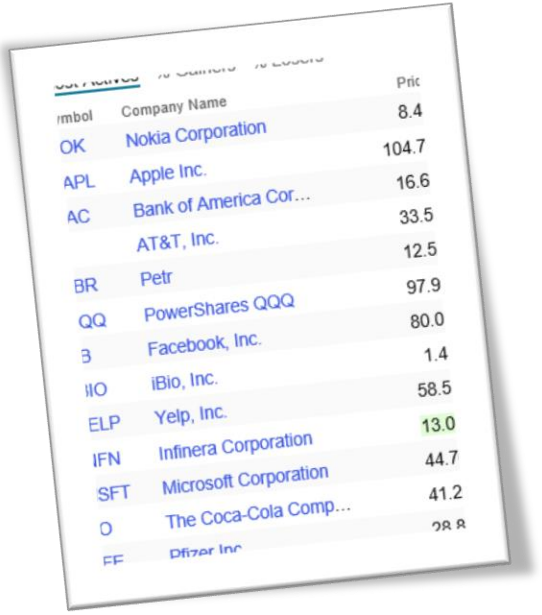
Where are data from?

网络数据获取入门

Zhang Li
Nanjing University

网络数据爬取

用Python获取网络数据



mbol	Company Name	Pric
OK	Nokia Corporation	8.4
APL	Apple Inc.	104.7
AC	Bank of America Cor...	16.6
	AT&T, Inc.	33.5
BR	Petr	12.5
QQ	PowerShares QQQ	97.9
3	Facebook, Inc.	80.0
IIO	iBio, Inc.	1.4
ELP	Yelp, Inc.	58.5
IFN	Infinera Corporation	13.0
SFT	Microsoft Corporation	44.7
O	The Coca-Cola Comp...	41.2
FE	Pfizer Inc	28.8

