DataScientist Society

データサイエンス100本ノック

構造化データ加工編

2022.03.25

商標について

- ✓ DockerおよびDockerのロゴは、米国およびその他の国におけるDocker, Inc.の商標または登録商標です。Docker, Inc. およびその他の当事者も、本書で使用されている他の用語で商標権を有している場合があります。
- ✓ 「 Jupyter」は、Project Jupyterが帰属するNumFOCUS財団の商標です。Jupyterのロゴ（いくつかのバリエーションがあります）は、NumFOCUSの商標でもあります。Jupyter商標は、米国特許商標庁に登録されています。
- ✓ Git Logo by Jason Long is licensed under the Creative Commons Attribution 3.0 Unported License.
- ✓ その他、本資料に登場する会社名・商品名は、それぞれの権利者の商標または登録商標となります。

Copyright © 2018 The Japan DataScientist Society. All Rights Reserved.

データサイエンス100本ノック（構造化データ加工編）の狙い

データ活用の重要性に対する認知が広がる中で、データサイエンティストの必要性も増々高まっています。そのような中で、書籍やWebサイトなど、自己研鑽に必要な情報源も多く提供されています。しかし、実践するための「データ」や「プログラミング実行環境」を持ち合わせていないため、「実践力」を身につけることが難しい、というケースが見られます。

そこでデータサイエンティスト協会では、データサイエンス100本ノック（構造化データ加工編）をリリースいたしました（以下、100本ノック）。データと実行環境構築スクリプト、そして演習問題をワンセットにして公開しています。

自然言語処理や画像処理、深層学習などの実践練習環境は、すでに素晴らしい題材が公開されています。例えば深層学習については、東京大学 松尾研究室「Deep Learning基礎講座演習コンテンツ 公開ページ」にてnotebookファイルが公開されており、Google Colaboratoryなどを利用することで無料で実践することができます。

一方、構造化データについては、無料で利用できるオープンデータが多数あるものの、データ分析力の実践という観点で環境整備がなされたものはあまり多くありません。そこで、分析実務においては構造化データの活用が多くを占めるという実態を鑑み、「構造化データ加工編」というサブタイトルで100本ノックを公開しました。

データの加工・集計、統計学や機械学習を駆使したモデリングの前処理など、基礎的なデータハンドリングの修行場として利用いただければと思います。

Copyright © 2018 The Japan DataScientist Society. All Rights Reserved.

演習問題の構成

- 実行環境のサポート言語はSQL、Python、Rとなります
- それぞれの言語で、同様の設問を100問用意しています
- 各設問に対する解答例のファイルも用意しています

注）設問によっては、向かない言語もありますが、実際の実務では企業におけるセキュリティやシステム実装の観点などから、どうしてもSQLでやらねばならないケース、Pythonでやらねばならないケース、Rでやらねばならないケースなどが出てきます。このような観点から、リソースを過剰に使ってしまうなどあまり良いコード例ではないものもありますが、分析で良く利用される3言語について、「この言語のときはこうすればできる」ということが学べるよう設問を揃えた構成としています。

No.	大区分	設問数
1	列に対する操作	3
2	行に対する操作	6
3	あいまい条件	7
4	ソート	4
5	集計	13
6	副問合せ	2
7	結合	7
8	縦横変換	2
9	データ変換	14
10	数値変換	4

