ダウンロード・データあります

Excelにデータを流し込んで集計する感覚で利用できるpandasライブラリで体験！

読み込み／整形／抽出…Pythonでデータ解析

土屋 健

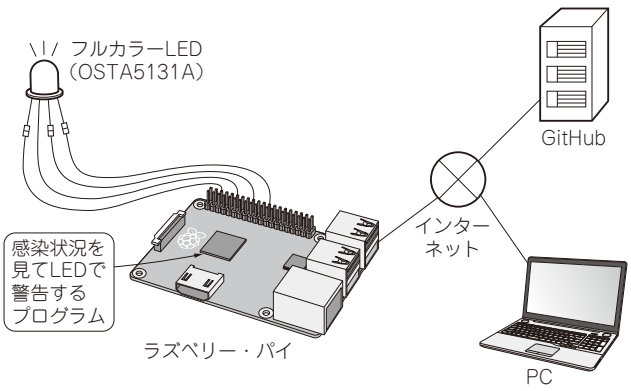

図1　今回作るもの

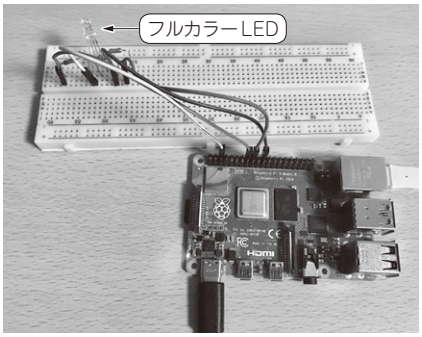

写真1　ラズベリー・パイに取り込んだ感染者データをPythonライブラリであるpandasで処理してLEDに表示する

ここでは，ラズベリー・パイを使ってPythonで大量データの整理を行い，その良さを体験します（**図1**）．

解析するデータとして，インターネットから入手可能なパブリック・データを利用します．さまざまなデータの中から今回は，新型コロナウイルスのこともあって見聞きする，

- 新型コロナウイルス感染者数(1)
- 人の移動情報(2)

などの情報を扱います．

ふぞろいのデータを扱うための前処理

● データの種類

データには定型データや非定型データがあります．世の中にあるデータはもともとコンピュータ処理にかける前提で用意されている訳ではないので，非定型データが圧倒的に多いです．

そういったデータを集めて解析処理するために前処理と呼ばれるデータの整形を行います．

重要な前処理ですが，結構面倒な処理でもあります．この処理によく使われるPython用ライブラリのpandas（`https://pandas.pydata.org/`）の使用例を示します．

● pandasでデータの前処理をする

pandasはデータ解析を行うために使われるライブラリです．CSVの読み込み，データ整形，データ抽出などデータ解析に便利な機能を持っています．

▶データ構造

pandasではデータ・フレームという構造でデータを扱います．これはExcelの表みたいなもの（行，列でアクセスできる）です．

データ・フレームには行と列にそれぞれラベル（名前）を付与でき，行ラベルはindex，列ラベルはcolumnsと呼ばれます．それらを指定することで特定の行や列のデータを参照できます．**図2**にデータ・フレームの構造を示します．

pandasデータ・フレームは，見た目にはこのような2次元の表で，行・列を名前や位置で指定することで任意の範囲のデータにアクセスできるデータ構造です．

内部的には高速なデータ・アクセスを実現するために工夫されているはずですが，利用者はそういったことを意識せず，アクセスしたいデータの取り出し方を指定するだけで，後はデータ・フレームが面倒をみてくれます．

▶プログラムに不慣れでも実装しやすい

データの編集処理や統計量計算などを行う集計処理