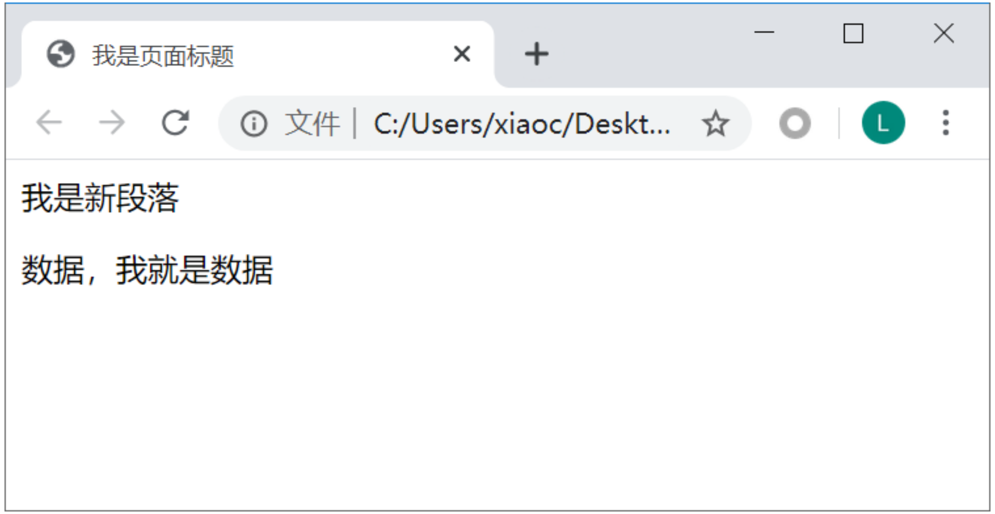
API获取数据

网络数据如何获取（爬取）？

抓取网页，解析网页内容

- 抓取
  - Requests第三方库
  - Scrapy框架
- 解析
  - Beautiful Soup库
  - re模块

html页面



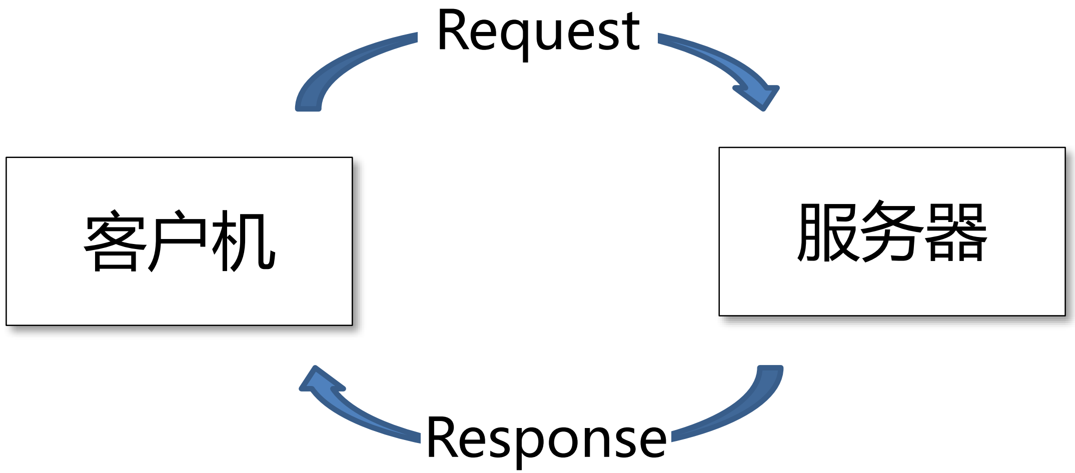
我是新段落

数据，我就是数据

```
<html>

<head>
<title>我是页面标题</title>
</head>

<body>
<p>我是新段落</p>
数据，我就是数据
</body>

</html>
```

参考：
http://www.w3school.com.cn/html/html_backgrounds.asp

网页抓取

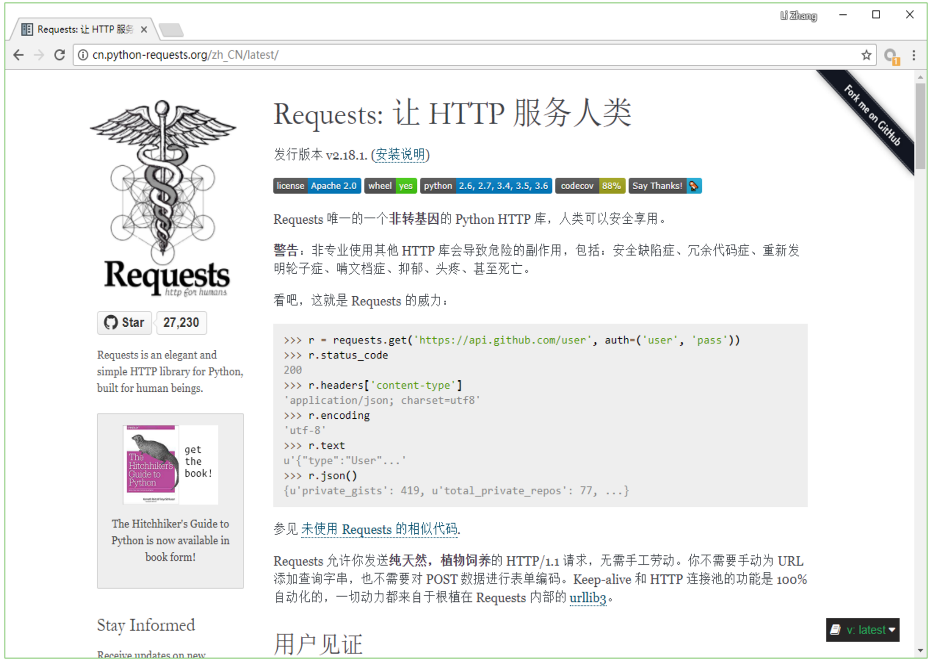


Requests库

- Requests库是更简单、方便和人性化的Python HTTP第三方库
- Requests官网：http://www.python-requests.org/

$ pip install requests

（Anaconda中预装）



Requests is an elegant and simple HTTP library for Python, built for human beings.

The Hitchhiker's Guide to Python is now available in book form!