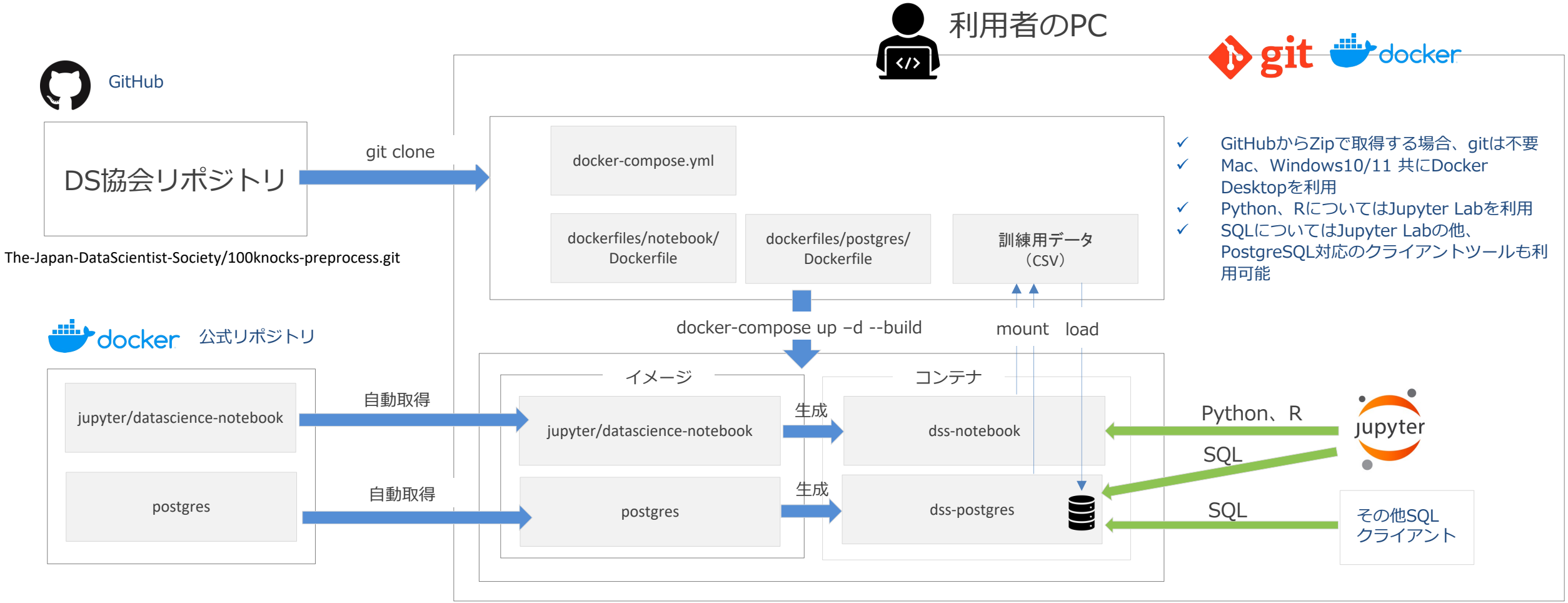
No.	大区分	設問数
11	四則演算	7
12	日付型の計算	5
13	サンプリング	2
14	外れ値・異常値	2
15	欠損値	5
16	除算エラー対応	1
17	座標データ	2
18	名寄せ	2
19	データ分割	2
20	不均衡データ	1

No.	大区分	設問数
21	正規化・非正規化	2
22	ファイル入出力	7
合計		100

Copyright © 2018 The Japan DataScientist Society. All Rights Reserved.

実践環境

100本ノックの仕組みは以下の通りとなります。



Copyright © 2018 The Japan DataScientist Society. All Rights Reserved.

必要なPCリソースおよびソフトウェア

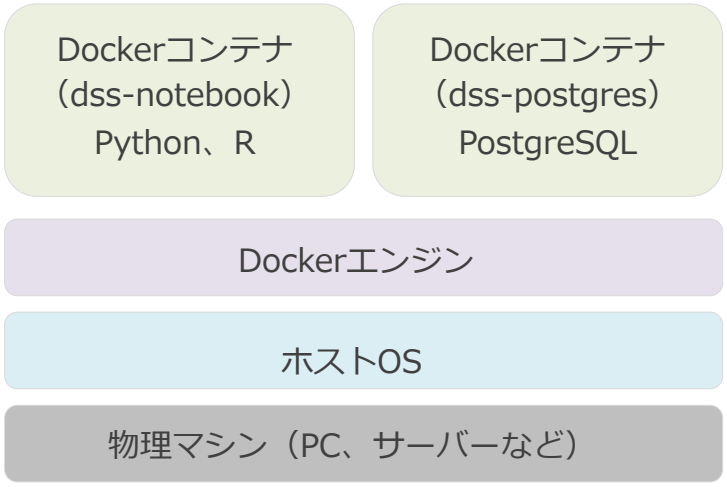
OS	✓ macOS 10.13 以上 ✓ Windows10/11 Professional Edition ✓ Windows10/11 Home Edition
メモリ	✓ 8GB以上（Dockerへの割当4GB以上）を推奨
HDD	✓ 15GB以上の空きスペース
ソフトウェア	✓ Docker Desktop※ ✓ git（GitHubからZIPで取得する場合は不要）

Copyright © 2018 The Japan DataScientist Society. All Rights Reserved.

