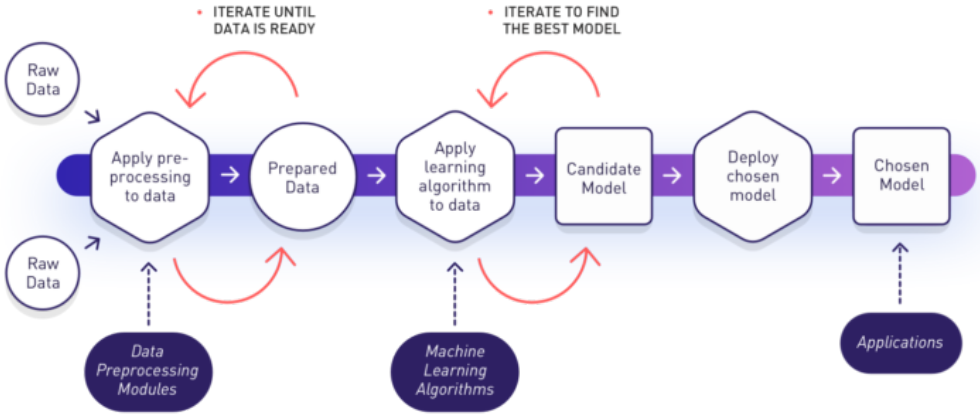
Đọc dữ liệu vào `DataFrame`

23-12-2020

1 Giới thiệu



Quy trình làm việc với dữ liệu

Các bước trong quá trình tiền xử lý dữ liệu bao gồm:

- Làm sạch dữ liệu (data cleaning).
- Chuyển đổi dữ liệu (data transformation).
- Thu gọn dữ liệu (data reduction)

Nhưng trước khi bắt đầu tiền xử lý dữ liệu, chúng ta phải nhập (import) dữ liệu.

2 Nhập dữ liệu

2.1 Định dạng `csv`

CSV (**C**omma **S**eparated **V**alues) là một loại định dạng văn bản đơn giản mà trong đó, các giá trị được ngăn cách với nhau bằng dấu phẩy. Định dạng CSV thường xuyên được sử dụng để lưu các bảng tính quy mô nhỏ như danh bạ, danh sách lớp, báo cáo...

Thông thường, một file `csv` có đuôi là `.csv`.

2.2 Cách đọc file `.csv` trong pandas.

Để đọc file `.csv` trong python, bạn có thể dùng đến hàm `pandas.read_csv()` như sau :

```
pandas.read_csv(<đường_dẫn_đến_file>, [các_tham_số_khác])
```

Trong đó, đường_dẫn_đến_file có thể là đường dẫn một tập tin trong máy (local) hoặc là một url (remote).

```
In [1]: # ví dụ :
        import pandas as pd

        df = pd.read_csv('F:\\Jupyter\\Datasets\\HNX.csv')
        df.head()
```

```
Out[1]:   Ticker        Date  Open  High   Low  Close   Volume
        0    AAV  2018-12-21  10.0  10.2  10.0   10.2    20550
        1    ACB  2018-12-21  29.5  29.5  29.1   29.4  1670803
        2    ACM  2018-12-21   0.8   0.8   0.7    0.7    25700
        3    ALV  2018-12-21   2.1   2.2   2.1    2.2     5200
        4    AMC  2018-12-21  21.5  21.7  19.1   19.1      300
```

2.3 Một số tham số của pandas.read_csv()

Vì dữ liệu thô có muôn hình vạn trạng, nên pandas.read_csv() cung cấp đến hơn 40 tham số để giúp đỡ quá trình đọc dữ liệu.

Tuy nhiên, chúng ta cùng xem qua một số tham số thường dùng.

2.3.1 Tham số sep

Tham số này dùng để chỉ ra cách các dữ liệu được phân tách như thế nào. Mặc định là dấu phẩy ','.

```
In [2]: df1 = pd.read_csv(
            'https://raw.githubusercontent.com/Levytan/MIS.2019/master/Data/sample_da
        )
        df1
```

```
Out[2]:   Region;Age;Income;Online Shopper
        0                 India;49;86400;No
        1               Brazil;32;57600;Yes
        2                  USA;35;64800;No
        3                Brazil;43;73200;No
        4                       USA;45;;Yes
        5                India;40;69600;Yes
        6                  Brazil;;62400;No
        7                India;53;94800;Yes
        8                  USA;55;99600;No
        9                India;42;80400;Yes
```

2.3.2 Tham số header

Tham số này được dùng để chỉ ra dòng (một hoặc nhiều) được dùng để làm header (tên cột). Giá trị mặc định là 'infer' (tự suy ra từ file).

```
In [3]: # Ví dụ, dùng None để báo rằng file .csv không có header
        df2 = pd.read_csv(
```

```
        'https://raw.githubusercontent.com/Levytan/MIS.2019/master/Data/sample_da
        header = None
    )
    df2
```