Stay Informed

Receive updates on new

Requests: 让 HTTP 服务人类

发行版本 v2.18.1. (安装说明)

Requests 唯一的一个**非转基因**的 Python HTTP 库，人类可以安全享用。

警告：非专业使用其他 HTTP 库会导致危险的副作用，包括：安全缺陷症、冗余代码症、重新发明轮子症、啃文档症、抑郁、头疼、甚至死亡。

看吧，这就是 Requests 的威力：

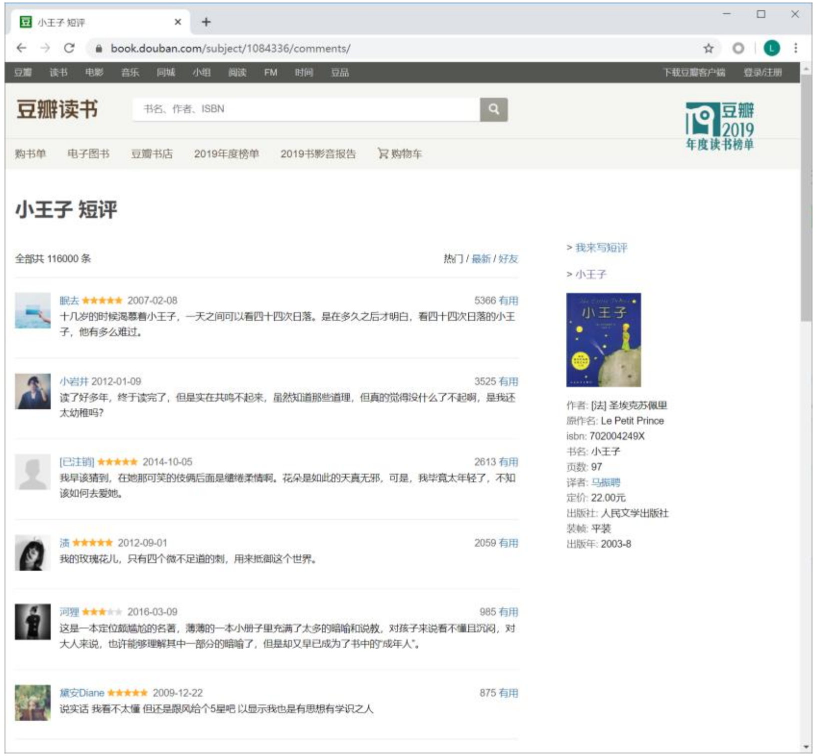
```
>>> r = requests.get('https://api.github.com/user', auth=('user', 'pass'))
>>> r.status_code
200
>>> r.headers['content-type']
'application/json; charset=utf8'
>>> r.encoding
'utf-8'
>>> r.text
u'{"type":"User"...'
>>> r.json()
{u'private_gists': 419, u'total_private_repos': 77, ...}
```

参见 未使用 Requests 的相似代码.

Requests 允许你发送**纯天然，植物饲养**的 HTTP/1.1 请求，无需手工劳动。你不需要手动为 URL 添加查询字串，也不需要对 POST 数据进行表单编码。Keep-alive 和 HTTP 连接池的功能是 100% 自动化的，一切动力都来自于根植在 Requests 内部的 urllib3。

用户见证

Requests库



小王子 短评
book.douban.com/subject/1084336/comments/
豆瓣 读书 电影 音乐 同城 小组 阅读 FM 时间 豆品
下载豆瓣客户端 登录/注册
豆瓣读书
书名、作者、ISBN
豆瓣 2019 年度读书榜单
购书单 电子图书 豆瓣书店 2019年度榜单 2019书影音报告 购物车

小王子 短评

全部共 116000 条　　热门 / 最新 / 好友

眠去 ★★★★★ 2007-02-08　　5366 有用
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。

小岩井 2012-01-09　　3525 有用
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?

[已注销] ★★★★★ 2014-10-05　　2613 有用
我早该猜到，在她那可笑的伎俩后面是缱绻柔情啊。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。

渍 ★★★★★ 2012-09-01　　2059 有用
我的玫瑰花儿，只有四个微不足道的刺，用来抵御这个世界。

河狸 ★★★☆☆ 2016-03-09　　985 有用
这是一本定位颇尴尬的名著，薄薄的一本小册子里充满了太多的暗喻和说教，对孩子来说看不懂且沉闷，对大人来说，也许能够理解其中一部分的暗喻了，但是却又早已成为了书中的"成年人"。

黛安Diane ★★★★★ 2009-12-22　　875 有用
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