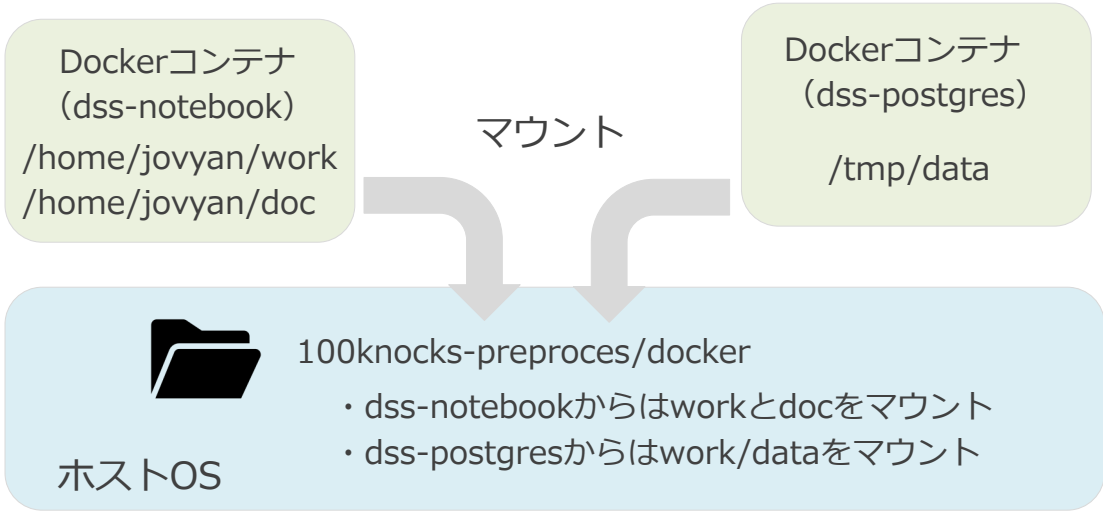
Dockerの概要とファイル共有

- Dockerは、Linuxのコンテナ技術を使った仮想環境であり、PCの中であたかも別のLinux OSマシンがあるかのように動かすことができるソフトウェアです
- Dockerの中のLinuxをゲストOSと呼び、PC本体のOSをホストOSと呼びます
- OSおよび様々なソフトウェアがインストールされたDockerイメージとよばれるファイルが公開されており、DockerイメージからDockerコンテナと呼ばれるプロセスを立ち上げることでPC内に仮想環境を実現します
- 100本ノックでは、以下２つのコンテナが起動するように設定されています（図①）
  - PythonおよびRの実行環境であるdss-notebook（ポート番号：8888）
  - PostgreSQLによるSQL実行環境であるdss-postgres（port番号：5432）
- 複数のコンテナを管理する際に便利なDocker Composeという機能を利用しています
- ゲストOSからホストOSのディレクトリをマウントすることが可能であり、これによりホストOSとゲストOSとのファイル受け渡しを容易に行うことができます
- それぞれのコンテナから、100本ノックのディレクトリが自動的にマウントされるよう設定されています(図②)
- PCと仮想環境を隔離し、分析環境構築に伴うライブラリインストールなどでマシン本体を汚してしまうことがないため、実際の分析実務でもよく利用されます

図①：Dockerの構成イメージ



図②：コンテナによるディレクトリ共有



Copyright © 2018 The Japan DataScientist Society. All Rights Reserved.

環境構築手順

以下の手順で環境を構築します。なお、③の手順ではgitの機能を使わずzipファイルで100本ノック関連ファイルを取得することも可能です。その場合は、①の手順が不要となります。

① gitをインストールする（詳細は関連書籍やWebサイト等を参照願います）

② Docker Desktopをインストールする※1（詳細は関連書籍やWebサイト等を参照願います）

③ 100本ノックリポジトリをクローンする※

- HTTPSの場合（通常はこちら）　　：
- SSHの場合（SSHの設定をしている方）　：git clone git@github.com:The-Japan-DataScientist-Society/100knocks-preprocess.git

④ ターミナル等※3でdocker-composeファイルのあるディレクトリまで移動（cd 100knocks-preprocess）

⑤ ターミナル等※3でコンテナ作成コマンドを実行する（docker-compose up -d --build）

※1: 従来、Windows10 Home Editionの場合はDocker Toolboxを利用する必要がありましたが、その後Docker Desktopを利用できるようになりました。ただし、WindowsでDocker Desktopを使う場合はWSL2（Windows上でLinuxの実行環境を構築できるシステム）をインストールすることが推奨されています。Windowsにおけるインストールの注意点は次ページに記載します。

