

爬虫工程师（二）

——初识Python爬虫



DC学院
class.pkbigdata.com

造数

联合制作

造数



导师介绍



张世润（小歪）

- 【爱好者】喜欢Python，擅长爬虫，数据处理。
- 【创作者】知乎原创文章近百篇，拥有数千粉丝。
- 【学习者】爱学习的“萌新”，欢迎大家来交流。





本节知识点概述

初识Python爬虫

环境搭建

Python2与Python3的差异

Python下载和安装

使用谷歌浏览器

下载安装PyCharm

创建第一个实例

百度首页爬虫

工具比较

爬虫三步走

下载数据

解析数据

保存数据



DC学院
class.pkbigdata.com

造数



Python 2.x 和 3.x 版本差别？

1. 语法
2. 编码
3. Print用法
4. Xrange等一系列函数改变
5.



林灿斌

南山区程序员一个 94cb.com/

收录于 知乎周刊 · 815 人赞同了该回答

看到现在，我一直很好奇为什么会有人因为纠结学Py 3还是Py 2而浪费大量时间。

编程的话最重要的是编程思想，Python 3和Python 2虽然是两门完全不同的语言（故意黑），但是它的思想基本是共通的，只有少量的语法差异。**而编程中，语法只是细枝末节的东西。**

那么无论你学2还是3，都没有区别。会Python的人，一般2和3都会。当你学会了3，你只要稍微花上一点时间学习Python 2的语法，那么Python 2和Python 3这两门语言你也就都学会了——而认识语法差距花的时间，一般也不会比你纠结学哪个花的时间更多。

不要纠结学什么了，想到就去学，不要在这方面纠结太多时间，选Py 2或者Py 3并没有什么差异。





为什么选择Python3?

编码更加
简单

Python 2
只维护到
2020年

Python 3
是未来的
趋势



我使用Python3，因为Python2在未来停止维护，而Python3则是未来的趋势。——造数爬虫工程师小X
我使用Python3，主要是成本考虑。迁移成本，还有Python3的优化与改进。——造数后端工程师慢慢





环境搭建

下载Python

<https://www.python.org/downloads/release/python-362/>

Version		Operating System
Gzipped source tarball		Source release
XZ compressed source tarball		Source release
Mac OS X 64-bit/32-bit installer		Mac OS X
Windows help file		Windows
Windows x86-64 embeddable zip file		Windows
Windows x86-64 executable installer	64位	Windows
Windows x86-64 web-based installer		Windows
Windows x86 embeddable zip file		Windows
Windows x86 executable installer	32位	Windows
Windows x86 web-based installer		Windows

<http://pan.baidu.com/s/1kVsWdE3>



python-3.6.2 (32位) .exe

类型: 应用程序



python-3.6.2 (64位) .exe

类型: 应用程序



python-3.6.2-macosx10.6.pkg

类型: PKG 文件



DC学院
class.pkbdata.com

造数

其他工具

下载谷歌浏览器

<https://www.google.cn/chrome/browser/desktop/index.html>

(<http://pan.baidu.com/s/1bp2GgDh>)



