Pythonで文書処理

資料のPDF化、文字認識、検索、その他さまざまな作業をプログラミングで解決

北山洋幸◉著

CUTT
カットシステム

■**サンプルファイルのダウンロードについて**

本書掲載のサンプルファイルは、下記 URL からダウンロードできます。

https://----------

・本書の内容についてのご意見、ご質問は、お名前、ご連絡先を明記のうえ、小社出版部宛文書（郵送または E-mail）でお送りください。
・電話によるお問い合わせはお受けできません。
・本書の解説範囲を越える内容のご質問や、本書の内容と無関係なご質問にはお答えできません。
・匿名のフリーメールアドレスからのお問い合わせには返信しかねます。

はじめに

本書は Python で PDF や画像の文字認識などを扱う入門書です。プラットフォームや開発環境へ依存しないように心がけましたが、複数のプラットフォームへ対応させると煩雑になるため、Windows で開発と実行テストを行いました。

Python とは、プログラミング言語の一種です。Python はプログラミングに不慣れな人であっても習得しやすい言語です。習得しやすいから非力かというと、そうではなく初級者なら初級者的な使い方が、そして上級者には上級者らしい使い方のできる万能なプログラミング言語と言って良いでしょう。Python には文字認識、PDF、そして画像を扱う多数のモジュールが用意されています。

本書は、Python で PDF を画像へ変換、その逆、文字認識、PDF への透かし追加、PDF の保護解除、PDF のサイズ変更や結合、カメラで撮影した資料を透視投影し PDF 化、大量の PDF へ grep 処理、複数の PDF を 1 つの PDF へ結合、逆に PDF を 1 ページ単位へ分解、PDF へパスワード設定や解除などを行うプログラムを紹介し、Python の理解を深めます。同じ処理を、異なったモジュールで実現する方法も解説します。

Python は、短くて簡潔なコードで、高度な処理を行うことができる言語です。基本的な知識を身に付ければ、高度な応用プログラムを開発することが期待できます。

本書の対象読者は、

- Python 初級者
- PDF を Python で操作したい人
- 文字認識に興味ある人

などです。

微力ながら、本書が Python を理解するきっかけになることを期待します。

2021 年 8 月　新型コロナ対応のため自粛中の自宅にて
北山洋幸

謝辞

出版にあたり、お世話になった株式会社カットシステムの石塚勝敏氏に深く感謝いたします。

環境

いくつかの環境で開発・実行しましたが、すべてのチェックを行ったのは下記の構成です。

OS	Windows 10 Pro	
Anaconda	2020.02 版	
Python	3.7.6	
Spyder	4.0.1	
Visual Studio Code	1.59.0	
Tesseract OCR	v5.0.0-alpha.20201127	
PyOCR	0.8	conda list で確認
Numpy	1.18.1	conda list で確認
pdf2image	1.14.0	conda list で確認
poppler	0.68.0	
img2pdf	0.4.0	conda list で確認
opencv-python	4.5.1.48	conda list で確認
pymupdf	1.18.15	conda list で確認
pypdf2	1.26.0	conda list で確認

すべての実行を確認したのは、Anaconda Prompt 上です。Spyder と Visual Studio Code は、一部のプログラムしか確認していません。

Spyder	4.0.1
Visual Studio Code	1.59.0

URL

書籍中に記述されている URL は原稿執筆時点のものです。URL の変更や Web サイトの構造は頻繁に変更されますので、記述した URL が存在するとは限りません。もし、ページなどが見つからない場合は、トップページへ移動して探すか、インターネットでキーワードを検索してください。記述した URL に必ず紹介した内容が記載されていることを保証するものではありません。

