hci+d lab.

Week 05 • 소셜네트워크 데이터마이닝과 분석

Data Analysis Using NumPy and Pandas 1

Joonhwan Lee
human-computer interaction + design lab.

오늘 다룰 내용

- NumPy

Using NumPy and Pandas in Data Analysis

Last Week

- Python에서 데이터를 불러올 때 다음과 같은 코드를 사용

```
import unicodecsv
def read_csv(filename):
  with open(filename, 'rb') as f:
    reader = unicodecsv.DictReader(f)
    return list(reader)

daily_engagement =
    read_csv("daily_engagement.csv")
```

Last Week

- 간단한 분석을 위한 코드

```
def get_unique_students(data):
  unique_students = set()
  for d in data:
    unique_students.add(d['acct'])
  return unique_students

unique_engagement_students =
  get_unique_students(daily_engagement)
len(unique_engagement_students)
```

Using NumPy and Pandas

- NumPy와 Pandas는 수치분석 및 데이터 분석을 위한 쉬운 도구를 제공해준다. (=> compared to R or Matlab)

```
import pandas as pd
daily_engagement =
    pd.read_csv('daily_engagement.csv')
len(daily_engagement['acct'].unique())
```

Using NumPy and Pandas

- NumPy는 데이터의 연산에 도움을 준다 (지난 시간에 간단하게 살펴본 내용)

```
import numpy as np
total_minutes =
    total_minutes_by_account.values()
print('Mean:',
    np.mean(list(total_minutes)))
print('Standard deviation:',
    np.std(list(total_minutes)))
```

NumPy

- Numpy는 Numerical Python의 약자로 이름에서 알수 있듯이 파이썬에서 과학적 계산을 하기 위해 수치연산기능을 제공함.
- 고성능 다차원 배열 객체와 이들과 함께 사용할 수 있는 다양한 수치연산 메소드를 제공하여 파이썬에서 Matlab 혹은 R의 기능을 사용할 수 있게 함.

NumPy

- NumPy는 고성능 연산을 위해 자체적으로 데이터구조를 제공하는데 파이썬이 기본적으로 제공하는 데이터구조와 유사점/차이점은 다음과 같다.
- 유사점
  - index를 사용하여 요소에 접근할 수 있다.

    ```
    a = ['a', 'b', 'c', 'd', 'e']
    a[3] → ‘d’
    ```

  - range를 사용하여 요소에 접근할 수 있다.

    ```
    a[1:3] → ['b', 'c']
    ```

  - loop를 사용할 수 있다

    ```
    for x in a:
    ```

NumPy

- 차이점
  - 하나의 array에는 같은 type의 데이터만 담을 수 있다.
    - array can holds `string`, `int`, `float64`, `boolean`, etc.
  - array와 함께 사용할 수 있는 손쉬운 수치연산 메소드 들을 제공한다.
    - `std()`, `mean()`, `log()`, `sin()`, etc.
  - 다차원의 array를 만들 수 있다.
    - 2D Array, 3D Array, etc.

Questions?