※2: 自身のユーザーホームディレクトリ配下に取得すると設定作業がもっとも簡単になります。それ以外のディレクトリに格納する場合、macOSではDockerの共有ディレクトリ設定が必要となる場合があります（「macOSでのDocker Desktopのリソース・共有設定」のページを参照願います）。

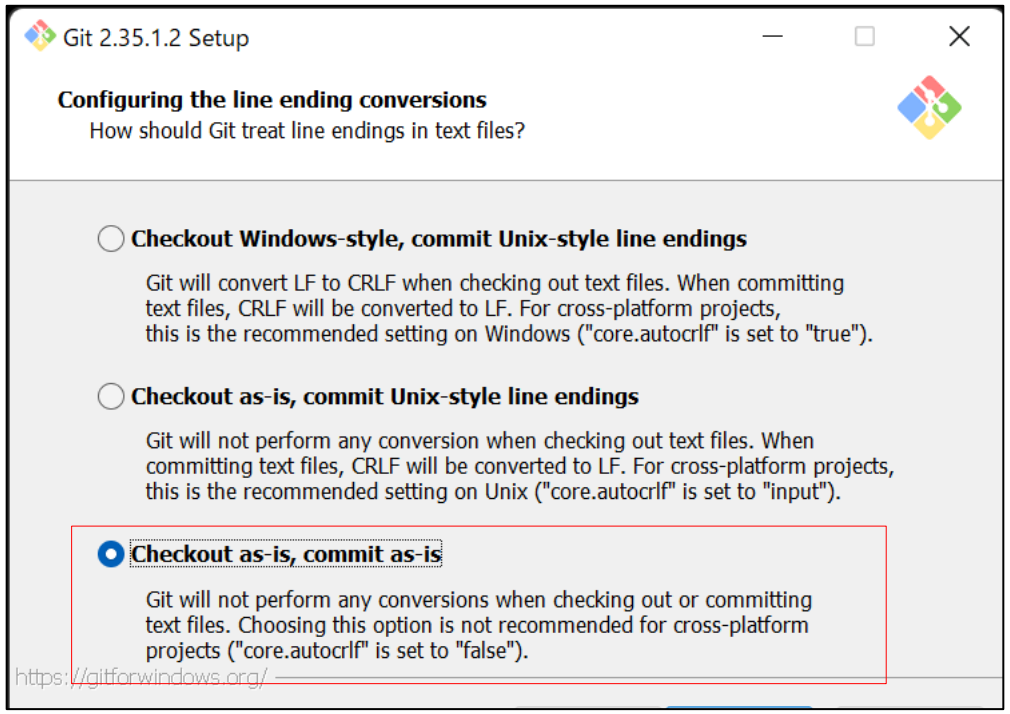
※3、ターミナル（macOS）またはコマンドプロンプト（Windows）を使用します。

ダウンロード時間を除けば作業時間は10分程度となります。

Copyright © 2018 The Japan DataScientist Society. All Rights Reserved.

WindowsにおけるGitインストールの注意点

Gitをインストールする際、以下図のようにcheckout/commit時の改行コード編集に関する挙動を設定することになります。100本ノックの環境を構築するためのスクリプト群は改行コードが「LF」となっていることを前提に動作します。したがって、インストール時には図の赤枠部分を選択してすすめる必要があります。初期設定では一番上が選択されていますが、この場合、Windowsでは改行コードが「CR+LF」に変換されてしまい、コンテナ作成時に処理がエラーとなります。既にGitがインストール済みの場合、あとからこの設定を変更することもできます。変更方法についてはWebサイトなどで探してみてください。