■ 参考文献（Web サイト）

Anaconda	https://www.anaconda.com/
Python	https://www.python.org/
Python documentation	https://docs.python.org/3/
The Python Tutorial	https://docs.python.org/3/tutorial/
pathlib reference	https://docs.python.org/ja/3/library/pathlib.html#
Unicode 入門 Unicode HOWTO	https://docs.python.org/ja/3/howto/unicode.html#
Pillow	https://pillow.readthedocs.io/en/stable/#
numpy reference	https://docs.scipy.org/doc/numpy/reference/
spyder	https://www.spyder-ide.org/
Tesseract	https://github.com/UB-Mannheim/tesseract/wiki
PyOCR	https://pypi.org/project/pyocr/
pdf2image	https://pypi.org/project/pdf2image/
img2pdf	https://pypi.org/project/img2pdf/
Poppler for Windows	http://blog.alivate.com.au/poppler-windows/
PyMuPDF	https://pypi.org/project/PyMuPDF/
MuPDF	https://en.wikipedia.org/wiki/MuPDF
PyMuPDF Tutorial	https://pymupdf.readthedocs.io/en/latest/tutorial.html
PyMuPDF Document	https://pymupdf.readthedocs.io/en/latest/document.html
PyPDF2 Documentation	https://pythonhosted.org/PyPDF2/

目次

第1章

開発環境の準備（Anaconda 編）

1

本章では、環境の設定、簡単な開発環境の使用法、そして環境が正しいことを確かめるプログラムの作成について解説します。Python はマルチプラットフォームで使用可能です。Python で開発したプログラム自体はポータビリティがありますが、Python 対応のプログラムを開発・実行させる環境は、それぞれが自身でセットアップしなければなりません。

Python のインストールは簡単になりました。あるいはプラットフォームによっては、Python は最初からインストールされています。ただ、各種モジュールなどは別途インストールが必要なものもあります。開発環境のバージョンは、日々変化するためインストールの詳細を説明しても価値があるとは思えません。しかし、何も解説しないのも初心者には不親切です。そこで、本書では Windows に環境を構築する方法を解説します。

なお、Python、各モジュール、オペレーティングシステムの組み合わせやバージョンの違いなどで、正常に環境をインストールできない、あるいは正常にプログラムが動作しないこともあります。そのような場合は、動作を確認できているバージョンに合わせる、あるいは公式サイトなどに掲載されている情報や、フォーラムなどを覗き解決法がないか調べるとよいでしょう。利用したモジュールには、比較的小規模でソースコードを簡単に入手できるものもありますので、正常に動作しない場合は自身で修正を行ってみるのもよいでしょう。筆者もいくつかのモジュールはドキュメントが不十分で理解が困難であったため、ソースコードを参照しました。基本的に、オープンソースで提供されているものは、自身で解決する努力は必要です。

1.1 Anaconda

Python 本体のみを使うには、いくつかの環境が提供されているため、それほど難しくありません。例えば Python の本家（https://www.python.org/）から Python のみをインストールするだけです。ここでは、IDE やコンソールなどもパッケージされて広く使われている Anaconda を、Windows にインストールする手順について紹介します。

開発環境を、Visual Studio Code をメインに使用したい人は、Anaconda をインストールせず、Visual Studio Code のみをインストールするのもよいでしょう。Visual Studio Code については後述します。

1.1.1 ダウンロード

Anaconda の公式ホームページからダウンロードします。まず、Anaconda のダウンロードページにアクセスします（https://www.anaconda.com/download/）。表示されたページの中ほどに「User interface makes learning easier」と表示され、その下の方の「Install Anaconda」ボタンをクリックします。ダウンロードページの先頭には表示されない場合もありますので、「Install Anaconda」ボタンが見えるまでスクロールしてください。