> 我来写短评

> 小王子

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

豆瓣读书
《小王子》短评

Requests库

requests.get()	请求获取指定URL位置的资源，对应HTTP协议的GET方法，返回一个Response对象

Source

```
>>> import requests
>>> r = requests.get('https://www.nju.edu.cn')
>>> r.status_code
200
>>> print(r.text)
>>> r = requests.get('https://book.douban.com/subject/1084336/comments/')
>>> r.status_code
418
```

Requests

Source

```
>>> headers = {'User-Agent': 'Mozilla/5.0 (Macintosh: Intel Mac OS X 10_14_0) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/71.0.3578.98 Safari/537.36'}
>>> r = requests.get(url, headers = headers)  #定制请求头
>>> payload = {'key1': 'value1', 'key2': 'value2'}
>>> r = requests.get('http://httpbin.org/get', params = payload)
>>> r = requests.post(url, data = payload)
```

Requests库

```
>>> r.encoding          # 根据HTTP头部自动推测
'UTF-8'
>>> r.encoding = 'gb2312'
>>> r.encoding = r.apparent_encoding
>>> r.content           # 以字节方式访问Response对象
>>> r.json()
```

Requests库

r.content　　　　*# 以字节方式访问Response对象*

F ile

```
r = requests.get('http://...sample.jpg')
with open('pic.jpg', 'wb') as f:
    f.write(r.content)
```

JSON格式

- **JSON格式**
  - JavaScript Object Notation，JS对象标记)
  - 一种轻量级的数据交换格式

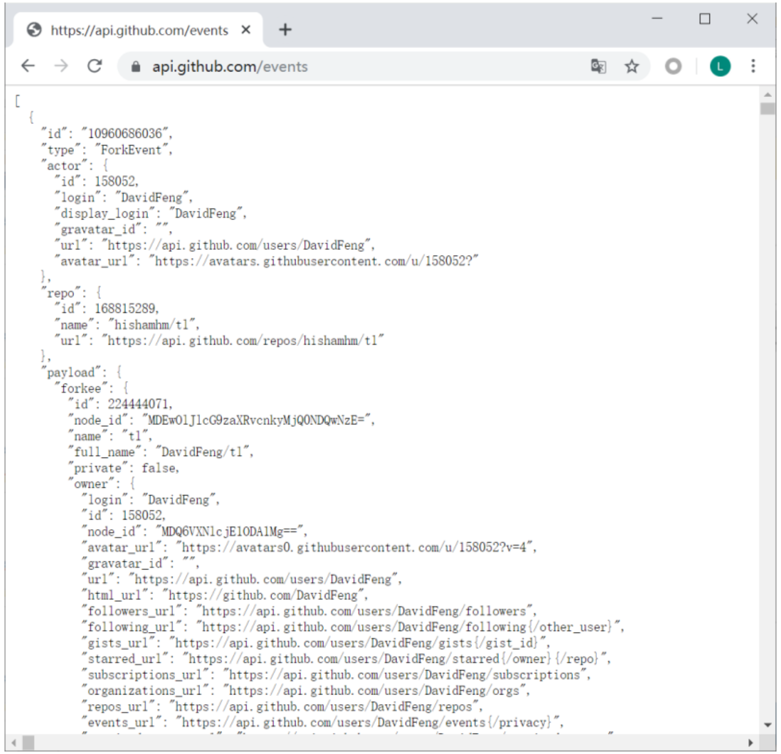
r.json()

'{"name":"Niuyun", "address":{"city": "Beijing", "street": "Chaoyang Road"}}'

解析后

```
>>> x = {"name":"Niuyun",
         "address":{"city":"Beijing","street":"Chaoyang Road"}
        }
>>> x['address']['street']
'Chaoyang Road'
```

