

实例2：淘宝商品比价定向爬虫

WS08



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



Beautiful Soup

解析HTML页面



Re

正则表达式详解
提取页面关键信息



Projects

实战项目A

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



“淘宝商品比价定向爬虫”实例介绍

淘宝网
Taobao.com

更多市场

宝贝书包

搜索

上传图片就能搜同款啦!

掌柜热卖

所有宝贝 天猫 二手 我的搜索 今日发现

所有分类

品牌:

FENRUIEN/梵瑞恩 沃曼威斯 阿迪达斯 MCYS&JPN/沐村耀司 耐克 瑞士军刀 SUNMAN CYZM. 1983 多选 更多

Disney 阳光树 更新 Circle of friends/朋友圈 李宁 艾私奔 Jansport 格格宝贝 雨凡 米森

材质:

帆布 牛津纺 涤纶 PU 锦纶 牛皮 牛仔布 PVC PC 丝绒 呢子 麻 多选 更多

包内部结构:

拉链暗袋 手机袋 证件袋 电脑插袋 夹层拉链袋 网袋 相机插袋 暗格 钥匙位 零钱位 多选 更多

功能箱包:

双肩背包 帆布背包 运动背包 单肩背包 箱包配件 零钱包 休闲钱包 旅行包 更多

筛选条件:

女包 玩具/童车/益智/积木/模型 大小 适用性别 相关分类

您是不是想找:

书包中学生女 书包男 书包小学生 书包中学生男 学生书包 幼儿园书包 儿童书包 书包小学生女 开学书包 sb书包 ulzzang书包 原宿书包

综合排序

人气 销量 信用 价格

¥ - ¥

发货地

1/100

☐ 女王节 ☐ 包邮 ☐ 赠送退货运费险 ☐ 货到付款 ☐ 新品 ☐ 海外商品 ☐ 二手 ☐ 天猫 ☐ 正品保障 更多

合并同款宝贝

您好 myboss56, “书包”相关的宝贝: 购买过店铺 找到 14 件 | 黄钻爱买店铺 找到 500+ 件 | 回头客爱买店铺 找到 500+ 件

deli 得力办公

WENGER 瑞士军刀威戈

SEPT WOLVES 官方旗舰店 品牌直营

包邮爆款

送五件

¥65.00

功能描述

目标：获取淘宝搜索页面的信息，提取其中的商品名称和价格

理解：

淘宝的搜索接口

翻页的处理



技术路线：`requests-bs4-re`

“书包”

https://s.taobao.com/search?q=书包&js=1&stats_click=search_radio_all%3A1&initiative_id=staobaoz_20170105&ie=utf8 起始页

https://s.taobao.com/search?q=书包&js=1&stats_click=search_radio_all%3A1&initiative_id=staobaoz_20170105&ie=utf8&bcoffset=0&ntoffset=0&p4ppushleft=1%2C48&s=44 第2页

https://s.taobao.com/search?q=书包&js=1&stats_click=search_radio_all%3A1&initiative_id=staobaoz_20170105&ie=utf8&bcoffset=-3&ntoffset=-3&p4ppushleft=1%2C48&s=88 每页44个商品 第3页

搜索接口和翻页的URL对应属性

定向爬虫可行性

<https://s.taobao.com/robots.txt>

User-agent: *

Disallow: /

请注意：这个例子仅探讨技术实现，请不要不加限制的爬取该网站

程序的结构设计

步骤1：提交商品搜索请求，循环获取页面

步骤2：对于每个页面，提取商品名称和价格信息

步骤3：将信息输出到屏幕上

“淘宝商品比价定向爬虫”实例编写

main()

```
import requests
import re

def getHTMLText(url):
    print("")

def parsePage(ilt, html):
    print("")

def printGoodsList(ilt):
    print("")

def main():
    goods = '书包'
    depth = 2
    start_url = 'https://s.taobao.com/search?q=' + goods
    infoList = []
    for i in range(depth):
        try:
            url = start_url + '&s=' + str(44*i)
            html = getHTMLText(url)
            parsePage(infoList, html)
        except:
            continue
    printGoodsList(infoList)

main()
```

getHTMLText()

```
def getHTMLText(url):  
    try:  
        r = requests.get(url, timeout=30)  
        r.raise_for_status()  
        r.encoding = r.apparent_encoding  
        return r.text  
    except:  
        return ""
```

parsePage()

```
def parsePage(ilt, html):  
    try:  
        plt = re.findall(r'"view_price"\':"\d\.'*"',html)  
        tlt = re.findall(r'"raw_title"\':"\. *?\"',html)  
        for i in range(len(plt)):  
            price = eval(plt[i].split(':')[1])  
            title = eval(tlt[i].split(':')[1])  
            ilt.append([price , title])  
    except:  
        print("")
```

printGoodsList()

```
def printGoodsList(ilt):  
    tplt = "{:4}\t{:8}\t{:16}"  
    print(tplt.format("序号", "价格", "商品名称"))  
    count = 0  
    for g in ilt:  
        count = count + 1  
        print(tplt.format(count, g[0], g[1]))
```

