LAB 9: Thư viện Pandas trong Python

Thư viện pandas trong python là một thư viện mã nguồn mở, hỗ trợ đắc lực trong thao tác dữ liệu. Đây cũng là bộ công cụ phân tích và xử lý dữ liệu mạnh mẽ của ngôn ngữ lập trình python.

Tại sao lại dùng thư viện pandas?

- *DataFrame đem lại sự linh hoạt và hiệu quả trong thao tác dữ liệu và lập chỉ mục;*
- *Là một công cụ cho phép đọc/ ghi dữ liệu giữa bộ nhớ và nhiều định dạng file: csv, text, excel, sql database, hdf5;*
- *Liên kết dữ liệu thông minh, xử lý được trường hợp dữ liệu bị thiếu. Tự động đưa dữ liệu lộn xộn về dạng có cấu trúc;*
- *Dễ dàng thay đổi bố cục của dữ liệu;*
- *Tích hợp cơ chế trượt, lập chỉ mục, lấy ra tập con từ tập dữ liệu lớn.*
- *Có thể thêm, xóa các cột dữ liệu;*
- *Tập hợp hoặc thay đổi dữ liệu với group by cho phép bạn thực hiện các toán tử trên tập dữ liệu;*
- *Hiệu quả cao trong trộn và kết hợp các tập dữ liệu;*
- *Lập chỉ mục theo các chiều của dữ liệu giúp thao tác giữa dữ liệu cao chiều và dữ liệu thấp chiều;*

a. Cài đặt thư viện Pandas

- Sử dụng pip và gõ lệnh: `pip install pandas`
- Hoặc bằng Anaconda, dùng lệnh: `conda install pandas`

b. Khai báo thư viện Pandas

```
import pandas as pd
```

c. Thao tác với cấu trúc dữ liệu

Đọc csv file vào dataframe: dùng hàm read_csv():

```
Ví dụ: sinhvien_df = pd.read_csv('./sinhvien2020.csv')
```

In ra n bản ghi đầu tiên của dataframe sử dụng hàm head(). Ngược lại của hàm head là hàm tail():

```
Ví dụ: sinhvien_df.head(5)
```

- **Một vài tham số của hàm read_csv như:**

encoding: chỉ định encoding của file đọc vào. Mặc định là utf-8.

sep: thay đổi dấu ngăn cách giữa các cột. Mặc định là dấu phẩy (‘,’)

header: chỉ định file đọc có header(tiêu đề các cột) hay không. Mặc định là infer.
index_col: chỉ định chỉ số cột nào là cột chỉ số(số thứ tự). Mặc định là None.
nrows: chỉ định số bản ghi sẽ đọc vào. Mặc định là None – đọc toàn bộ.

Pandas có 3 cấu trúc dữ liệu cơ bản là:

- Series (1 chiều)
- DataFrame (2 chiều).
- Panel (3 chiều)

❖ Series

```
Series([data, index, dtype, name, copy, . . . ])
```

Series có thể được khởi tạo thông qua NumPy, kiểu Dict hoặc các dữ liệu vô hướng bình thường. Series có nhiều thuộc tính như index, array, values, dtype, v.v.
Chuyển đổi Series sang dạng dtype xác định, tạo bảng copy, trả về dạng bool của một thành phần, chuyển Series từ DatetimeIndex sang PeriodIndex, v.v.

Ví dụ 1: Không truyền index

```
import pandas as pd
s = pd.Series([0,1,2,3])
print(s)
```

Ví dụ 2: Có truyền index

```
import pandas as pd
s = pd.Series([0,1,2,3], index=["a","b","c","d"])
print(s)
```

Ví dụ 3: Tạo Series từ dict

```
import pandas as pd
data = {'a' : -1.3, 'b' : 11.7, 'd' : 2.0, 'f': 10, 'g': 5}
ser = pd.Series(data,index=['a','c','b','d','e','f'])
print(ser)
```

❖ DataFrame

```
DataFrame([data, index, columns, dtype, copy])
```

Dataframe là cấu trúc dữ liệu được gắn nhãn hai chiều với các cột và hàng như bảng tính (spreadsheet) hoặc bảng (table). Giống như Series, DataFrame có thể chứa bất kỳ loại dữ liệu nào.

Ví dụ: Tạo DataFrame từ dict các Series

```
import pandas as pd
# tạo dict từ các series
s ={'một':pd.Series([1.,2.,3.,5.],index=['a','b','c','e']),
    'hai':pd.Series([1. 2.,3.,4.],index=['a','b','c','d'])}

# tạo DataFrame từ dict
df = pd.DataFrame(s)
print(df)
```

THỰC HÀNH

Câu 1: Tạo 1 pandas DataFrame có 40 phần tử ngẫu nhiên, giá trị <200 và có 4 cột ABCD

a. In ra dataframe

b. In ra hàng có giá trị tăng dần

c. Thêm 1 cột E có giá trị ngẫu nhiên < 300 vào DataFrame và in ra DataFrame mới

d. In ra 3 cột B, C và E

Giải mẫu:

```
import pandas as pd
import numpy as np

df = pd.DataFrame(np.random.randint(200, size=(10, 4)), columns=list('ABCD'))
#a
print(df)

#b
dong_tang= df[(df.A < df.B) & (df.B < df.C) &(df.C < df.D)]
print(dong_tang)
#c
cot =np.random.randint(1,300)
df["E"]=cot
print(df)
# d
print(df.drop(["A","D"], axis=1))
```

Câu 2. Nhập dữ liệu từ file k2020.csv (file kèm)

a. In dữ liệu ra màn hình
b. In 5 dòng đầu tiên và 5 dòng cuối cùng của dữ liệu ra màn hình
c. Thống kê xem lớp có bao nhiêu bạn điểm loại giỏi (điểm từ 8 trở lên)
d. Thông kê xem lớp có bao nhiêu bạn trượt môn (điểm dưới 4 hoặc không có điểm)