JSON文件


https://api.github.com/events

api.github.com/events

```
[
  {
    "id": "10960686036",
    "type": "ForkEvent",
    "actor": {
      "id": 158052,
      "login": "DavidFeng",
      "display_login": "DavidFeng",
      "gravatar_id": "",
      "url": "https://api.github.com/users/DavidFeng",
      "avatar_url": "https://avatars.githubusercontent.com/u/158052?"
    },
    "repo": {
      "id": 168815289,
      "name": "hishamhm/tl",
      "url": "https://api.github.com/repos/hishamhm/tl"
    },
    "payload": {
      "forkee": {
        "id": 224444071,
        "node_id": "MDEwOlJlcG9zaXRvcnkyMjQ0NDQwNzE=",
        "name": "tl",
        "full_name": "DavidFeng/tl",
        "private": false,
        "owner": {
          "login": "DavidFeng",
          "id": 158052,
          "node_id": "MDQ6VXNlcjE1ODA1Mg==",
          "avatar_url": "https://avatars0.githubusercontent.com/u/158052?v=4",
          "gravatar_id": "",
          "url": "https://api.github.com/users/DavidFeng",
          "html_url": "https://github.com/DavidFeng",
          "followers_url": "https://api.github.com/users/DavidFeng/followers",
          "following_url": "https://api.github.com/users/DavidFeng/following{/other_user}",
          "gists_url": "https://api.github.com/users/DavidFeng/gists{/gist_id}",
          "starred_url": "https://api.github.com/users/DavidFeng/starred{/owner}{/repo}",
          "subscriptions_url": "https://api.github.com/users/DavidFeng/subscriptions",
          "organizations_url": "https://api.github.com/users/DavidFeng/orgs",
          "repos_url": "https://api.github.com/users/DavidFeng/repos",
          "events_url": "https://api.github.com/users/DavidFeng/events{/privacy}",
```

```
>>> r = requests.get('https://api.github.com/events')
>>> data = r.json()
>>> data[0]['id']
```

Robots协议

- Robots协议也称为爬虫协议，全称为爬虫排除协议（The Robots Exclusion Protocol）
- 检查站点根目录下是否存在robots.txt

Robots协议-豆瓣网

```
User-agent: *
Disallow: /subject_search
Disallow: /amazon_search
Disallow: /search
…
Disallow: /link2/
Disallow: /recommend/
Disallow: /doubanapp/card
Disallow: /update/topic/
Allow: /ads.txt
Sitemap: https://www.douban.com/sitemap_index.xml
Sitemap: https://www.douban.com/sitemap_updated_index.xml
# Crawl-delay: 5

User-agent: Wandoujia Spider
Disallow: /
…
```

网页数据解析

网页数据解析

- **Beautiful Soup**是一个可以从HTML或XML文件中提取数据的Python库
- 官方网站：https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Beautiful Soup

- $ pip install beautifulsoup4 （Anaconda中预装）

```
>>> import requests
>>> from bs4 import BeautifulSoup
>>> r = requests.get('https://book.douban.com/subject/1084336/comments/', headers = headers)
>>> soup = BeautifulSoup(r.text, 'lxml')
```

lxml: HTML解析器
$ pip install lxml

Python内置的HTML解析器
BeautifulSoup(markup, 'html.parser')

Beautiful Soup

- BeautifulSoup对象
  - Tag
  - NavigableString
  - BeautifulSoup
  - Comment

```
>>> markup = '<p class="title"><b>The Little Prince</b></p>'
>>> soup = BeautifulSoup(markup, 'lxml')
>>> soup.b
<b>The Little Prince</b>
```

标签内容访问方式
BeautifulSoup对象.Tag

Beautiful Soup

```
>>> markup = '<p class="title"><b>The Little Prince</b></p>'
>>> soup = BeautifulSoup(markup, 'lxml')
>>> tag = soup.p
>>> tag
<p class="title"><b>The Little Prince</b></p>
>>> tag.name
'p'
>>> tag['class']
['title']
>>> tag.attrs
{'class': ['title']}
>>> tag.string
'The Little Prince'
>>> soup.find_all('b')
[<b>The Little Prince</b>]
```