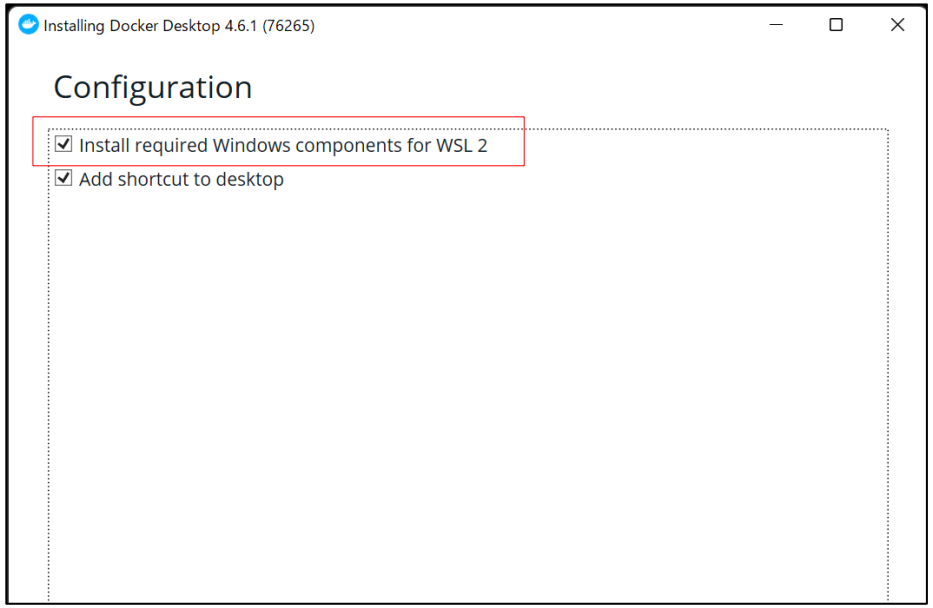


Copyright © 2018 The Japan DataScientist Society. All Rights Reserved.

WindowsにおけるDocker Desktopインストールの注意点

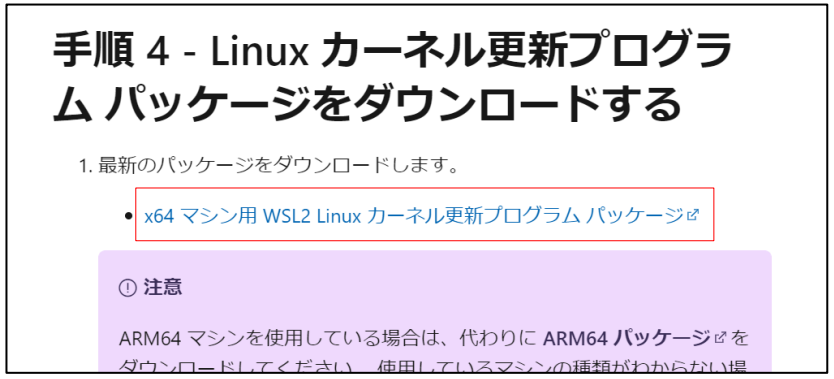
WindowsでDockerを利用する場合、WSL2をインストールすることが推奨されています。Docker Desktopをインストールする際、以下図①の「Install required Windows components for WSL2」にチェックをつけます。インストール完了後、Docker Desktopを起動すると、WSL2が未インストールの場合は図②のメッセージが表示されます。記載されているURLのリンクをクリックすると図③のサイトが開かれるので、「x64 マシン用 WSL2 Linuxカーネル更新プログラムパッケージ」をダウンロードしてインストールします。

図①



図②

図③



Copyright © 2018 The Japan DataScientist Society. All Rights Reserved.

docker compose 基本コマンド

docker composeというコマンドを使うことで、複数のコンテナ管理を行うことができます。docker-compose.ymlファイルが存在するディレクトリで実行するか、当ファイルをオプションで指定する必要があります。以下に基本コマンドを記載します。

- docker compose up –d
  - docker-compose.ymlファイルの定義に従いDockerイメージを取得・ビルドし、Dockerコンテナを作成・起動します。"--build"オプションを付与することで、イメージがすでに存在する場合も強制的にビルドします
- docker compose down
  - docker-compose.ymlの定義に従い、関連するコンテナを削除します。"--rmi all"オプションを付与することで、関連するイメージも削除することができます。
- docker compose stop
  - コンテナを停止します。
- docker compose start
  - コンテナを起動します。
- docker compose exec サービス名（notebook | db） bash
  - コンテナ内のLinux OSに入ることができます。dss-notebnookに入る場合はnotebookを、dss-postgresに入る場合はdbをサービス名として指定します。

Copyright © 2018 The Japan DataScientist Society. All Rights Reserved.