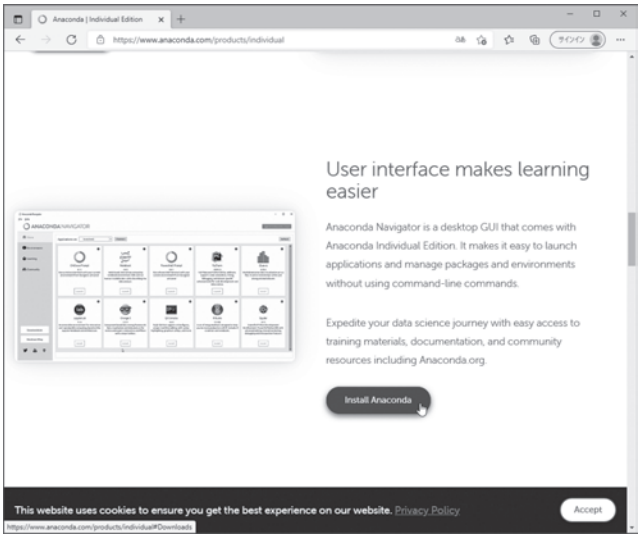

図1.1●Anacondaのダウンロードページ

「Anaconda Installers」が現れますので、「Windows」の「Python 3.x」の「64-Bit Graphical Installer」をクリックします。バージョン番号は、アップデートされますので、時期によって変化します。

図1.2●「Windows」の「Python 3.x」を選択

古いバージョンなどを利用したい場合は、先ほどの画面の下部に「The archive has older versions of Anaconda Individual Edition installers. The Miniconda installer homepage can be found here.」と表示されますので、"archive" や "here" をクリックし、目的のバージョンをインストールできます。

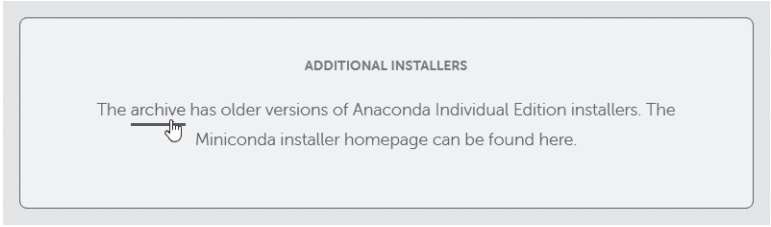

図1.3●古いバージョンなどを利用したい場合

1.1.2 インストール

ダウンロードが終わると、ブラウザの下部や上部にインストールを促す表示が現れます。表示形式はブラウザの種類などに依存します。この例では、「ファイルを開く」をクリックします。

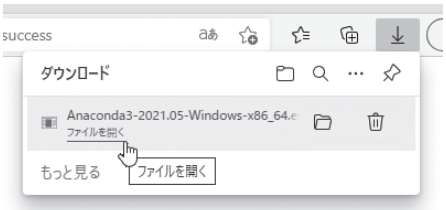

図1.4●「ファイルを開く」をクリック

しばらくすると、インストーラーが起動します。Anaconda のビット数表示や、ほかのアプリケーションを閉じるように案内されますので、間違いがないことを確認したら「Next >」ボタンをクリックして次の画面に進みましょう。

図1.5●インストーラーの起動画面

ライセンスの確認画面が表示されますので、最後までスクロールしながら、ライセンスを一通り確認し、「I Agree」ボタンをクリックして次の画面に進みます。

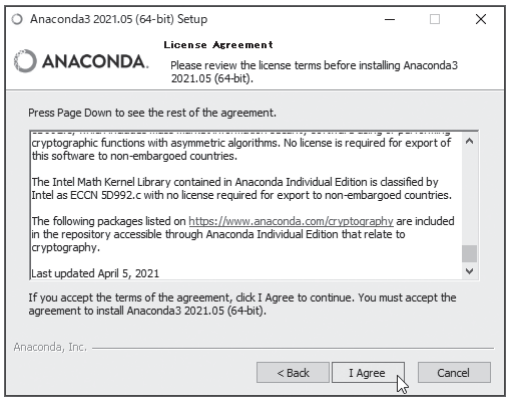

図1.6●ライセンスの確認画面

現在の単一ユーザーとしてインストールするか、全ユーザーへインストールするかを選択する画面が表示されます。単一ユーザーとして（Just Me）インストールするのが推奨されています。複数の人でパソコンを共用し、全員が同じバージョンの Anaconda を使用するような特殊な環境でない場合、「Just Me」を選択したまま「Next >」ボタンをクリックして次の画面に進みましょう。

