

4.1 Đọc file định dạng Text

Tạo file text với nội dung như sau:

```
In [52]: 1 import pandas as pd
          2 #Tạo file text:
          3 # !cat ex1.csv
          4 # a, b, c, d, message
          5 # 1, 2, 3, 4, hello
          6 # 5, 6, 7, 8, world
          7 # 9, 10, 11, 12, foo
          8
```

Load file text vào 1 DataFrame với read_csv

```
In [53]: 1 df = pd.read_csv('ex1.csv')
          2 df
          3
```

```
Out[53]:
```

	a	b	c	d	message
0	1	2	3	4	hello
1	5	6	7	8	world
2	9	10	11	12	foo

Hoặc với read_table phải có dấu hiệu ngăn cách

Hoặc với read_table phải có dấu hiệu ngăn cách

```
In [54]: 1 df = pd.read_table('ex1.csv', sep = ',')
          2 df
```

Out[54]:

	a	b	c	d	message
0	1	2	3	4	hello
1	5	6	7	8	world
2	9	10	11	12	foo

Tạo file text như sau:

```
In [55]: 1 # !cat ex2.csv
          2 # 1, 2, 3, 4, hello
          3 # 5, 6, 7, 8, world
          4 # 9, 10, 11, 12, foo
          5
          6
```

```
In [56]: 1 f = pd.read_csv('ex2.csv', header=None)
          2 f
```

Out[56]:

	0	1	2	3	4
0	1	2	3	4	hello
1	5	6	7	8	world
2	9	10	11	12	foo

```
In [57]: 1 f = pd.read_csv('ex2.csv', names=['a', 'b', 'c', 'd', 'message'])
          2 f
```

```
Out[57]:
```

	a	b	c	d	message
0	1	2	3	4	hello
1	5	6	7	8	world
2	9	10	11	12	foo

```
In [58]: 1 names = ['a', 'b', 'c', 'd', 'message']
          2 f = pd.read_csv('ex2.csv', names=names, index_col = 'message')
          3 f
```

```
Out[58]:
```

	a	b	c	d
message				
hello	1	2	3	4
world	5	6	7	8
foo	9	10	11	12

Thực hành đọc file có chỉ số phân cấp như sau:

```
In [59]: 1 parsed = pd.read_csv('csv_mindex.csv', index_col = ['key1', 'key2'])
          2 parsed
```

```
Out[59]:
```

	value1	value2
key1 key2		
one a	1	2

Thực hành đọc file có chỉ số phân cấp như sau:

```
In [59]: 1 parsed = pd.read_csv('csv_mindex.csv', index_col = ['key1', 'key2'])
         2 parsed
```

Out[59]:

		value1	value2
key1 key2			
one	a	1	2
	b	3	4
	c	5	6
	d	7	8
two	a	9	10
	b	11	12
	c	13	14
	d	15	16

```
In [60]: 1 list(open('ex3.txt'))
```

```
Out[60]: ['\tA\tB\tC\n',
          'aaa\t-0.264438\t-1.026059\t-0.6195\n',
          'bbb\t0.927272\t0.302904\t-0.032399\n',
          'ccc\t-0.264273\t-0.386314\t-0.217601\n',
          'ddd\t-0.871858\t-0.348382\t1.100491']
```

```
In [61]: 1 result = pd.read_table('ex3.txt', sep = '\s+')
         2 result
```

Out[61]:

A	B	C
---	---	---

```
In [61]: 1 result = pd.read_table('ex3.txt', sep = '\s+')
         2 result
```

```
Out[61]:
```

	A	B	C
aaa	-0.264438	-1.026059	-0.619500
bbb	0.927272	0.302904	-0.032399
ccc	-0.264273	-0.386314	-0.217601
ddd	-0.871858	-0.348382	1.100491

```
In [62]: 1 f = pd.read_csv('ex4.csv', skiprows=[0, 2, 3])
         2 f
```

```
Out[62]:
```

	a	b	c	d	message
0	1	2	3	4	hello
1	5	6	7	8	world
2	9	10	11	12	foo

Missing data và các lệnh trong Table 6-2 trang 160

```
In [63]: 1 result = pd.read_csv('ex6.csv')
         2 result
```

```
Out[63]:
```

	something	a	b	c	d	message
0	one	1	2	3.0	4	NaN
1	two	5	6	NaN	8	world
2	three	9	10	11.0	12	foo

Missing data và các lệnh trong Table 6-2 trang 160

```
In [63]: 1 result = pd.read_csv('ex6.csv')
          2 result
```