网页数据解析

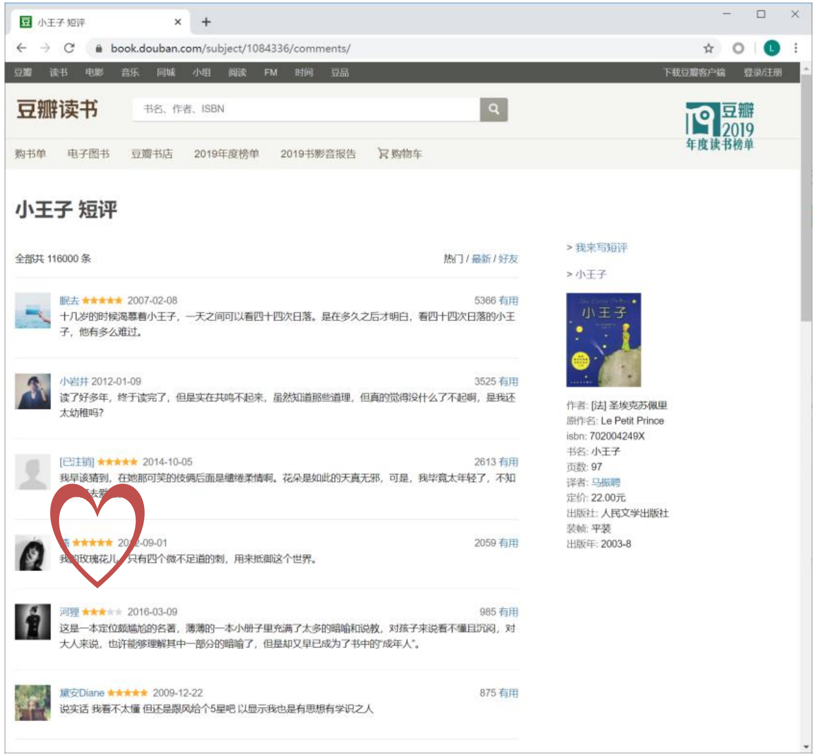
<span class="short">不知道第几次重读。每过一段时间再读，都有新的收获。心变得很柔软，脑里的迷雾被驱散。更多的关注他人，关心这个世界，自私是多么无趣的事情啊。我想，写一本能温暖人心，帮助困难的人们的书，比世界上很多事情都有意义。</span>

```
pattern = soup.find_all('span', {'class':'short'})
for item in pattern:
    print(item.string)
```

网页数据解析

```
r = requests.get('https://book.douban.com/subject/1084336/comments/', headers=headers)
soup = BeautifulSoup(r.text, 'lxml')
pattern = soup.find_all('span', {'class':'short'})
for item in pattern:
    print(item.string)
```

网页数据解析


小王子 短评
book.douban.com/subject/1084336/comments/
豆瓣 读书 电影 音乐 同城 小组 阅读 FM 时间 豆品
下载豆瓣客户端 登录/注册
豆瓣读书
书名、作者、ISBN
豆瓣 2019 年度读书榜单
购书单 电子图书 豆瓣书店 2019年度榜单 2019书影音报告 购物车

小王子 短评

全部共 116000 条　　　　热门 / 最新 / 好友

眠去 ★★★★★ 2007-02-08　　5366 有用
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。

小岩井 2012-01-09　　3525 有用
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？

[已注销] ★★★★★ 2014-10-05　　2613 有用
我早该猜到，在她那可笑的伎俩后面是缱绻柔情啊。花朵是如此的天真无邪，可是，我毕竟太年轻了，不知[illegible]去爱[illegible]

[illegible] ★★★★★ 20[illegible]2-09-01　　2059 有用
我[illegible]玫瑰花儿[illegible]只有四个微不足道的刺，用来抵御这个世界。

河狸 ★★★☆☆ 2016-03-09　　985 有用
这是一本定位颇尴尬的名著，薄薄的一本小册子里充满了太多的暗喻和说教，对孩子来说看不懂且沉闷，对大人来说，也许能够理解其中一部分的暗喻了，但是却又早已成为了书中的"成年人"。

黛安Diane ★★★★★ 2009-12-22　　875 有用
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

> 我来写短评

> 小王子

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

豆瓣读书
《小王子》推荐星级

网页数据解析

- 正则表达式是对字符串（包括普通字符和特殊字符）操作的一种逻辑公式
- **re**正则表达式模块进行各类正则表达式处理
- 参考网站：https://docs.python.org/3.5/library/re.html

元字符	描述
.	匹配除换行符外的任意字符
*	重复前面的子表达式0次或多次
+	重复前面的子表达式1次或更多次
?	重复前面的子表达式0次或1次
^	匹配字符串的开始
$	匹配字符串的结束
{n}	重复n次
{n, }	重复n次或更多次
{n, m}	重复n到m次
\b	匹配单词的开始或结尾即单词边界，“\B”匹配非单词边界
\d	匹配数字，“\D”匹配任意非数字字符
\s	匹配任意空白符，“\S”匹配任意非空白符
\w	匹配任意字母、数字或下划线的标识符字符，“\W”匹配任意非标识符字符
[a-z]	匹配指定范围内的任意字符
[^a-z]	匹配任何不在指定范围内的任意字符