全代码

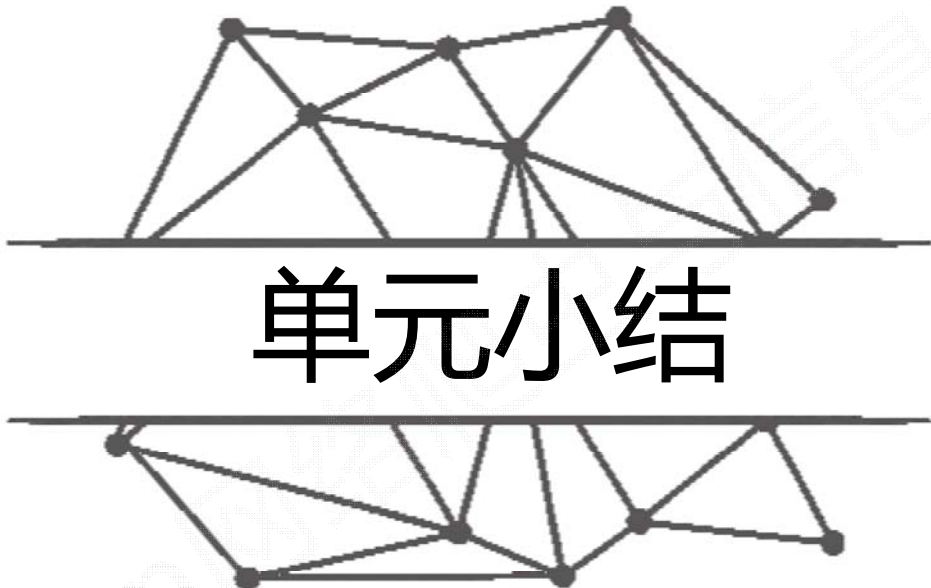
请阅读全代码

Python网络爬虫与信息提取


输出结果

>>>

序号	价格	商品名称
1	142.00	得力小学生男女减负护脊双肩背包儿童书包
2	69.00	迪士尼书包小学生男女1-3-4-6年级米奇减负背包儿童书包8-10-12岁
3	119.00	韩国kk树书包小学生男1-3-4-6年级儿童开学女6-12周岁双肩包护脊
4	168.00	新款韩版男士双肩包时尚潮流电脑背包男休闲高初中学生开学书包男
5	368.00	尼龙防水双肩包男户外登山旅游包背包学院风高中生初中学生书包女
6	96.00	欧扉双肩包女韩版pu书包时尚百搭学院风女士包包2016新款潮背包女
7	156.00	男士青年背包时尚潮流双肩包男韩版学生书包电脑包休闲运动旅游包
8	289.00	米森双肩包男士背包大高中学生书包商务休闲电脑旅行包韩版潮正品
9	168.00	潮道双肩包男士大学生书包男时尚潮流旅行背包韩版休闲包包电脑包
10	298.00	瑞戈瑞士军刀双肩包男背包男士旅行包女中学生书包商务休闲韩版潮
11	199.00	沃曼威斯韩版大高中学生书包男时尚潮流背包学院风2017旅行双肩包
12	139.00	男士双肩包 书包男女中学生大学生 旅行韩版时尚潮流休闲电脑背包
13	599.00	新款牛津布韩版尼龙双肩包女包简约百搭休闲旅行小背包女帆布书包
14	272.00	英伦帆布双肩包女韩版初中学生学院风书包高中生校园简约百搭背包
15	199.00	双肩包女2016新款时尚铆钉韩版牛津布旅行背包夏季百搭尼龙书包
16	142.00	木村耀司韩版双肩包书包中学生女双肩书包男休闲帆布背包男女包潮
17	259.00	新潮代韩版双肩包休闲旅行包男背包帆布包书包大容量运动旅游男包
18	199.00	新款韩版高初中学生书包休闲背包街头潮流双肩包男个性旅行包
19	399.00	铆钉双肩包女包日韩版pu学院风书包2016新款潮大容量休闲旅行背包
20	168.00	时尚潮流功能双肩包男高中大学生开学书包男背包包男电脑包旅行包
21	96.00	欧扉双肩包女韩版书包时尚百搭女士包包2016新款潮流尼龙小背包女
22	89.00	韩国原宿风ulzzang双肩包女帆布中学生书包可爱萌日系背包小清新
23	59.80	帆布双肩包女学院风书包韩版潮流简约字母休闲大容量男女学生背包



单元小结



A diagram of a truss structure with nodes and members, positioned below a horizontal line.

实例2：淘宝商品比价定向爬虫

采用requests-re路线实现了淘宝商品比价定向爬虫

熟练掌握正则表达式在信息提取方面的应用