Out[3]:

	0	1	2	3
0	India	49.0	86400.0	No
1	Brazil	32.0	57600.0	Yes
2	USA	35.0	64800.0	No
3	Brazil	43.0	73200.0	No
4	USA	45.0	NaN	Yes
5	India	40.0	69600.0	Yes
6	Brazil	NaN	62400.0	No
7	India	53.0	94800.0	Yes
8	USA	55.0	99600.0	No
9	India	42.0	80400.0	Yes

2.3.3 Tham số `usecols`

Tham số này dùng này chỉ ra cần đọc từ file `.csv`, Có thể dùng tên cột (nếu có) hoặc số thứ tự cột.

```
In [4]: # lấy các cột có số thứ tự 0, 1, 2, 3
        df3 = pd.read_csv('https://raw.githubusercontent.com/Levytan/MIS.2019/master/
        df3
```

Out[4]:

	ID	Method	a1	a2
0	1	Soil	2	3
1	2	Soil	6	9
2	3	Soil	5	2
3	1	Soil	5	0
4	2	Soil	4	0

```
In [5]: # lấy các cột có tên ID, Method, a1, a3, a5
        df4 = pd.read_csv(
            'https://raw.githubusercontent.com/Levytan/MIS.2019/master/Data/sample_da
            usecols = ['ID', 'Method', 'a1', 'a3', 'a5']
        )
        df4
```

Out[5]:

	ID	Method	a1	a3	a5
0	1	Soil	2	4	0
1	2	Soil	6	2	0
2	3	Soil	5	4	0
3	1	Soil	5	0	4
4	2	Soil	4	0	6

2.3.4 Tham số `index_col`

Tham số này dùng để chỉ ra cột (một hoặc nhiều) dùng làm index. Giá trị mặc định là None, nghĩa là không dùng cột nào.

```
In [6]: df5 = pd.read_csv(
            'https://raw.githubusercontent.com/Levytan/MIS.2019/master/Data/sample_da
            index_col = 'ID'
        )
        df5
```

Out[6]:

ID	Method	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11
1	Soil	2	3	4	5	0	3	6	4	8	0	0
2	Soil	6	9	2	7	0	4	3	4	4	0	0
3	Soil	5	2	4	9	0	1	1	2	3	0	0
1	Soil	5	0	0	0	4	7	8	2	1	3	4
2	Soil	4	0	0	0	6	3	8	7	2	2	1

2.3.5 Tham số `na_values`

Được dùng để chỉ ra thêm giá trị `NaN` trong file `.csv`.

Mặc định, các giá trị sau sẽ được hiểu là `NaN`: `'#N/A'`, `'#N/A N/A'`, `'#NA'`, `'-1.#IND'`, `'-1.#QNAN'`, `'-NaN'`, `'-nan'`, `'1.#IND'`, `'1.#QNAN'`, `'N/A'`, `'NA'`, `'NULL'`, `'NaN'`, `'n/a'`, `'nan'`, `'null'`.

```
In [7]: df5 = pd.read_csv(
            'https://raw.githubusercontent.com/Levytan/MIS.2019/master/Data/sample_da
        )
        df5
```

Out[7]:

	ID	Method	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11
0	1	Soil	2	3	4	5	0	3	Null	4	8	0	0
1	2	Soil	6	9	2	7	0	4	3	4	4	0	0
2	3	Soil	5	2	4	9	0	1	Null	2	3	0	0
3	1	Soil	5	0	0	0	4	7	8	2	1	3	4
4	2	Soil	4	0	0	0	Null	3	8	7	2	2	1
5	1	Soil	5	0	Null	0	4	7	8	2	1	3	4

```
In [9]: df6 = pd.read_csv(
            'https://raw.githubusercontent.com/Levytan/MIS.2019/master/Data/sample_da
            na_values = 'Null'
        )
        df6
```

Out[9]:

	ID	Method	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11
0	1	Soil	2	3	4.0	5	0.0	3	NaN	4	8	0	0
1	2	Soil	6	9	2.0	7	0.0	4	3.0	4	4	0	0
2	3	Soil	5	2	4.0	9	0.0	1	NaN	2	3	0	0
3	1	Soil	5	0	0.0	0	4.0	7	8.0	2	1	3	4
4	2	Soil	4	0	0.0	0	NaN	3	8.0	7	2	2	1
5	1	Soil	5	0	NaN	0	4.0	7	8.0	2	1	3	4

Bạn có thể đọc về các tham số khác của `pandas.read_csv()` tại đây.