コンテナを完全に作り直したい場合

Dockerイメージに変化がない場合、コンテナを再作成してもキャッシュが利用されます。後述の「注意事項」のようなケースによりコンテナがうまく作成できなかった場合、原因を取り除いた上でコンテナの再作成をする必要がありますが、キャッシュが利用されてしまうことでうまく再作成できないことがあります。このような状況では、以下の手順でコンテナを再作成する必要があります。

- docker compose down --rmi all
  - docker-compose.ymlの定義に従い、関連するコンテナを削除します。"--rmi all"オプションを付与することで、関連するイメージも削除することができます。
- docker compose build --no-cache
  - キャッシュを使わずにイメージをビルドします。
- docker compose up –d
  - コンテナを起動します。

Copyright © 2018 The Japan DataScientist Society. All Rights Reserved.

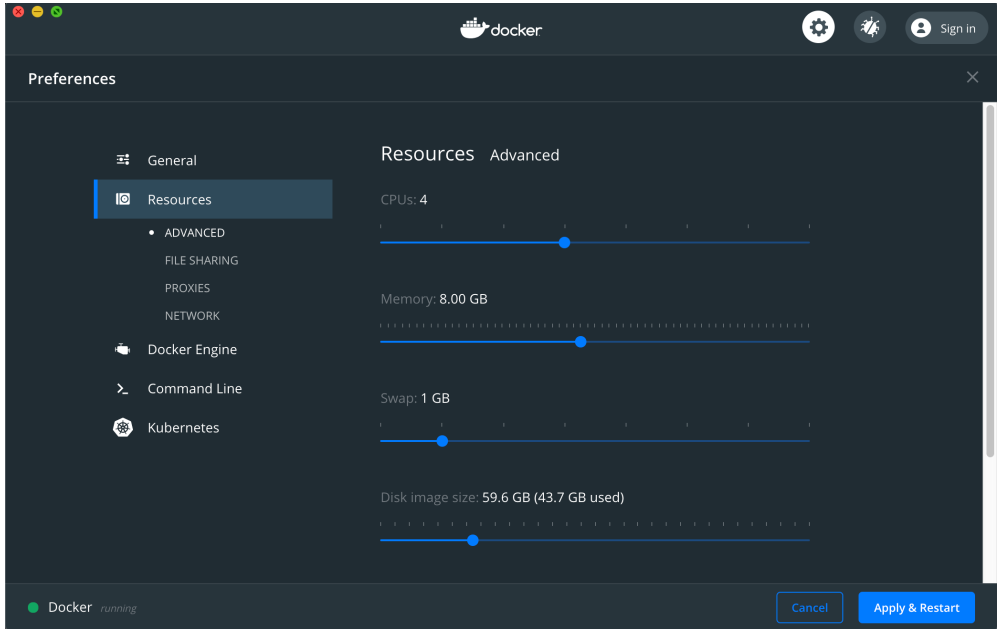
macOSでのDocker Desktopのリソース・共有設定

macOSの場合、初期設定ではご利用の環境により利用可能リソースの上限が低く設定されることがあります。また、ユーザのホームディレクトリ配下以外に100本ノックのファイル群を配置した場合、Dockerのファイル共有設定が必要となることがあります。このような場合はDocker Desktopの設定でCPUやメモリの使用上限、ファイルの共有設定を行います。

