

必修ライブラリ③… データ分析「pandas」

ご購入はこちら

高橋 知宏

```

CSVの読み込み

In [1]: %%writefile test.csv
        ,A,B,C
        2014,1,2,3
        2015,4,5,6
        2016,7,8,9

Writing test.csv

In [134]: import pandas as pd
          pd.read_csv('test.csv', index_col=0)

Out[134]:


|      | A | B | C |
|------|---|---|---|
| 2014 | 1 | 2 | 3 |
| 2015 | 4 | 5 | 6 |
| 2016 | 7 | 8 | 9 |


```

図1 データ・ファイルの読み込み&解析(%%writefileコマンドでCSVファイルを作成し、pandasのread_csv関数で読み込む)

特徴&準備

● データ型はNumPyのarrayが基本

Pythonでデータ分析を行うときは、pandasというライブラリが定番です。数表や、日付や時刻に紐づいた時系列データの取り扱いに優れています。NumPyのデータ構造arrayをベースに実装されているため、大きなデータでも効率良く取り扱うことができます。array操作の概念であるスライス、ファンシー・インデックス、ビューなどがほぼそのまま適用できます。

● 準備

本特集で紹介しているPythonソフトウェアを使うには、自前でパッケージをインストールしている場合はimportが必要です。Anacondaやクラウド・サービスの場合は基本的にはそのまま使えます。

データ分析向け機能を確認する

● その1：データ・ファイルの読み込み&解析

さっそく、CSVファイルからデータをpandasで読み

```

In [3]: import pandas_datareader.data as web

ImportErrorTraceback (most recent call last)
<ipython-input-3-01a61b993f3d> in <module>()
----> 1 import pandas_datareader.data as web
      2 N225 = web.DataReader('N225', 'yahoo
      3 N225.tail()

ImportError: No module named pandas_dataread
モジュールがないことがある

In [4]: !pip install pandas_datareader

Collecting pandas_datareader
Using cached pandas_datareader-0.10.0-py2.py3-none-any.whl
Requirement already satisfied: pandas in /op
(from pandas_datareader)
pipコマンドでインストールする

```

図2 Web APIを使ってオープンデータなどにアクセスするための拡張機能pandas_datareaderを準備する

pipコマンドを使ってインストールする。root権限が必要な場合はsudoを付ける

込んでみます。まずは小さなCSVファイルをJupyter Notebookのセルからマジック・コマンド(コラム1, コラム2)で作ります(図1)。

そのファイルをread_csv関数で読み込みます。

CSVファイルを読み込んだデータは表形式で表示されます。これはDataFrameというオブジェクトです。行や列、あるいはその両方を指定してデータを取り出すことができます。

データを取り出すだけでなく、行や列を対象にした計算が可能です。行や列同士の計算やブロードキャストが可能など、このあたりの感じはNumPyと同様です。

● その2：Web APIによる時系列オープンデータの取得&解析

ネットにはいろいろなデータが公開されています。オープンデータとして、APIが公開されているものもあります。pandasにはそのようなデータにアクセスするためのpandas_datareaderという拡張が提供されています。ここでは経済データとして世界銀行が提供している年次のGDPデータを取得してグラフ化してみましょう。