



本节知识点概述

使用pandas保存豆瓣短评数据

文件保存方法

文件处理——open函数

数据分析利器——pandas

“十分钟” 搞定pandas

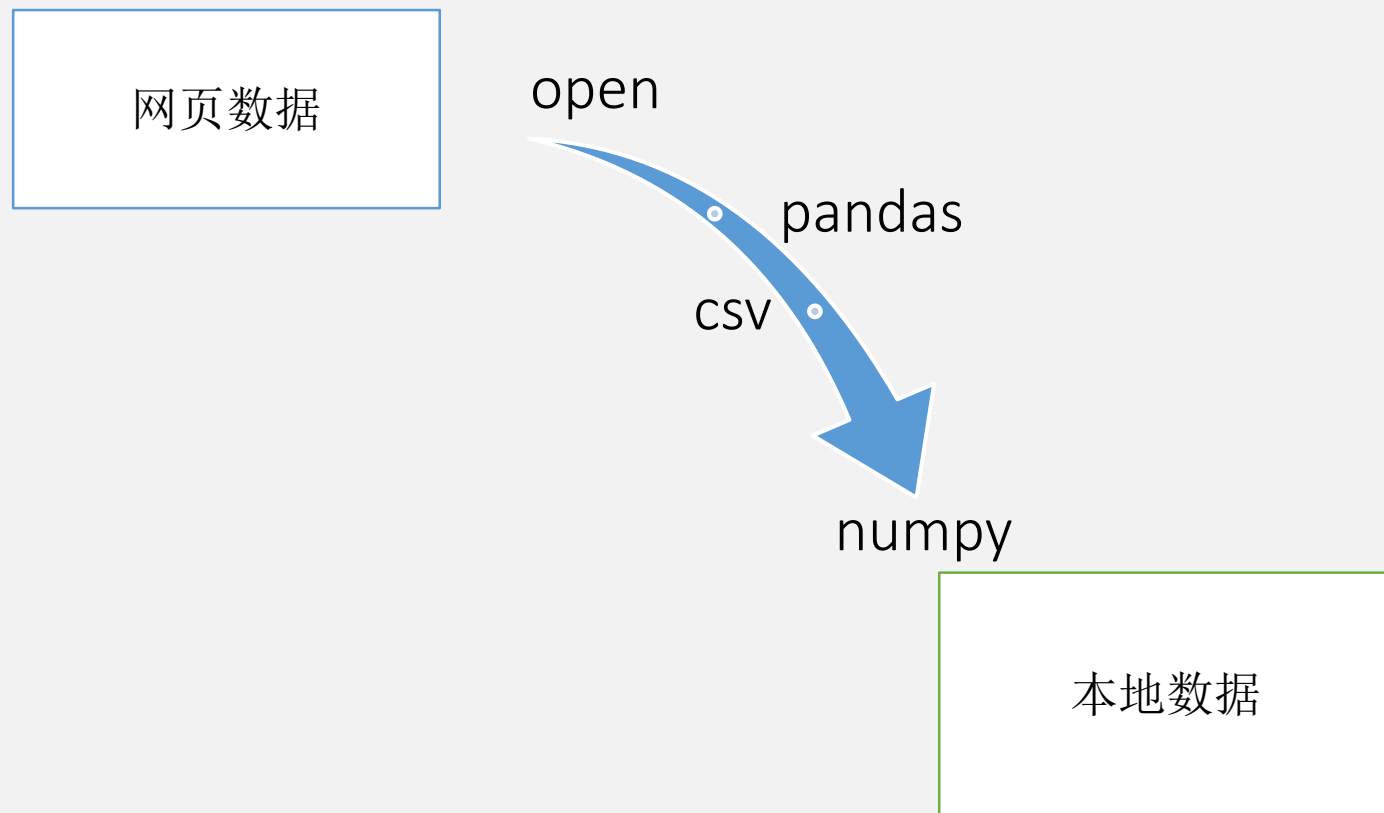
实战环节

使用pandas保存数据

造数与爬虫的对比



文件保存





open函数用法

```
import requests
from lxml import etree

url = 'https://book.douban.com/subject/1084336/comments/'
r = requests.get(url).text

s = etree.HTML(r)
file = s.xpath('//div[@class="comment"]/p/text()')

with open('pinglun.txt', 'w', encoding='utf-8') as f:
    for i in file:
        f.write(i)
```

新建对象 f

写入数据



DC学院
class.pkbigdata.com

造数

open函数的打开模式

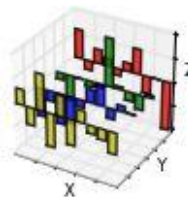
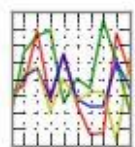
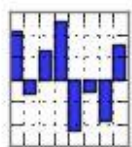
r	只读。若不存在文件会报错。
w	只写。若不存在文件自动新建。
a	附加到文件末尾。
rb, wb, ab	操作二进制文件
r+	读写模式打开



数据分析利器——pandas

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



pandas

numpy

matplotlib



DC学院
class.pkbigdata.com

造数



“十分钟”搞定pandas

[pandas中文文档](#)

10分钟了解pandas

这是对pandas的简短介绍，主要面向新用户。您可以在Cookbook中查看更复杂的手册

通常，我们导入如下：

```
In [1]: import pandas as pd
In [2]: import numpy as np
In [3]: import matplotlib.pyplot as plt
```

