Data Analysis

Velen Kong

摘要

基于Python的数据分析入门笔记，希望能自学掌握基本的数据分析能力，毕竟是统计学的基础之一。

内容安排主要参考

(美) Wes McKinney 著. Python for Data Analysis:2nd Edition [M] USA: O'Reilly Media 2017

该书作者同时也是pandas库的作者，同时借用其在GitHub上的数据资料。

第三方库与Python入门

IPython & Jupyter

用到的库：ipython，jupyter，numpy，matplotlib，pandas，scipy，scikit-learn，statsmodels。

jupyter确实好用，主要扩展了$Tab, ?, *$的功能。

还有一些快捷键仅限IPython使用。

另外就是一些常用的Magic Command，比如

```
%run
%load
%debug
%xmode Plain # less detailed
%xmode Verbose # more detailed
%matplotlib inline # draw in Jupyter
import numpy as np
import matplotlib.pyplot as plt
plt.plot(np.random.rand(50).cumsum())
```

```
# ----- Day 0 -----
```

Python基础

#注释。

```
# this is comment
```

所有变量都是对象(object)。

基本的函数调用，利用函数批量处理。

```
def append_element(some_list, element):
    some_list.append(element)
    return

data = [1, 2, 3]
append_element(data, 4)
data
# Out: [1, 2, 3, 4]
```

变量赋值类似引用(reference)。

动态类型，利用type()和isinstance()进行类型检查。后者可以传入tuple代替逻辑或操作。

```
a = 1.5
isinstance(a, (int, float))
# Out: True
```

print配合format输出。

```
a = 1; b = 1.5
print('a is {1}, b is {0}'.format(type(b),
type(a)))
# Out: a is <class 'int'>, b is <class 'float'>
```

Python中的object拥有各自的属性(attributes)和方法(methods)，可通过getattr, hasattr, setattr操作。其中getattr可以直接使用返回对象，setattr不改变原class。

```
class A(object):
    def set(self, a, b):
        x = a
        a = b
        b = x
        print a, b

a = A()
c = getattr(a, 'set')
c(a='1', b='2')
# Out: 2 1
```

Numpy库