

Requests库网络爬取实战

WS03



嵩天

www.python123.org

The Website is the API ...



Requests

自动爬取HTML页面
自动网络请求提交

robots.txt

网络爬虫排除标准



Projects

实战项目

掌握定向网络数据爬取和网页解析的基本能力

Python网络爬虫与信息提取

python
弹指之间 · 享受创新

04X -Tian

实例1：京东商品页面的爬取

实例2：亚马逊商品页面的爬取

实例3：百度/360搜索关键字提交

实例4：网络图片的爬取和存储

实例5：IP地址归属地的自动查询

实例1：京东商品页面的爬取

https://item.jd.com/2967929.html



iPhone 7

搜索

我的购物车 0

亿元红包 100元神券 新机专场 红米Note4X 荣耀8 乐S3 1元800M 美图T8

全部商品分类

首页

手机首页

新机发布

网上营业厅

京选卖家

配件中心

手机社区

手机 > 手机通讯 > 手机 > 华为 (HUAWEI) > 华为荣耀8

荣耀官方旗舰店 JD自营 联系供应商 JIMI 关注店铺



荣耀8 4GB+64GB 全网通4G手机 魅海蓝

双镜头，双2.5D玻璃，双功能指纹！【荣耀爆品领券立减200！点击查看~】

选择移动/联通/电信优惠购，流量话费送不停，惊喜不断，优惠多多！！

京东价 **¥2499.00** 降价通知

累计评价
31万+

促销 **加价购** 满100.00另加9.90元，或满200.00另加25.00元，或满300.00另加45.00元，即可在购物车换购热

销商品 详情 >>

支持 **以旧换新，闲置手机回收** **4G套餐18元起** **礼品购**

配送至 北京朝阳区管庄 有货 支持 99元免基础运费(50kg内) 货到付款 京准达

由 京东 发货，并提供售后服务。23:00前完成下单，预计明天(02月19日)送达

重量 0.473kg

选择颜色



流光金



珠光白



魅海蓝



幻夜黑



樱语粉

```
>>> import requests
>>> r = requests.get("https://item.jd.com/2967929.html")
>>> r.status_code
200
>>> r.encoding
'gbk'
>>> r.text[:1000]
'<!DOCTYPE HTML>\n<html lang="zh-CN">\n<head>\n    <meta http-equiv
="Content-Type" content="text/html; charset=gbk" />\n    <title>【华
为荣耀8】荣耀8 4GB+64GB 全网通4G手机 魅海蓝【行情 报价 价格 评测】-京东<
/title>\n    <meta name="keywords" content="HUAWEI荣耀8,华为荣耀8,华
为荣耀8报价,HUAWEI荣耀8报价"/>\n    <meta name="description" content=
"【华为荣耀8】京东JD.COM提供华为荣耀8正品行货，全国价格最低，并包括HUAWEI
荣耀8网购指南，以及华为荣耀8图片、荣耀8参数、荣耀8评论、荣耀8心得、荣耀8技
```

全代码

```
import requests
url = "https://item.jd.com/2967929.html"
try:
    r = requests.get(url)
    r.raise_for_status()
    r.encoding = r.apparent_encoding
    print(r.text[:1000])
except:
    print("爬取失败")
```

实例2：亚马逊商品页面的爬取

https://www.amazon.cn/gp/product/B01M8L5Z3Y



亚马逊
amazon.cn
免费试享Prime

图书

浏览全部商品分类

我的亚马逊 Z秒杀 礼品卡 我要开店 海外购 帮助 In English

您好, 登录我的帐户 免费试享Prime 购物车

图书 高级搜索 所有分类 新品排行榜 销售排行榜 新书店 教材教辅 少儿 文学 小说 历史 经管 励志 人文社科 生活 科普 进口图书 促销

< 返回结果



在线试读

极简
THE MORE OF LESS
在你拥有的一切之下,发现你想要的生活

极简:在你拥有的一切之下,发现你想要的生活 平装 – 2016年11月5日

乔舒亚·贝克尔 (Joshua Becker) (作者)

★★★★☆ 47 条商品评论 | 分享 | 自营

显示所有 格式和版本

平装
¥23.80

配送至: 北京东城区 现在有货

送达日期: 明天(2月19日), 请在9小时9分钟内下单并选择“快递送货上门”。
(精确送达时间请于结账页面查询)

