

Week 05 • 데이터저널리즘

Data Analysis Using NumPy and Pandas 2

Joonhwan Lee

human-computer interaction + design lab.

오늘 다룰 내용

- Pandas

Using NumPy and Pandas in Data Analysis

Pandas

- ♦ Pandas는 numpy 기반에서 개발된 자료구조이다.
- ♦ R과 비슷한 데이터프레임과 이를 조작하기 위한 메소드를 제공한다.

Pandas의 자료구조

◆ Series

- ◆ Series는 요소(객체)를 담을 수 있는 1차원 배열 자료구조.

- ◆

```
>>> a = pd.Series([1, 2, 3, 4])
```

```
>>> a
```

```
0    1
```

```
1    2
```

```
2    3
```

```
3    4
```

```
dtype: int64
```

```
>>> a.values
```

```
array([1, 2, 3, 4])
```

```
>>> a.index
```

```
RangeIndex(start=0, stop=5, step=1)
```

Pandas의 자료구조

◆ DataFrame

- ◆ DataFrame은 스프레드시트의 표같은 형식의 자료구조로 여러 column으로 구성되어 있다. 각각의 컬럼은 다른 형식의 데이터를 담을 수 있다.

```
data = {  
    'state': ["PA", "NY", "CO", "CA"],  
    'population': [100, 200, 300, 400],  
    'size': [10, 20, 30, 40]  
}  
  
>>> b = pd.DataFrame(data)
```

Pandas의 자료구조

```
>>> b = pd.DataFrame(data)
```

	population	size	state
0	100	10	PA
1	200	20	NY
2	300	30	CO
3	400	40	CA

```
>>> b.population.mean()  
250.0
```

기본 Python에서의 matrix 만들기

S.No	Student Name	Science	English	History	Arts	Maths
1	Roy	80	75	85	90	95
2	John	75	80	75	85	100
3	Dave	80	80	80	90	95

```
✧ a = [['Roy', 80, 75, 85, 90, 95],  
        ['John', 75, 80, 75, 85, 100],  
        ['Dave', 80, 80, 80, 90, 95]]
```

10 Minutes to Pandas

10 Minutes to pandas

This is a short introduction to pandas, geared mainly for new users. You can see more complex recipes in the [Cookbook](#)

Customarily, we import as follows:

```
In [1]: import pandas as pd
In [2]: import numpy as np
In [3]: import matplotlib.pyplot as plt
```

Object Creation

See the [Data Structure Intro](#) section

Creating a [Series](#) by passing a list of values, letting pandas create a default integer index:

```
In [4]: s = pd.Series([1,3,5,np.nan,6,8])
In [5]: s
Out[5]:
```

<http://pandas.pydata.org/pandas-docs/stable/10min.html>

Questions...?
