Librabuch Librabuch

- HOME
- AUTHOR

- HOME
- AUTHOR

データの前処理や加工に使えるPython csvkit

2014-12-24 Wed

Takahiro Ikeuchi

みなさまこんばんは。Python Advent Calendar 2014 24日目の記事です。

先日のpyhackで@atelierhideに教えてもらった、データ前処理スト垂涎のライブラリの紹介をすることにしました。

csvkit とは

csvkitは、コマンドラインでCSVやTSVファイルを取り扱うのに便利なライブラリです。データの前処理や加工をLinux/UNIXのコマンドラインで行っている環境もあると思いますが、それを代替する、あるいは組み合わせて使うとよいのがcsvkitです。

- csvkit

pipでインストール出来ます。Python3.4にもインストールは可能ですが、一部の機能が動作しないことを確認しています。今回は2.7にインストールしました。

```
pip install csvkit
```

具体的な使い方を見ていきます。

基本的な使い方

ここからはiris.csvのデータをcsvkitで触っていきます。

csvcut

```
csvcut -n iris.csv
```

上記のコマンドを実行すると下記の様に出力されます。

```
1: Sepal Length
2: Sepal Width
3: Petal Length
4: Petal Width
5: Species
```

ヘッダ行を出力してくれています。

「5: Species」の列だけ抽出してみます。

```
csvcut -c Species iris.csv

# 列数指定でもOK
csvcut -c 5 iris.csv
```

結果は以下の通り。

```
Species
setosa
```

```
setosa
setosa
setssa
〜略〜
```

ちなみに、Delimiter はオプションで指定することが出来ますので、ファイルがカンマ区切りでなくとも同様に扱えます。

csvlook

```
# csvファイルを整形して出力します。
csvlook iris.csv
```

csvlookは、次の様にCSVファイルを整形してくれるコマンドです。

```
|--------------+-------------+--------------+-------------+-------------|
|  Sepal Length | Sepal Width | Petal Length | Petal Width | Species     |
|--------------+-------------+--------------+-------------+-------------|
|  5.1          | 3.5         | 1.4          | 0.2         | setosa      |
|  4.9          | 3.0         | 1.4          | 0.2         | setosa      |
|  4.7          | 3.2         | 1.3          | 0.2         | setosa      |
〜略〜
|  5.9          | 3.0         | 5.1          | 1.8         | virginica   |
|--------------+-------------+--------------+-------------+-------------|
```

csvkitはLinux/UNIXコマンド同等にパイプで処理を渡すことによって柔軟な処理が行えます。

```
# 先頭の5行・1列目、5列目を整形して出力します。
head -5 iris.csv | csvcut -c 1,5 | csvlook
```

上記の実行結果は下記の通りです。

```
|--------------+----------|
|  Sepal Length | Species  |
|--------------+----------|
|  5.1          | setosa   |
|  4.9          | setosa   |
|  4.7          | setosa   |
|  4.6          | setosa   |
|--------------+----------|
```

csvgrep

```
# Sepal Length列が 5.1 の行だけを出力します。
csvgrep -c 1 -m "5.1" iris.csv | csvlook
```

grepコマンドのように、指定する文字列で絞り込むことができます。

```
|--------------+-------------+--------------+-------------+-------------|
|  Sepal Length | Sepal Width | Petal Length | Petal Width | Species     |
|--------------+-------------+--------------+-------------+-------------|
|  5.1          | 3.5         | 1.4          | 0.2         | setosa      |
|  5.1          | 3.5         | 1.4          | 0.3         | setosa      |
|  5.1          | 3.8         | 1.5          | 0.3         | setosa      |
|  5.1          | 3.7         | 1.5          | 0.4         | setosa      |
|  5.1          | 3.3         | 1.7          | 0.5         | setosa      |
|  5.1          | 3.4         | 1.5          | 0.2         | setosa      |
|  5.1          | 3.8         | 1.9          | 0.4         | setosa      |
|  5.1          | 3.8         | 1.6          | 0.2         | setosa      |
|  5.1          | 2.5         | 3.0          | 1.1         | versicolor  |
|--------------+-------------+--------------+-------------+-------------|
```