Out[63]:

	something	a	b	c	d	message
0	one	1	2	3.0	4	NaN
1	two	5	6	NaN	8	world
2	three	9	10	11.0	12	foo

```
In [64]: 1 #Kiểm tra giá trị có rỗng không
          2 check = pd.isnull(result)
          3 check
```

Out[64]:

	something	a	b	c	d	message
0	False	False	False	False	False	True
1	False	False	False	True	False	False
2	False	False	False	False	False	False

```
In [65]: 1 result = pd.read_csv('ex6.csv',na_values = ['NULL'])
          2 result
```

Out[65]:

	something	a	b	c	d	message
0	one	1	2	3.0	4	NaN
1	two	5	6	NaN	8	world
2	three	9	10	11.0	12	foo

```
In [66]: 1 sentinels = {'message': ['foo', 'NA'],
2               'something': ['two']}
3 pd.read_csv('ex6.csv', na_values = sentinels)
```

Out[66]:

	something	a	b	c	d	message
0	one	1	2	3.0	4	NaN
1	NaN	5	6	NaN	8	world
2	three	9	10	11.0	12	NaN

```
In [67]: 1 #Xuất bảng tối dòng 2.
2 result = pd.read_csv('ex6.csv', nrows=2)
3 result
```

Out[67]:

	something	a	b	c	d	message
0	one	1	2	3.0	4	NaN
1	two	5	6	NaN	8	world

```
In [68]: 1 #kiểm tra kích thước của file chunk.
2 chunker = pd.read_csv('ex6.csv', chunksize=100)
3 chunker
```

Out[68]: <pandas.io.parsers.TextFileReader at 0x178f8b259a0>

```
In [69]: 1 iterator = pd.read_csv('ex6.csv', iterator=100 )
2 iterator
```

Out[69]: <pandas.io.parsers.TextFileReader at 0x178f8b25f40>

```
In [70]: 1 squeeze = pd.read_csv('ex6.csv', squeeze=0)
2 squeeze
```

```
In [70]: 1 squeeze = pd.read_csv('ex6.csv', squeeze=0)
2         squeeze
```

```
Out[70]:
```

	something	a	b	c	d	message
0	one	1	2	3.0	4	NaN
1	two	5	6	NaN	8	world
2	three	9	10	11.0	12	foo

```
In [71]: 1 encoding = pd.read_csv('ex6.csv', encoding = 'ascii')
2         encoding
```

```
Out[71]:
```

	something	a	b	c	d	message
0	one	1	2	3.0	4	NaN
1	two	5	6	NaN	8	world
2	three	9	10	11.0	12	foo

4.2 Lưu file định dạng Text

```
In [72]: 1 data = pd.read_csv('ex5.csv')
2         data
```

```
Out[72]:
```

	something	a	b	c	d	message
0	one	1	2	3.0	4	NaN
1	two	5	6	NaN	8	world
2	three	9	10	11.0	12	foo

4.2 Lưu file định dạng Text

```
In [72]: 1 data = pd.read_csv('ex5.csv')
          2 data
```

```
Out[72]:
```

	something	a	b	c	d	message
0	one	1	2	3.0	4	NaN
1	two	5	6	NaN	8	world
2	three	9	10	11.0	12	foo

```
In [73]: 1 f = data.to_csv('out.csv')
          2
```

4.3 Đọc file định dạng Excel

```
In [74]: 1 xls_file = pd.ExcelFile('data.xls')
          2 table = xls_file.parse('Sheet1')
          3 table
```

```
Out[74]:
```

xin	chao
-----	------

```
In [77]: 1 table1 = xls_file.parse('Sheet2')
          2 table1
```

```
Out[77]:
```

hello	hola
-------	------



Quit Logout

Files Running Clusters

Select items to perform actions on them.

Upload New ↕

<input type="checkbox"/> 0	/ TNHien / Xử lý số liệu thống kê / 24-11-2021	Name ↓	Last Modified	File size
	..		seconds ago	
<input type="checkbox"/>	19110315_TrinhNgocHien_BTuan6.ipynb	Running	4 minutes ago	40.3 kB
<input type="checkbox"/>	csv_mindex.csv		3 hours ago	110 B
<input type="checkbox"/>	data.xls		2 hours ago	9.51 kB
<input type="checkbox"/>	ex1.csv		3 hours ago	74 B
<input type="checkbox"/>	ex2.csv		3 hours ago	54 B
<input type="checkbox"/>	ex3.txt		2 hours ago	137 B
<input type="checkbox"/>	ex4.csv		2 hours ago	170 B
<input type="checkbox"/>	ex5.csv		2 hours ago	81 B
<input type="checkbox"/>	ex6.csv		24 minutes ago	81 B
<input type="checkbox"/>	out.csv		4 minutes ago	92 B