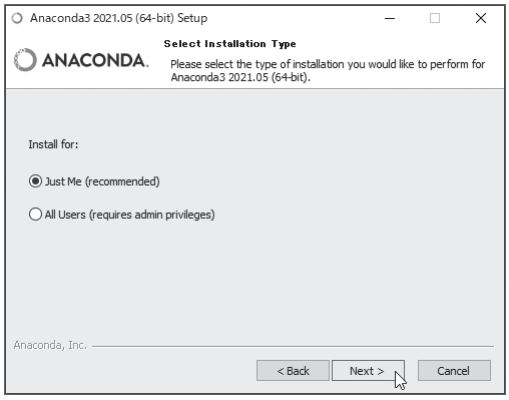

図1.7●インストールの種類を選択

インストール先を選択する画面が表示されます。特別な理由がないかぎり変更する必要はないでしょう。インストール先を確認・選択して「Next >」ボタンをクリックして次の画面に進みましょう。「¥User」の配下にインストールしたくない場合は、「Browse...」ボタンをクリックし、適切な場所へインストールしましょう。

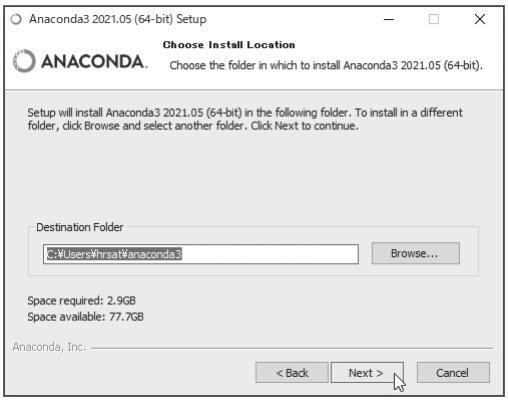

図1.8●インストール先を選択

インストールのオプションを選択する画面が表示されます。特に変更の必要はありませんので、このまま「Install」ボタンをクリックしインストールを開始します。「Add Anaconda to my PATH environment variable」は、環境変数 PATH に Anaconda のフォルダーを追加するか否かを決める選択肢です。最初の方に「Not recommended ...」と記述されており推奨されていません。これをチェックしなくても、コンソールを使用したいときは「Anaconda Prompt」を使用すると Python へのパスは通っています。また「Register Anaconda as my default Python 3.x」は「Anaconda をデフォルトの Python 3.x として登録するか」の選択肢です。

これにチェックを付けておくと、インストールした Python がシステム上のプライマリとして扱われますので、ほかの開発ツールは、この Python（Anaconda）を自動で認識します。詳細は表示されているメッセージを参照してください。

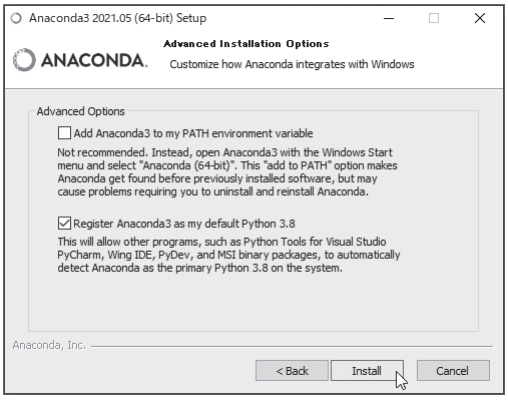

図1.9●インストールのオプションを選択

「Install」をクリックすると、しばらくインストール作業が続きますので、完了するのを待ちます。

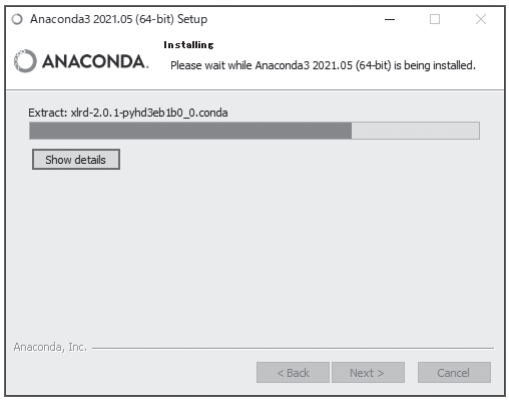

図1.10●インストール中

インストールが完了すると、図のように「Completed」という表示に変わります。「Next>」をクリックして次の画面に進みます。示した画面は「Show details」ボタンをクリックした状態です。「Show details」ボタンをクリックしないと、インストール状況は表示されません。