なお、WindowsでWSL2を利用する場合、これによりリソースの管理が行われますが100本ノックを利用する限りにおいて特段の設定は不要です。

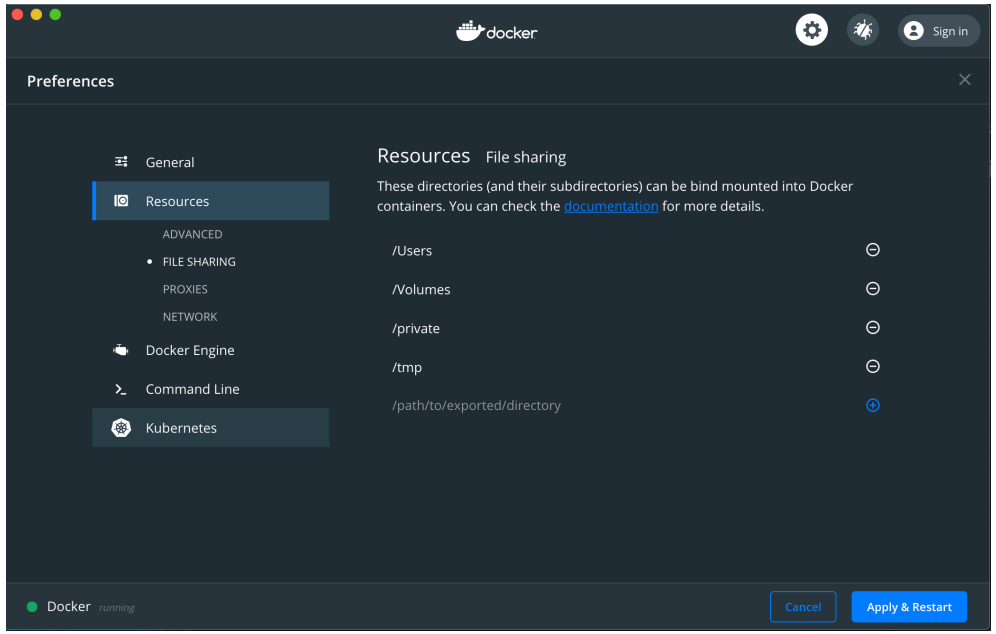
■リソース設定

Docker Desktopを起動します。上部にある設定アイコン（歯車のアイコン）をクリックし、Preference → Resource → ADVANCEDと進んで設定します。CPUとメモリそれぞれPC搭載量の半分程度を推奨します。



■共有設定

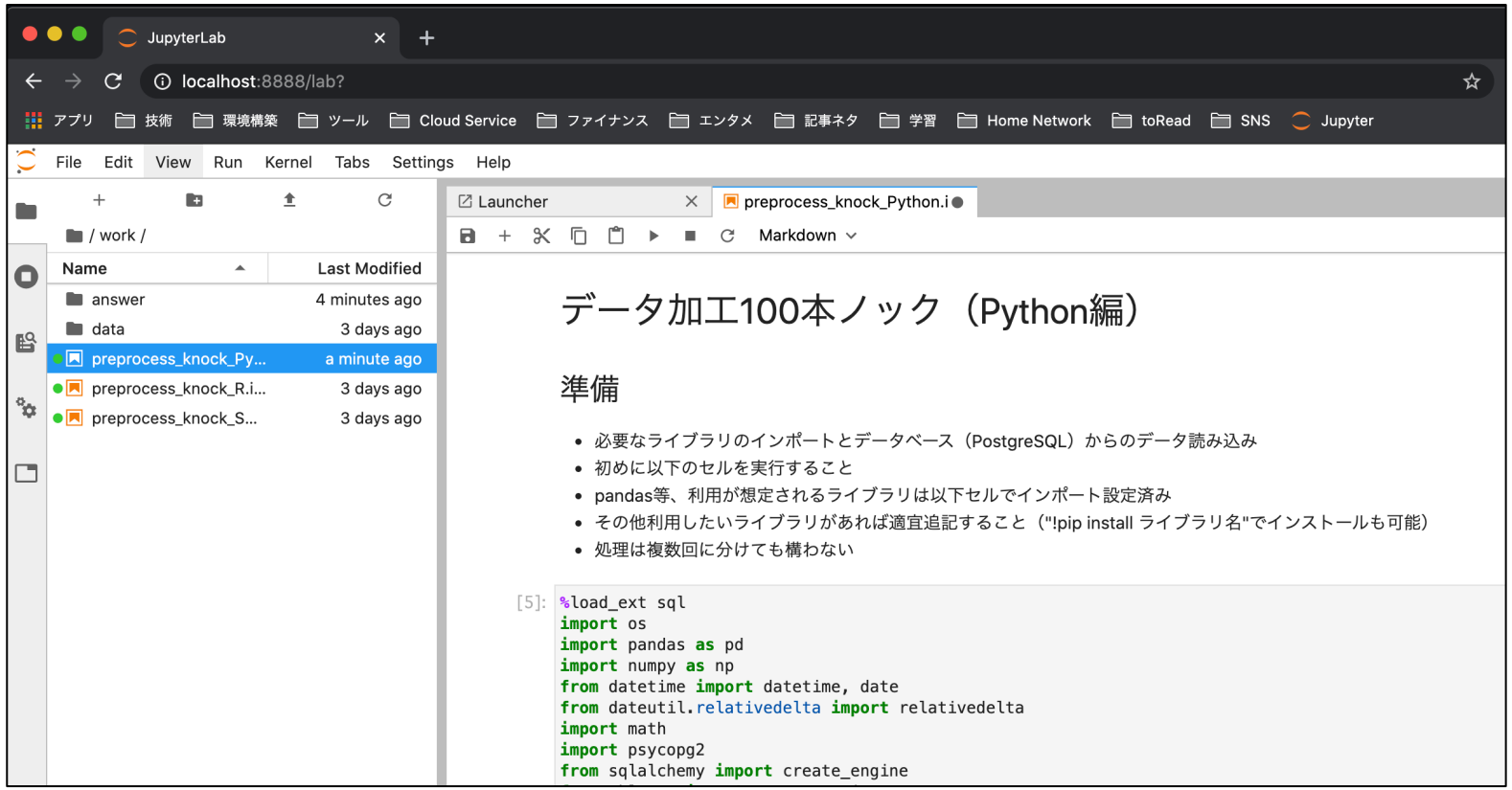
Docker Desktopを起動します。上部にある設定アイコン（歯車のアイコン）をクリックし、Preference → Resource → FILE SHARINGと進んで設定します。100本ノックファイル群を格納したディレクトリを追加します。