创建对象

请参见Data Structure Intro section

通过传递值列表创建Series，让pandas创建一个默认整数索引：

```
In [4]: s = pd.Series([1, 3, 5, np.nan, 6, 8])
In [5]: s
Out[5]:
0    1.0
1    3.0
2    5.0
3     NaN
4    6.0
5    8.0
dtype: float64
```



DC学院
class.pkbigdata.com

造数

pandas与Excel

Excel

读取和写入MS Excel

写入excel文件

```
In [145]: df.to_excel('foo.xlsx', sheet_name='Sheet1')
```

从excel文件读取

```
In [146]: pd.read_excel('foo.xlsx', 'Sheet1', index_col=None, na_values=['NA'])
```

Out[146]:

	A	B	C	D
2000-01-01	0.266457	-0.399641	-0.219582	1.186860
2000-01-02	-1.170732	-0.345873	1.653061	-0.282953
2000-01-03	-1.734933	0.530468	2.060811	-0.515536
2000-01-04	-1.555121	1.452620	0.239859	-1.156896
2000-01-05	0.578117	0.511371	0.103552	-2.428202
2000-01-06	0.478344	0.449933	-0.741620	-1.962409
2000-01-07	1.235339	-0.091757	-1.543861	-1.084753

to_excel()实例方法

用于将DataFrame保存到Excel

DataFrame 是一个表格或者类似二维数组的结构,它的各行表示一个实例,各列表示一个变量



pandas与Excel

```
import pandas as pd
import numpy as np

df = pd.DataFrame(np.random.randn(6,3))
print(df.head())

df.to_csv('numpppy.csv')
```

导入相关库

创建随机值

保存到Excel



DC学院
class.pkbigdata.com

造数



实战环节

```
import requests
from lxml import etree

url = 'https://book.douban.com/subject/1084336/comments/'
r = requests.get(url).text

s = etree.HTML(r)
file = s.xpath('//div[@class="comment"]/p/text()')

import pandas as pd
df = pd.DataFrame(file)
df.to_excel('pinglun.xlsx')
```

爬虫三步
你还记得吗



DC学院
class.pkbigdata.com

造数



实战环节

```
import pandas as pd  
df = pd.DataFrame(file)  
df.to_excel('pinglun.xlsx')
```

导入pandas

创建DataFrame

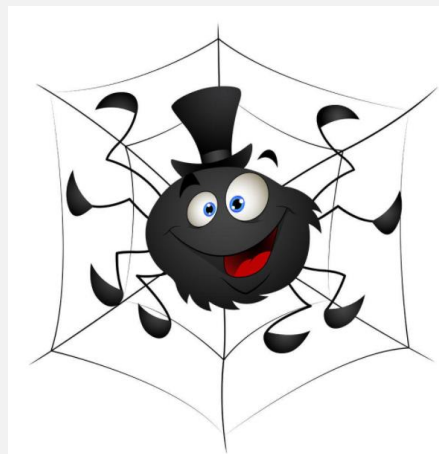
保存为Excel



造数与爬虫的对比



VS





造数与爬虫的对比

返回

正在爬取: 小王子 短评
查看使用说明

完成创建

眠去 ★★★★★ 2007-02-08
十几岁的时候渴慕着小王子，一天之间可以看四十四次日落。是在多久之后才明白，看四十四次日落的小王子，他有多么难过。

1461 有用

小岩井 2012-01-09
读了好多年，终于读完了，但是实在共鸣不起来，虽然知道那些道理，但真的觉得没什么了不起啊，是我还太幼稚吗？

784 有用

[已注销] ★★★★★ 2014-10-05
我早该猜到，在她那可笑的伎俩后面是缜密柔情。花朵是如此的天真无邪，可是，我实在太年轻了，不知道如何去爱她。

707 有用

渡 ★★★★★ 2012-09-01
我的玫瑰花儿，只有四个微不足道的刺，用来抵御这个世界。

526 有用

陈朝丞 ★★★★★ 2013-08-10

318 有用

作者: [法] 圣埃克苏佩里
原作名: Le Petit Prince
isbn: 702004249X
书名: 小王子
页数: 97
译者: 马振聘
定价: 22.00元
出版社: 人民文学出版社
装帧: 平装
出版年: 2003-8

咨询

已选中 1 列数据，爬取结果预览如下

匹配效果不理想？告诉我们这个问题

列-1	
十几岁的时候渴慕着小王子，一天...	
读了好多年，终于读完了，但是实...	
我早该猜到，在她那可笑的伎俩后...	

造数的可视化界面



造数与爬虫的对比

<https://book.douban.com/subject/1084336/comments/hot?p={{1-3}}>

2017 年 09 月 05 日 17:30:16

爬取 3 个网页，共 60 条数据
29 天后过期

在线预览

下载数据

造数的多页设置



DC学院
class.pkbigdata.com

造数



完成作业

- 使用csv保存数据
- 学习如何使用造数
- 思考如何在Python爬虫中翻页



DC学院
class.pkbigdata.com

造数

爬虫工程师

更多数据科学课程，上DC学院：class.pkbigdata.com



关注 DataCastle



关注造数

造数