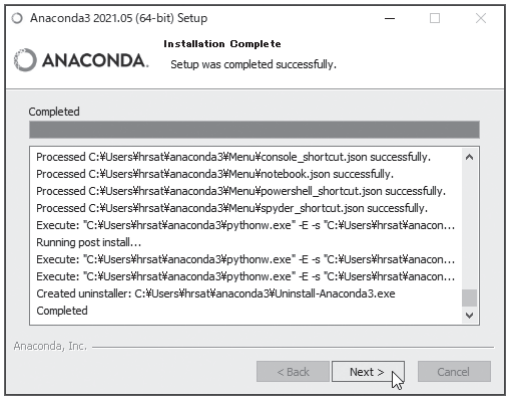

図1.11●インストール完了

PyCharm に関する案内メッセージが現れます。「PyCharm for Anaconda」が available になっているサイトの URL も表示されます。今回は、PyCharm などは使用しません。内容を読んで「Next >」をクリックして次の画面に進みます。

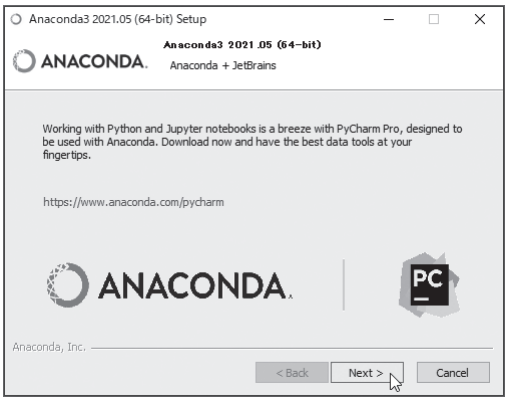

図1.12●インストーラーを終了する

インストール作業が終わったことを表す画面が表示されます。「Finish」ボタンをクリックし、インストーラーを終了させます。チェックボックスにチェックされたままだと関連情報が表示されますが、必要なければチェックを外してください。いずれにしても Anaconda 3 はインストールされています。

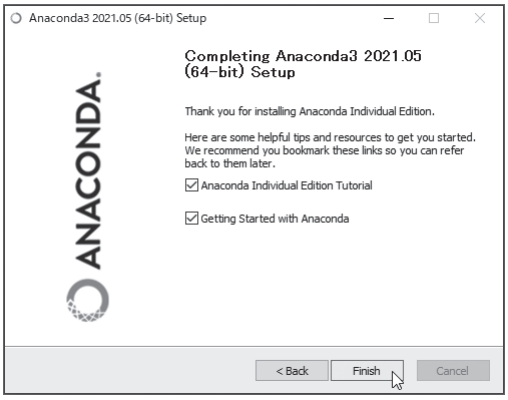

図1.13●インストール完了

「Finish」ボタンをクリックしたあとに、「Welcome to Anaconda!」の web サイトを表示したブラウザが現れるときがあります。インストール作業と関係ありませんので、すぐに閉じて構いません。表示内容に興味のある人は、読んでみるのもよいと思います。

これで Anaconda のインストールは完了です。

1.2 Anaconda Navigator

Anaconda のインストールが完了すると、スタートメニューに Anaconda のフォルダーが追加されます。この中に「Anaconda Navigator」という項目があります。これは Anaconda を管理するためのツールです。まず、Anaconda Navigator を起動してみましょう。Anaconda Navigator は、Anaconda のフォルダーだけでなく、スタートメニューの「最近追加されたもの」にも表示されますので、そちらから起動します。