销售配送: 由亚马逊直接销售和发货。

全新品15 售价从 ¥22.80起

退换承诺: 此商品支持30天免费退换 详情

售价: ¥23.80 (6.3折)
定价: ¥38.00

图书满¥59免运费且可货到付款

数量: 1

☐ 试享亚马逊Prime免费配送

加入购物车

登录 即可开启一键下单。

加入心愿单

亚马逊的其他卖家

```
>>> import requests
>>> r = requests.get("https://www.amazon.cn/gp/product/B01M8L5Z3Y")
>>> r.status_code
503
>>> r.encoding
'ISO-8859-1'
>>> r.encoding = r.apparent_encoding
>>> r.text
'<!--\n          To discuss automated access to Amazon data please cont
act api-services-support@amazon.com.\n          For information about m
igrating to our APIs refer to our Marketplace APIs at https://develop
er.amazonservices.com.cn/index.html/ref=rm_5_sv, or our Product Adver
tising API at https://associates.amazon.cn/gp/advertising/api/detail/
main.html/ref=rm_5_ac for advertising use cases.\n-->\n<html><head><m
eta http-equiv="Content-Type" content="text/html; charset=utf-8"><titl
e>亚马逊</title><body style="text-align:center;"><br><div style="width
:600px;margin:0 auto;text-align:left;"><h2>意外错误</h2></div><br><div
style="width:500px;margin:0 auto;text-align:left;"><font color="red">
抱歉，由于程序执行时，遇到意外错误，您刚刚操作没有执行成功，请稍后重试。或将
此错误报告给我们的客服中心：<a href="mailto:service_bj@cs.amazon.cn">ser
```

[illegible]

全代码

```
import requests
url = "https://www.amazon.cn/gp/product/B01M8L5Z3Y"
try:
    kv = {'user-agent': 'Mozilla/5.0'}
    r = requests.get(url, headers=kv)
    r.raise_for_status()
    r.encoding = r.apparent_encoding
    print(r.text[1000:2000])
except:
    print("爬取失败")
```

实例3：百度/360搜索关键词提交



<http://www.baidu.com>

百度一下



<http://www.so.com>

搜一下

搜索引擎关键词提交接口

百度的关键词接口：

<http://www.baidu.com/s?wd=keyword>

360的关键词接口：

<http://www.so.com/s?q=keyword>


```
>>> import requests
>>> kv = {'wd': 'Python'}
>>> r = requests.get("http://www.baidu.com/s", params=kv)
>>> r.status_code
200
>>> r.request.url
'http://www.baidu.com/s?wd=Python'
>>> len(r.text)
302829
```


百度搜索全代码

```
import requests
keyword = "Python"
try:
    kv = {'wd':keyword}
    r = requests.get("http://www.baidu.com/s",params=kv)
    print(r.request.url)
    r.raise_for_status()
    print(len(r.text))
except:
    print("爬取失败")
```

```
>>> import requests
>>> kv = {'q': 'Python'}
>>> r = requests.get('http://www.so.com/s', params=kv)
>>> r.status_code
200
>>> r.request.url
'https://www.so.com/s?q=Python'
>>> len(r.text)
228253
```

360搜索全代码

```
import requests
keyword = "Python"
try:
    kv = {'q':keyword}
    r = requests.get("http://www.so.com/s",params=kv)
    print(r.request.url)
    r.raise_for_status()
    print(len(r.text))
except:
    print("爬取失败")
```

实例4：网络图片的爬取和存储

网络图片的爬取

网络图片链接的格式：

<http://www.example.com/picture.jpg>

国家地理：<http://www.nationalgeographic.com.cn/>

选择一个图片Web页面：

http://www.nationalgeographic.com.cn/photography/photo_of_the_day/3921.html

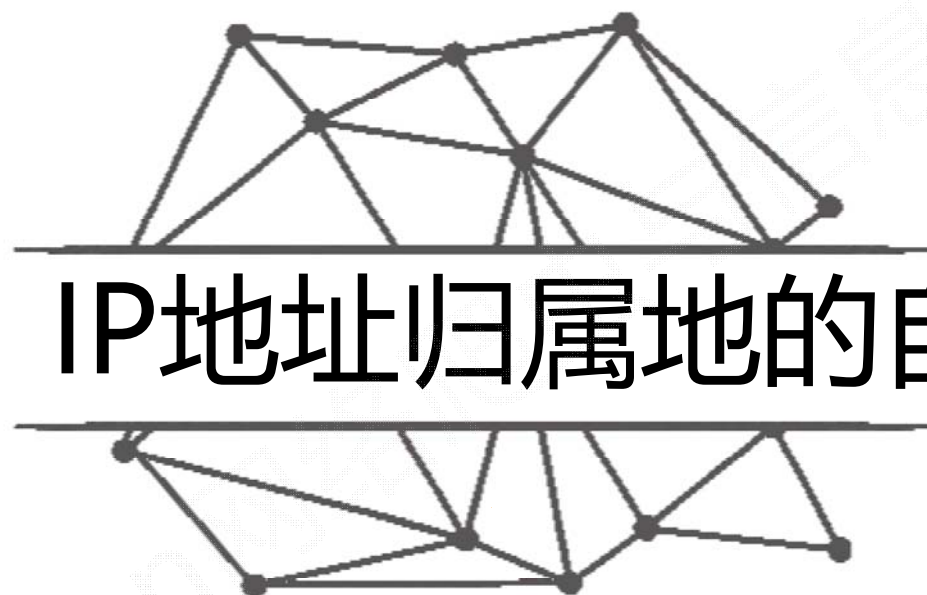
图片地址：<http://image.nationalgeographic.com.cn/2017/0211/20170211061910157.jpg>

```
>>> import requests
>>> path = "D://abc.jpg"
>>> url = "http://image.nationalgeographic.com.cn/2017/0211/20170211061910157.jpg"
>>> r = requests.get(url)
>>> r.status_code
200
>>> with open(path, 'wb') as f:
        f.write(r.content)
206476
>>> f.close()
```