网页数据解析

```
<span class="user-stars allstar50 rating" title="力荐"></span>
```

```
'<span class="user-stars allstar(.*?) rating"'
```

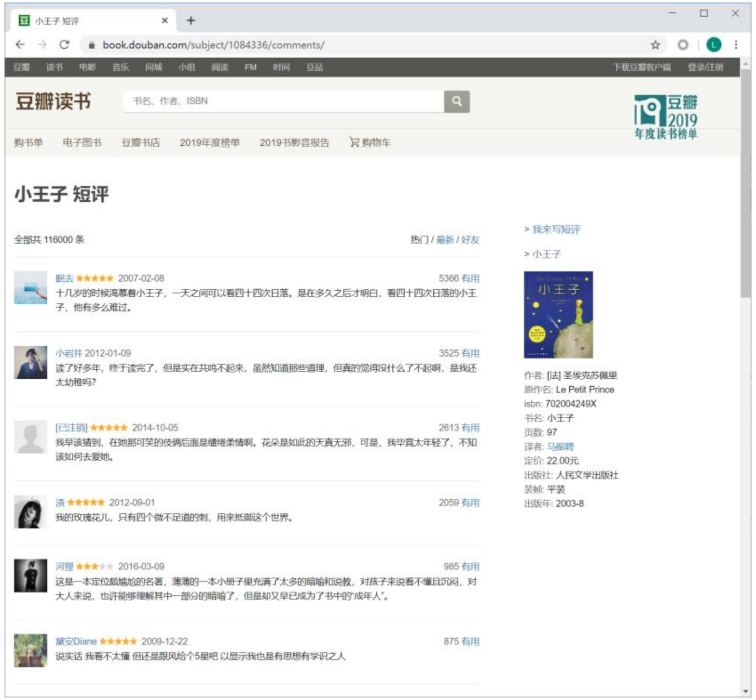
```
pattern = re.compile('<span class="user-stars allstar(.*?) rating"')
p = re.findall(pattern, r.text)
```

网页数据解析

```
r = requests.get('https://book.douban.com/subject/1084336/comments', headers = headers)
soup = BeautifulSoup(r.text, 'lxml')
pattern = soup.find_all('span', {'class':'short'})
for item in pattern:
    print(item.string)
```

```
pattern_s = re.compile('<span class="user-stars allstar(.*?) rating"')
p = re.findall(pattern_s, r.text)
```

思考：抓取图书短评前5页



小王子 短评

book.douban.com/subject/1084336/comments/

豆瓣 读书 电影 音乐 同城 小组 阅读 FM 时间 豆品 下载豆瓣客户端 登录/注册

豆瓣读书

购书单 电子图书 豆瓣书店 2019年度榜单 2019书影音报告 购物车

小王子 短评

全部共 116000 条 热门 / 最新 / 好友

眠去 ★★★★★ 2007-02-08 5366 有用
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。

小岩井 2012-01-09 3525 有用
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？

[已注销] ★★★★★ 2014-10-05 2613 有用
我早该猜到，在她那可笑的伎俩后面是缱绻柔情啊。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。

清 ★★★★★ 2012-09-01 2059 有用
我的玫瑰花儿，只有四个微不足道的刺，用来抵御这个世界。

河狸 ★★★☆☆ 2016-03-09 985 有用
这是一本定位颇尴尬的名著，薄薄的一本小册子里充满了太多的暗喻和说教，对孩子来说看不懂且沉闷，对大人来说，也许能够理解其中一部分的暗喻了，但是却又早已成为了书中的"成年人"。

黛安Diane ★★★★★ 2009-12-22 875 有用
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

> 我来写短评

> 小王子

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

https://book.douban.com/subject/1084336/

```
r = requests.get('https://book.douban.com/subject/1084336/comments/hot?p=' + str(i+1))
```

抓取例1

http://money.cnn.com/data/dow30/

抓取道指成分股数据并将30家公司的代码、公司名称和最近一次成交价放到一个列表中输出