図1.14●Anaconda Navigatorの起動

Anaconda Navigator が立ち上がると、図に示す画面が現れます。左側に Environments、Learning、Community や Documentation などが存在します。これらはとても有益ですので、覗いてみるとよいでしょう。ここでは、右側に表示されている Spyder を起動します。「Launch」ボタンをクリックして起動します。

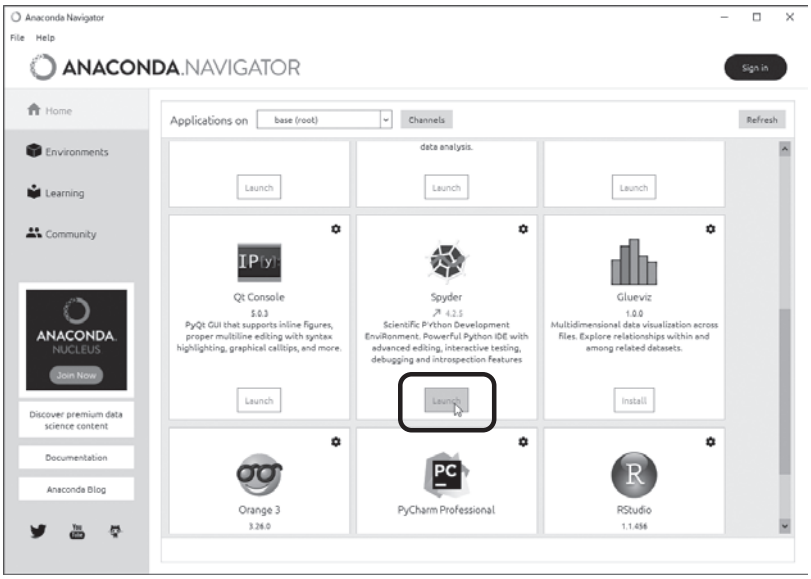

図1.15●Spyderの起動

起動前に、ファイヤーウォールが警告を発する場合や、Spyder の最新バージョンが存在する案内が表示される場合があります。ファイヤーウォールの警告が出たら、通信を許可してください。Spyder の更新は行う必要はありませんので、最新バージョンが案内されても構わず、×か「OK」ボタンをクリックします。ここでは、×をクリックし、最新バージョンへ変更しません。もちろん、最新バージョンを使用したいなら、案内メッセージに従って最新に更新してもよいでしょう。

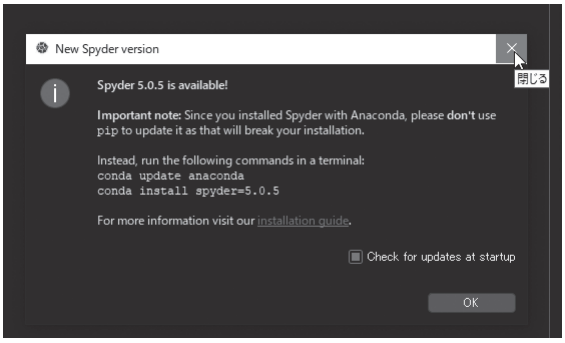

図1.16●Spyderのアップデート案内

次に、「Welcome to Spyder」の画面が現れます。interactive tour などを行いたいなら、「Start

tour」をクリックしてください。必要なければ、×か「Dismiss」ボタンをクリックします。ここでは、×をクリックし、tour は行いません。興味のある人は、「Start tour」をクリックするとよいでしょう。

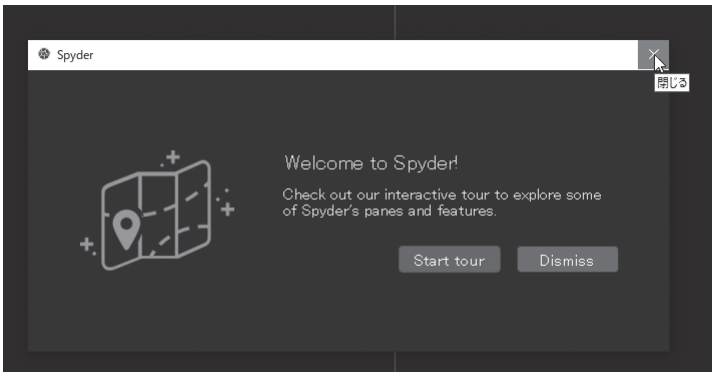

図1.17●Spyderのinteractive tourの案内

しばらくすると Spyder が起動します。Python や Spyder のインストールに問題がないか確認するために、「Hello Python」を表示する簡単なプログラムを作ってみましょう。