应用推荐

<https://www.zhihu.com/question/20054116>



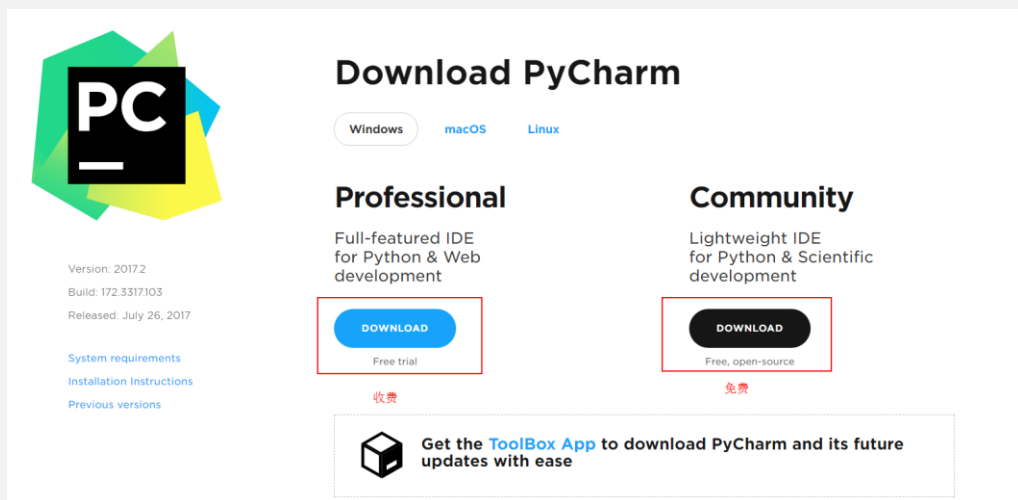


其他工具

下载PyCharm

<http://www.jetbrains.com/pycharm/download/#section=windows>

(<http://pan.baidu.com/s/1i4EzyGt>)



The screenshot shows the PyCharm download page. On the left is the PyCharm logo (a green and blue hexagon with 'PC' and a minus sign). Below it, text indicates 'Version: 2017.2', 'Build: 172.5317103', and 'Released: July 26, 2017'. There are links for 'System requirements', 'Installation instructions', and 'Previous versions'. The main section is titled 'Download PyCharm' and has tabs for 'Windows', 'macOS', and 'Linux'. Under 'Windows', there are two options: 'Professional' (Full-featured IDE for Python & Web development) and 'Community' (Lightweight IDE for Python & Scientific development). Each has a 'DOWNLOAD' button. Below the 'Professional' button is the text 'Free trial' and a red '收费' (Paid) label. Below the 'Community' button is the text 'Free, open-source' and a red '免费' (Free) label. At the bottom, there is a section for 'Get the Toolbox App to download PyCharm and its future updates with ease'.

风格和字体的调整

【File】 --> 【Settings】

```
import requests

r = requests.get('https://www.baidu.com/')
r.encoding = 'utf-8'
print(r.text)
```




创建第一个实例

urllib包

urllib是一个包，用于操作URL

<https://docs.python.org/3/library/urllib.html>

使用urllib包获取百度首页信息

```
>>> import urllib.request
>>> f =
urllib.request.urlopen('http://www.baidu.com/')
>>> f.read(500)
>>> f.read(500).decode('utf-8')
```





创建第一个实例

步骤详解

```
>>> import urllib.request
#导入urllib.request

>>> f = urllib.request.urlopen('http://www.baidu.com/')
#打开网址，返回一个类文件对象

>>> f.read(500)
#打印前500字符

>>> f.read(500).decode('utf-8')
#打印前500字符并修改编码为utf-8
```





创建第一个实例

Requests库

Requests是一个优雅而简单的HTTP库，适用于人类。

http://docs.python-requests.org/zh_CN/latest/user/quickstart.html

使用Requests库获取百度首页信息

```
>>> import requests
>>> r = requests.get('https://www.baidu.com/')
>>> r

>>> r.text

>>> r.encoding='utf-8'
>>> r.text
```



创建第一个实例

步骤详解

```
>>> import requests    #导入requests库

>>> r = requests.get('https://www.baidu.com/')
#使用requests.get方法获取网页信息

>>> r
>>> r.text    #打印结果

>>> r.encoding='utf-8'    #修改编码

>>> r.text    #打印结果
```





创建第一个实例

工具比较

把这两段代码在PyCharm中运行试试，你会发现什么？



urllib

```
import urllib.request
f = urllib.request.urlopen('http://www.baidu.com/')
print(f.read(500))

print(f.read(500).decode('utf-8'))
```

requests

```
import requests

r = requests.get('https://www.baidu.com/')
print(r)

print(r.text)

r.encoding = 'utf-8'
print(r.text)
```