正規表現も利用出来ます。

```
csvgrep -c 1 -r "^5.[12]" iris.csv | csvlook
```

csvsort

```
# 2列,5列で抽出して1列目（元の2列目）でソートします
csvcut -c 2,5 iris.csv | csvsort -c 1 |csvlook
```

指定した列でソートして出力されます。

```
|-------------+-------------|
|  Sepal Width | Species     |
|-------------+-------------|
|  2.0         | versicolor  |
|  2.2         | versicolor  |
|  2.2         | versicolor  |
|  2.2         | virginica   |
```

```
|  2.3         | setosa      |
|  2.3         | versicolor  |
〜略〜
|  4.4         | setosa      |
|--------------+-------------|
```

csvjoin

csvjoinを試すために、もう1つのデータ name.csv を用意しました。

```
|------------------+----------------|
|  Scientific name | Pronunciation  |
|------------------+----------------|
|  setosa          | せとさ           |
|  versicolor      | ばーじからー         |
|  virginica       | ばーじにか          |
|------------------+----------------|
```

csvjoinを行います。

```
# iris.csv と name.csv を指定したカラムでOuter Joinします
csvjoin -c "Species,Scientific name" --outer iris.csv name.csv | csvcut -c 1,5,6,7 | csvlook | head -8
```

結果は次の通り。

```
|---------------+------------+-----------------+----------------|
|  Sepal Length | Species    | Scientific name | Pronunciation  |
|---------------+------------+-----------------+----------------|
|  5.1          | setosa     | setosa          | せとさ           |
|  4.9          | setosa     | setosa          | せとさ           |
|  4.7          | setosa     | setosa          | せとさ           |
|  4.6          | setosa     | setosa          | せとさ           |
|  5.0          | setosa     | setosa          | せとさ           |
```

オプションを指定することで、LEFT JOIN、RIGHT JOINも行えます。

csvclean

```
csvclean -n name.csv
```

name.csvに対して、csvcleanコマンドを実行すると「No errors.」とだけ表示されます。

では、name.csvを編集し、「ばーじからー」を「ばーじ,からー」にして保存します。再度csvcleanを実行すると下記の様に出力されます。

```
Line 2: Expected 2 columns, found 3 columns
```

このように、フォーマットが正常であるか異常であるかの判定、異常発見の手がかりを教えてくれるのがcsvcleanです。

csvstat

```
csvcut -c 1,5 iris.csv | csvstat
```

csvstatは、データの概要を出力しれくれます。

```
1. Sepal Length
    <type 'float'>
    Nulls: False
    Min: 4.3
    Max: 7.9
    Sum: 876.5
    Mean: 5.84333333333
    Median: 5.8
    Standard Deviation: 0.825301291785
    Unique values: 35
    5 most frequent values:
        5.0:    10
        6.3:    9
        5.1:    9
        6.7:    8
        5.7:    8
  2. Species
    <type 'unicode'>
    Nulls: False
    Values: setosa, versicolor, virginica

Row count: 150
```

数値の場合は最大値や最頻値、中央値、文字列の場合は値の一覧が出力されます。うーん、便利ですねぇ。

csvsql

```
csvsql -i postgresql iris.csv
```

csvsqlは、データベースへの入力をサポートしてくれるコマンドです。

```
CREATE TABLE iris (
    "Sepal Length" FLOAT NOT NULL,
    "Sepal Width" FLOAT NOT NULL,
    "Petal Length" FLOAT NOT NULL,
    "Petal Width" FLOAT NOT NULL,
    "Species" VARCHAR(10) NOT NULL
);
```

指定したファイルを解析して、適切なCREATE文を作成してくれているようです。MySQLやORACLEにも対応しています。データベースへの接続情報を与えることで、データのINSERTまで行うこともできます。

まとめ

いかがでしたでしょうか。head, cut, join など Linux/UNIXコマンドと似た機能もありますが、csvstatやcsvsqlなど、独自の便利な機能も実装されています。利用するのにPythonプログラミングの知識は必要ありませんので、間口の広いライブラリではないかと思います。

それではメリークリスマス！

Today's Proverb

Η ΑΛΗΘΕΙΑ ΕΛΕΥΘΕΡΩΣΕΙ ΥΜΑΣ

Links

- Twitter
- Github
- Facebook