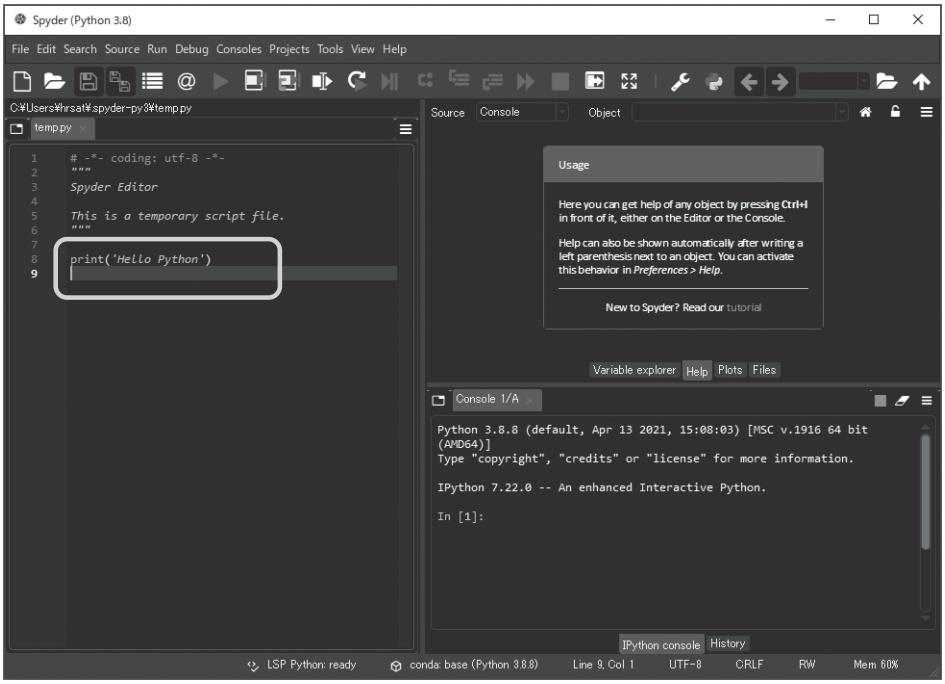

図1.18●「Hello Python」プログラムの作成

左側のエディターペインに print 文を入力します。実行するには ▶ をクリックします。すると、最初の実行で以降の画面が現れます。コンソールの選択、コマンドラインや作業ディレクトリなどを指定できます。ここでは何も変更せず「実行」ボタンをクリックします。

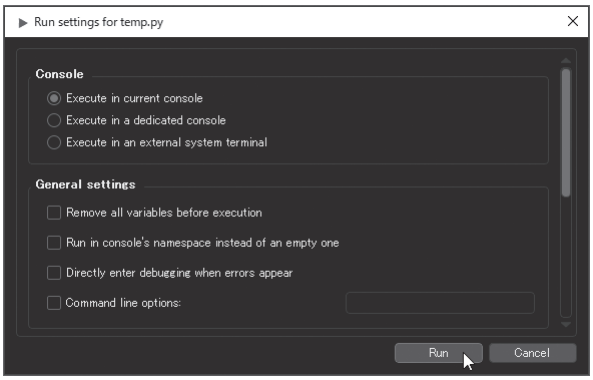

図1.19●実行設定の画面

すると、右下の IPython コンソールに実行結果が表示されます。

図1.20●実行結果の表示

今回は Spyder を Anaconda Navigator 経由で起動しましたが、Anaconda Navigator を使用する必要がないときは、直接 Spyder を起動するのもよいでしょう。