Cần nhắc lại một chút là trong lúc soạn thảo lệnh trong Spyder, để thực thi từng dòng lệnh thì bôi chọn từng dòng, nhấn Ctrl + Enter hoặc nhấn F9; trường hợp không bôi đoạn lệnh nào thì F9 sẽ thực thi dòng đang có con nháy. Sau đó theo dõi kết quả trong cửa sổ Console.



- Gọi hàm `np.median(x)`:

```
np.median(scores)
```

sẽ cho kết quả: 6.5

- Gọi hàm `np.sort(scores)`:

```
np.sort(scores)
```

sẽ cho kết quả: `array([4, 5, 5, 6, 6, 7, 7, 7, 8, 9])`

Phần mềm Spyder in ra kết quả có chữ `array()` và cặp dấu ngoặc `[]` để cho chúng ta biết đây là mảng.

- Gọi hàm `np.var(x)`:

```
np.var(scores, ddof = 1)
```

sẽ cho kết quả: 2.2666666666666666

Bạn sẽ thắc mắc là trong Python để tính phương sai thì gọi hàm `var` của thư viện NumPy phải có tham số "`ddof = 1`". `ddof` viết tắt của Delta Degrees of Freedom. Kết quả Python cũng hiển thị số lượng kí số phần thập phân cũng khác với R. Delta Degrees of Freedom là gì thì tạm thời lúc này hãy quên nó đi nhé. Chúng ta đang tập làm quen với việc gọi hàm trong Python. Chúng ta sẽ quay lại khái niệm này sau.

- Gọi hàm `np.std(x)`:

```
np.std(scores, ddof = 1)
```

Ứng dụng Phân tích dữ liệu và Trí tuệ nhân tạo với Python

sẽ cho kết quả: 1.505545305418162

- Gọi hàm `np.quantile(a, q)` để tính Bách phân vị:

```
np.quantile(scores, 0.5)
np.quantile(scores, 0.25)
np.quantile(scores, 0.75)
```

sẽ cho kết quả tương ứng của Bách phân vị 50%, 25%, 75%: 6.5, 5.25, 7.0

Bài tập thực hành

① *Làm quen với Biến và Phép gán.*

Trong bài 2 tôi có giới thiệu khái niệm Biến và Phép gán.

Đây là thời điểm để bạn mở Spyder thực hành các lệnh sau:

Python

```
import datetime
```

```
fullName = 'Lê Ngọc Thạch'
```

```
height = 165
```

```
weight = 72.5
```

```
sex = True
```

```
birthday = datetime.datetime.strptime('30/6/2019', '%d/%m/%Y')
```

```
favorNumbers = [1, 2, 5, 10, 20, 50]
```

```
favorSports = ['Bóng bàn', 'Bóng đá ', 'Quần vợt']
```

```
# Thử xem giá trị của vài biến
```

```
birthday
```

```
favorNumbers
```

```
# Xem phần tử đầu tiên của favorNumbers
```

```
favorNumbers[0]
```

```
# Đếm số phần tử của biến favorNumbers
```

```
len(favorNumbers)
```


Lấy ra phần tử cuối cùng của biến favorNumbers

`favorNumbers[len(favorNumbers) - 1]`

Bạn nên copy từng lệnh hoặc tốt nhất là tự gõ vào Spyder để chạy và quan sát.

Sau mỗi lệnh bạn nên gõ lệnh `type` để biết thêm kiểu dữ liệu của biến:

```
type(<tên biến>)
```

Ví dụ:

```
type(fullName)
```

Cho kết quả là: `str`

str có nghĩa là String (chuỗi)

🔗 Làm quen hàm thống kê

Khảo sát đoạn chương trình sau:

```
import pandas as pd
a = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]
s_a = pd.Series(a)
s_a.describe()
```

Kết quả như sau:

```
count    15.000000
mean      8.000000
std       4.472136
min       1.000000
25%       4.500000
50%       8.000000
75%      11.500000
max      15.000000
dtype: float64
```

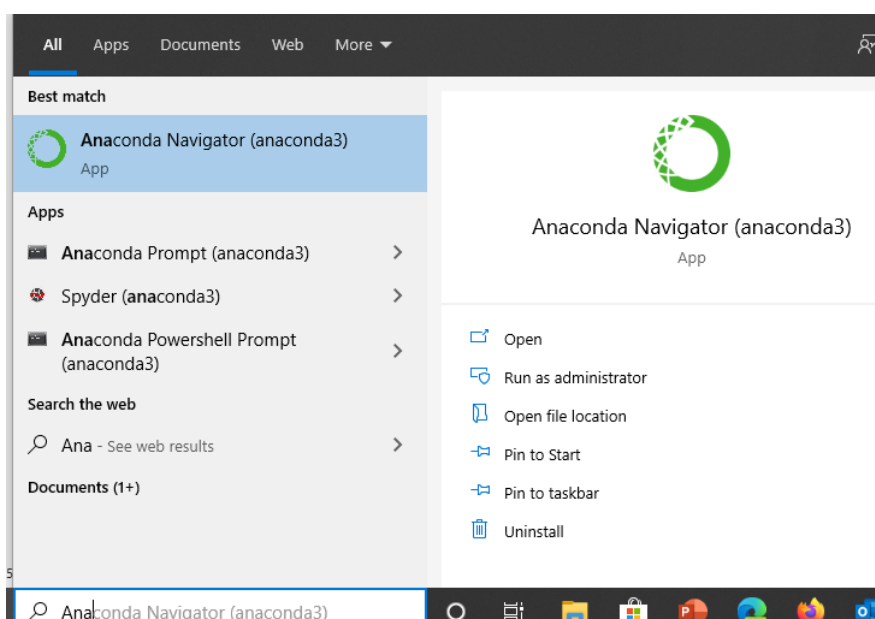
Diễn giải:

- Lệnh đầu tiên là khai báo sử dụng thư viện pandas với định danh là `pd`.
- Lệnh thứ hai khai báo một dãy số gồm 15 phần tử, mỗi phần tử có giá trị tương ứng từ 1 đến 15.
- Lệnh thứ ba chuyển dãy `a` thành kiểu dữ liệu gọi là Series – là một cột dữ liệu trong bảng dữ liệu (gọi là Data Frame). Kết quả lưu vào biến `s_a`.
- Sử dụng hàm `.describe()` của cột dữ liệu (Series) `s_a`.

- Kết quả hàm `describe()` sẽ cung cấp vài thông tin thống kê để mô tả về biến `s_a`. Cụ thể gồm:
 - `count`: tổng số phần tử.
 - `mean`: giá trị trung bình của các phần tử.
 - `std`: Độ lệch chuẩn (Xem lại mô tả khái niệm [Độ lệch chuẩn](#))
 - `min, max`: Giá trị nhỏ nhất, Giá trị lớn nhất.
 - `25%`: Giá trị bách phân vị 25%. Giá trị mà tại đó chia tập dữ liệu thành 2 phần $\frac{1}{4}$ và $\frac{3}{4}$.
 - `50%`: Giá trị bách phân vị 50%. Giá trị mà tại đó chia tập dữ liệu thành 2 phần bằng nhau $\frac{1}{2}$ và $\frac{1}{2}$.
 - `75%`: Giá trị bách phân vị 75%. Giá trị mà tại đó chia tập dữ liệu thành 2 phần $\frac{3}{4}$ (Phần các giá trị nhỏ) và $\frac{1}{4}$ (Phần các giá trị lớn..

Cài đặt thư viện

Một trong các lý do mà ngôn ngữ Python phổ biến nhất tại thời điểm eBook được viết trong lĩnh vực Machine Learning và AI là cộng đồng phát triển rất lớn. Trong đó có rất nhiều thư viện được cung cấp miễn phí. Trong Windows, để cài đặt thư viện Python thì mở cửa sổ của Anaconda Prompt hoặc Anaconda Powershell Prompt bằng cách bấm vào nút Windows Start, gõ chữ Ana thì ra màn hình bên dưới, sau đó bấm vào biểu tượng tương ứng (ví dụ Anaconda Prompt).

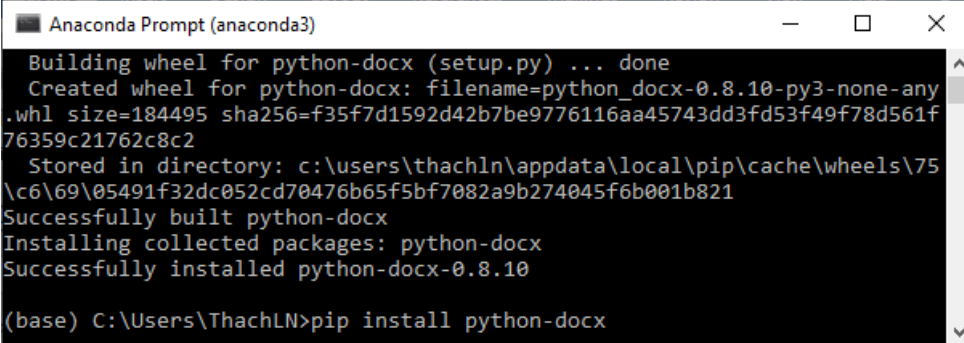


Trong cửa sổ Anaconda Prompt gõ lệnh:

```
pip install <tên thư viện>
```

Ví dụ cài thư viện python-docx để xử lý file .docx của Microsoft Word:

```
pip install python-docx
```

A screenshot of the Anaconda Prompt terminal window. The window title is "Anaconda Prompt (anaconda3)". The terminal output shows the process of building a wheel for python-docx, creating a wheel file, storing it in a cache directory, and successfully installing it. The prompt at the bottom shows the command being executed: (base) C:\Users\ThachLN>pip install python-docx.

```
Anaconda Prompt (anaconda3)
Building wheel for python-docx (setup.py) ... done
Created wheel for python-docx: filename=python_docx-0.8.10-py3-none-any
.whl size=184495 sha256=f35f7d1592d42b7be9776116aa45743dd3fd53f49f78d561f
76359c21762c8c2
Stored in directory: c:\users\thachln\appdata\local\pip\cache\wheels\75
\c6\69\05491f32dc052cd70476b65f5bf7082a9b274045f6b001b821
Successfully built python-docx
Installing collected packages: python-docx
Successfully installed python-docx-0.8.10

(base) C:\Users\ThachLN>pip install python-docx
```