Copyright © 2018 The Japan DataScientist Society. All Rights Reserved.

利用方法

http://localhost:8888にブラウザからアクセスします。workフォルダ内のnotebookファイルを開いてノックを実践しましょう。
解答ファイルはwork/answer配下に格納されています。



Copyright © 2018 The Japan DataScientist Society. All Rights Reserved.

その他注意事項

【コンテナ作成に失敗した場合】
notebookファイルが存在しない、またはPostgreSQLのデータベースにアクセスできない場合はコンテナの作成に失敗している可能性があります。

① 100本ノックファイル群を格納したディレクトリがマウントできていないケース
ホームディレクトリ配下以外にファイルを格納した場合、OSのアクセス権限設定やDockerのファイル共有設定が必要となります。関連する設定を確認しましょう。

② Windows版gitが勝手に改行コードを変えてしまうケース
PostgreSQLでデータベースを初期設定するためのスクリプトなど、本環境構築のスクリプト群は文字コードUTF-8、改行コードLFで記述されていますが、Windows版gitはインストール時の設定により改行コードをCR+LFに変換してしまうケースがあります。改行コードがCR+LFとならないよう設定するか、GitHubからZIPファイルで取得することで対応できます。

③ コンテナのbuildでエラーとなるケース
dockerを再起動してから実行することで解消することがあります。dockerを再起動してから実行してみてください。

④ Rのコードが実行できないケース
Rパッケージの取得先であるCRANサーバが停止していることで、R用のnotebookファイルからPostgreSQLにアクセスするためのパッケージなどがインストールできず、ノック用のデータを作成できないことがあります。その際はCRANサーバの再開を待ってからコンテナを再作成してください。このケースの場合はRのコードからPostgreSQLにアクセスする際に以下のエラーが発生します。
ERROR: Error in dbConnect(PostgreSQL(), host = host, port = port, dbname = dbname, : could not find function "dbConnect"

100本ノック環境が参照しているCRANサーバは以下の通りです。
https://cran.microsoft.com/snapshot/2021-10-05/

【免責】
データサイエンス100本ノック（構造化データ加工編）の利用に関するご質問等について、個別での対応は受けかねますので予めご了承ください。
また、データサイエンス100本ノック（構造化データ加工編）の利用により生じるいかなる問題についても、当協会は一切の責任を負いかねますのであらかじめご了承ください。

Copyright © 2018 The Japan DataScientist Society. All Rights Reserved.

参考図書

- データサイエンス100本ノック（データ加工編）を作成するにあたり、以下の書籍や各種Webサイトを参考とさせていただきました。

本橋智光[著] 株式会社ホクソエム[監修]
「前処理大全 データ分析のためのSQL/R/Python実践テクニック」 技術評論社 2018年

Alice Zheng、Amanda Casari [著] 株式会社ホクソエム[訳]
「機械学習のための特徴量エンジニアリング その原理とPythonによる実践」 技術評論社 2019年

- データサイエンス100本ノック（データ加工編）の解説本として以下の書籍が出版されています。

森谷和弘、鈴木雅也[著]
「データサイエンス100本ノック構造化データ加工編ガイドブック」 ソシム 2022年

Copyright © 2018 The Japan DataScientist Society. All Rights Reserved.