Trường ĐH Khoa Học Tự Nhiên Tp. Hồ Chí Minh
TRUNG TÂM TIN HỌC

Bài 4: Thư viện Pandas

Phòng LT & Mạng

2020

Nội dung

- Giới thiệu
- Series
- DataFrame

Giới thiệu

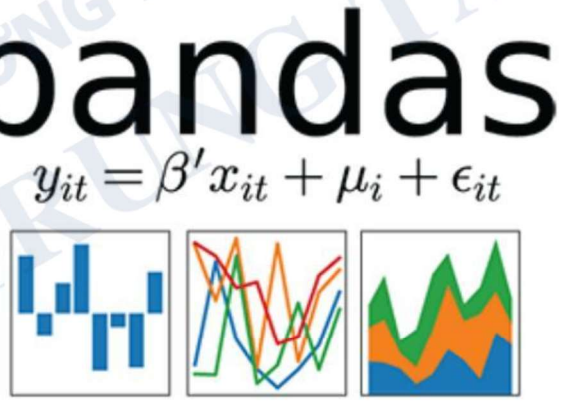
- **Pandas**
  - Là thư viện Python mã nguồn mở dùng để làm sạch, phân tích, khám phá, trực quan hóa và chuyển đổi dữ liệu.
  - Có tốc độ xử lý cao, dễ thao tác, rất hữu ích cho các chuyên gia phân tích tiếp thị, khoa học dữ liệu
  - Python với Pandas được sử dụng trong nhiều lĩnh vực bao gồm khoa học, kinh tế, phân tích thống kê…
  - Cài đặt: **pip install pandas**

https://pandas.pydata.org/

Giới thiệu

- **Pandas có 2 kiểu cấu trúc dữ liệu:**
  - Series
  - DataFrame

- **Các cấu trúc dữ liệu này được xây dựng trên Numpy array, có tốc độ xử lý nhanh.**



Giới thiệu

- **Ưu điểm**
  - Hỗ trợ đa dạng dữ liệu
  - Tích hợp dữ liệu
  - Chuyển đổi dữ liệu
  - Hỗ trợ dữ liệu time-series
  - Thống kê mô tả

Giới thiệu

Cấu trúc dữ liệu

- Có thể xem Data Frame là một container của các Series

Data Structure	Dimension	Mô tả
Series	1	1D array là đối tượng cấu trúc chứa một cột duy nhất
Data Frame	2	2D array là đối tượng cấu trúc bảng lưu trữ dữ liệu theo hàng và cột (như bảng tính - spreadsheet)

Ghi chú: DataFrame được sử dụng rộng rãi và là một trong những cấu trúc dữ liệu quan trọng nhất.

Giới thiệu

Ví dụ 1: Series

```
# tạo series age_ser là tuổi của 5 nhân viên
age_ser = pd.Series([25,27,24,28,30])
```

```
print(age_ser)
```

```
0    25
1    27
2    24
3    28
4    30
dtype: int64
```

```
print(type(age_ser)) # kiểu Series
```

```
<class 'pandas.core.series.Series'>
```

```
print(age_ser.dtype) # kiểu dữ liệu
```

```
int64
```

Giới thiệu

❑ Ví dụ 2: DataFrame

```
# tạo dataframe df gồm 5 nhân viên, có các cột: name, age, score
df = pd.DataFrame({'name':['Tom','Mike','Rose','Bill','Dick'],
                   'age':[25,27,24,28,30], 'score':[20.0,18.9,16.5,17.0,18.7]})
```

```
print(df)
```

```
   name  age  score
0   Tom   25   20.0
1  Mike   27   18.9
2  Rose   24   16.5
3  Bill   28   17.0
4  Dick   30   18.7
```

```
print(type(df)) # kiểu DataFrame
print(df.dtypes) # kiểu dữ liệu cột
```

```
<class 'pandas.core.frame.DataFrame'>
name      object
age        int64
score    float64
dtype: object
```