图片爬取全代码

```
import requests
import os
url = "http://image.nationalgeographic.com.cn/2017/0211/20170211061910157.jpg"
root = "D://pics//"
path = root + url.split('/')[-1]
try:
    if not os.path.exists(root):
        os.mkdir(root)
    if not os.path.exists(path):
        r = requests.get(url)
        with open(path, 'wb') as f:
            f.write(r.content)
            f.close()
            print("文件保存成功")
    else:
        print("文件已存在")
except:
    print("爬取失败")
```

实例5：IP地址归属地的自动查询



<http://m.ip138.com/ip.asp?ip=ipaddress>

 iP138

导航 

www.ip138.com IP查询

IP地址或者域名：

查询

手机号码所在地区强力查询

输入手机号即可知道用户所在的地区

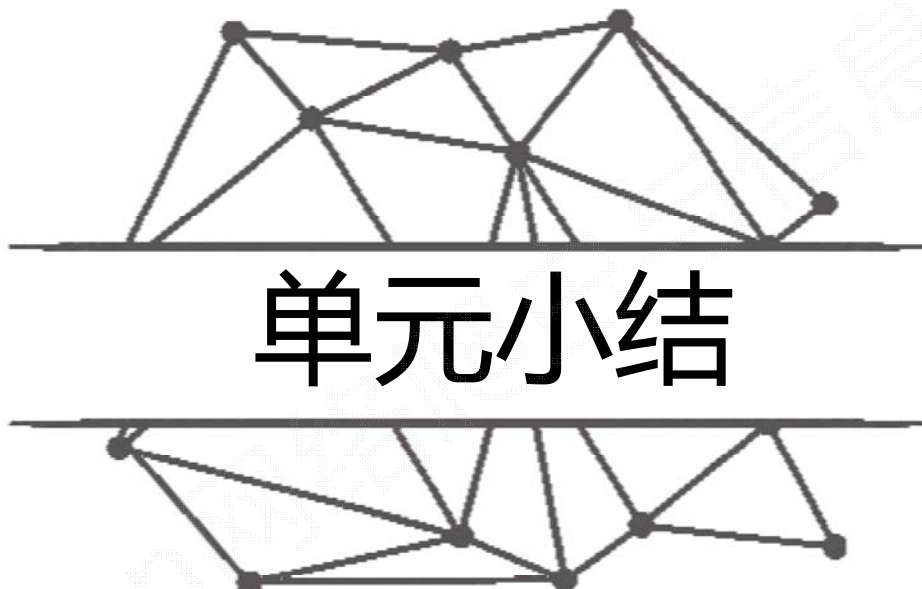
手机号码（段）：

查询

```
>>> import requests
>>> url = 'http://m.ip138.com/ip.asp?ip='
>>> r = requests.get(url+'202.204.80.112')
>>> r.status_code
200
>>> r.text[-500:]
'value="查询" class="form-btn" />\r\n\t\t\t\t\t</form>\r\n\t\t\t\t\t</div>\r\n\t\t\t\t\t<div class="query-hd">ip138.com IP查询(搜索IP地址的地理位置)</div>\r\n\t\t\t\t\t<h1 class="query">您查询的IP: 202.204.80.112</h1><p class="result">本站主数据: 北京市海淀区北京理工大学 教育网</p><p class="result">参考数据一: 北京市 北京理工大学</p>\r\n\r\n\t\t\t\t\t</div>\r\n\t\t\t\t\t</div>\r\n\r\n\t\t\t\t\t<div c
```

IP地址查询全代码

```
import requests
url = "http://m.ip138.com/ip.asp?ip="
try:
    r = requests.get(url+'202.204.80.112')
    r.raise_for_status()
    r.encoding = r.apparent_encoding
    print(r.text[-500:])
except:
    print("爬取失败")
```



单元小结

Requests库网络爬取实战

实例1：京东商品页面的爬取

实例2：亚马逊商品页面的爬取

实例3：百度/360搜索关键字提交

实例4：网络图片的爬取和存储

实例5：IP地址归属地的自动查询

以爬虫视角看
待网络内容