爬虫三步走

爬虫的操作步骤

获取数据

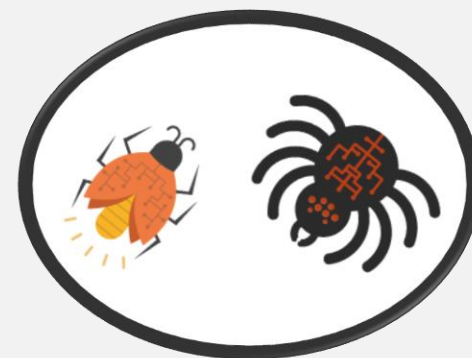
- Requests
- Urllib

解析数据

- Xpath
- BeautifulSoup4

保存数据

- 保存本地
- 数据库



DC学院
class.pkbigdata.com

造数



爬虫三步走

爬虫第一步：使用requests获得数据

导入
requests



requests.ge
t

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



DC学院
class.pkbigdata.com

造数



爬虫三步走

爬虫第二步：使用BeautifulSoup4解析数据

导入bs4



解析网页数据



寻找数据



for循环打印

```
from bs4 import BeautifulSoup
soup = BeautifulSoup(r,'lxml')
pattern = soup.find_all('p','comment-content')
for item in pattern:
    print(item.string)
```





爬虫三步走

爬虫第三步：使用pandas保存数据

导入pandas



新建list对象



使用to_csv写
入

```
import pandas
comments = []
for item in pattern:
    comments.append(item.string)
df = pandas.DataFrame(comments)
df.to_csv('comments.csv')
```





动手试试

一个完整的爬虫

```
#encoding='utf-8'

import requests

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

from bs4 import BeautifulSoup
soup = BeautifulSoup(r,'lxml')
pattern = soup.find_all('p','comment-content')
for item in pattern:
    print(item.string)

import pandas
comments = []
for item in pattern:
    comments.append(item.string)
df = pandas.DataFrame(comments)
df.to_csv('comments.csv')
```



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	0	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。												
2	1	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？												
3	2	我早该猜到，在她那可笑的伎俩后面是缜密柔情啊。花朵是如此的天真无邪，可是，我毕竟太年轻了，不知该如何去爱她。												
4	3	我的玫瑰花儿，只有四个微不足道的刺，用来抵御这个世界。												
5	4	谁能告诉我，小王子这种书有什么好看的？												
6	5	他像一颗树那样倒了下去												
7	6	不能理解的是，为什么它忽然红成这样？												
8	7	虽然我实在幼稚，但我并不怎么喜欢孩童般的纯净，我只爱风浪过后的平静，流水打磨出的清亮，大雪纷飞时的安宁，沧桑看透的纯真												
9	8	说实话 我看不太懂 但还是跟风给个5星吧 以显示我也是有思想有学识之人												
10	9	第一遍读时，我才4岁。等到真的读懂，才明白为什么这是一部“童话”。												
11	10	狐狸告诉小王子的秘密是：用心去看才看得清楚；是分离让小王子更思念他的玫瑰；爱就是责任。												
12	11	我老觉得小狐狸跟小王子是在搞GAY												
13	12	It is the time you have wasted for your rose that makes your rose so important.												
14	13	痛苦迷茫不是因为成为了“可笑”的大人，而是成为大人却没有真正长大。所以回过头来想要从怀念童年中解脱缓解痛苦那是本末倒置的												
15	14	爱屋及乌，爱一个人会让周围的一切变得美好，连麦浪的金黄都会让人心醉。爱同样意味着责任。因为“你现在要对												
16	15	漫山遍野的玫瑰，但真的，我最喜欢最初的那一朵，带刺儿的那一朵。我能不能回去继续浇灌那朵玫瑰？												
17	16	长这么大，读过次数最多的书就是《小王子》，我是那么的爱这本书，可是我都没有一本属于我的小王子，每次想读了就去图书馆借他												
18	17	其实我觉得小王子自私不负责任，没有看懂玫瑰，而且被勒地把小狐狸伤害了，小王子看懂很多人和事，却到最后才明白爱情的本质。												
19	18	原来在我不懂爱情的时候就爱上了你												
20	19	不知道第几次重读。每过一段时间再读，都有新的收获。心变得很柔软，脑里的迷雾被驱散。更多的关注他人，关心这个世界，自私是												
21														
22														
23														
24														
25														
26														
27														
28														
29														
30														
31														



完成作业

- 安装Python和PyCharm
- 爬取百度首页
- 尝试使用不同的代码编辑器
- 爬取豆瓣图书短评，你发现哪些问题



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