# python爬虫项目总结

小组成员：杨立伦、姜裕、王文慧、王龙均
黄思雨、王盼、吴凌霄、徐睿

汇报人　：杨立伦

# 目录

Content


1 项目需求

2 环境准备

3 详细设计

4 实现过程

5 总结

01

# 第一章　项目需求

# 项目需求

## 基础目标

掌握Python程序运行方法，会修改基本语法和逻辑错误。掌握Python语言中的if..elif、循环结构while、for、函数递归调用以及爬虫相关知识。

## 进阶需求

使用第三方库对网站数据进行爬取，汇总。并且不局限于一种库，锻炼多种库的混合使用

## 能力提升

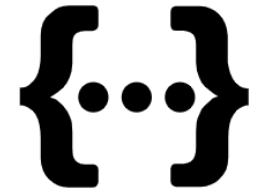
在爬取了数据之后，可以根据实际需求，存入MongoDB，或者存入csv文件中，以便于以后使用

02

# 第二章　环境准备

# 环境准备

| 软件名称 | 描述 |
| --- | --- |
| python3.8 编程语言 | 开发游戏主语言 |
| requests请求库 | 发送http请求保存，读取游戏记录 |
| selenium | 自动化控制浏览器的库 |
| pyquery | 能在python中使用选择器 |
| threading | python多线程库 |
| json | python中JSON数据格式的库 |
| Queue | 队列库，用于多线程收集结果 |
| pandas | 数据分析库，有导出csv的方法 |
| pycharm | 集成开发环境 |
| Fiddler 5.0 | 抓包工具 |
| MongoDB | 非关系型数据库 |



03

# 第三章　详细设计

# 详细设计 同步开发

杨立伦 --- 淘宝。可以输入关键词和数量进行爬取

吴凌霄 --- 虎扑。爬取NBA球队比赛分数信息

王文慧 --- 蝉大师。爬取手机app排行榜

王盼 --- 美妆品库。爬取美妆品牌对应商品

王龙均 --- 疫情地图。爬取新冠疫情各个地区的新增，现有，累计，治愈，
以及死亡患者的数据

姜裕 --- 中关村ZOL。爬取热门手机商品信息，以及价格

黄思雨 --- 114票务网。爬取火车站时刻表，每一个火车站一个线程，
可以设置要爬取的车站

徐睿 --- 天极产品库。爬取电子产品信息

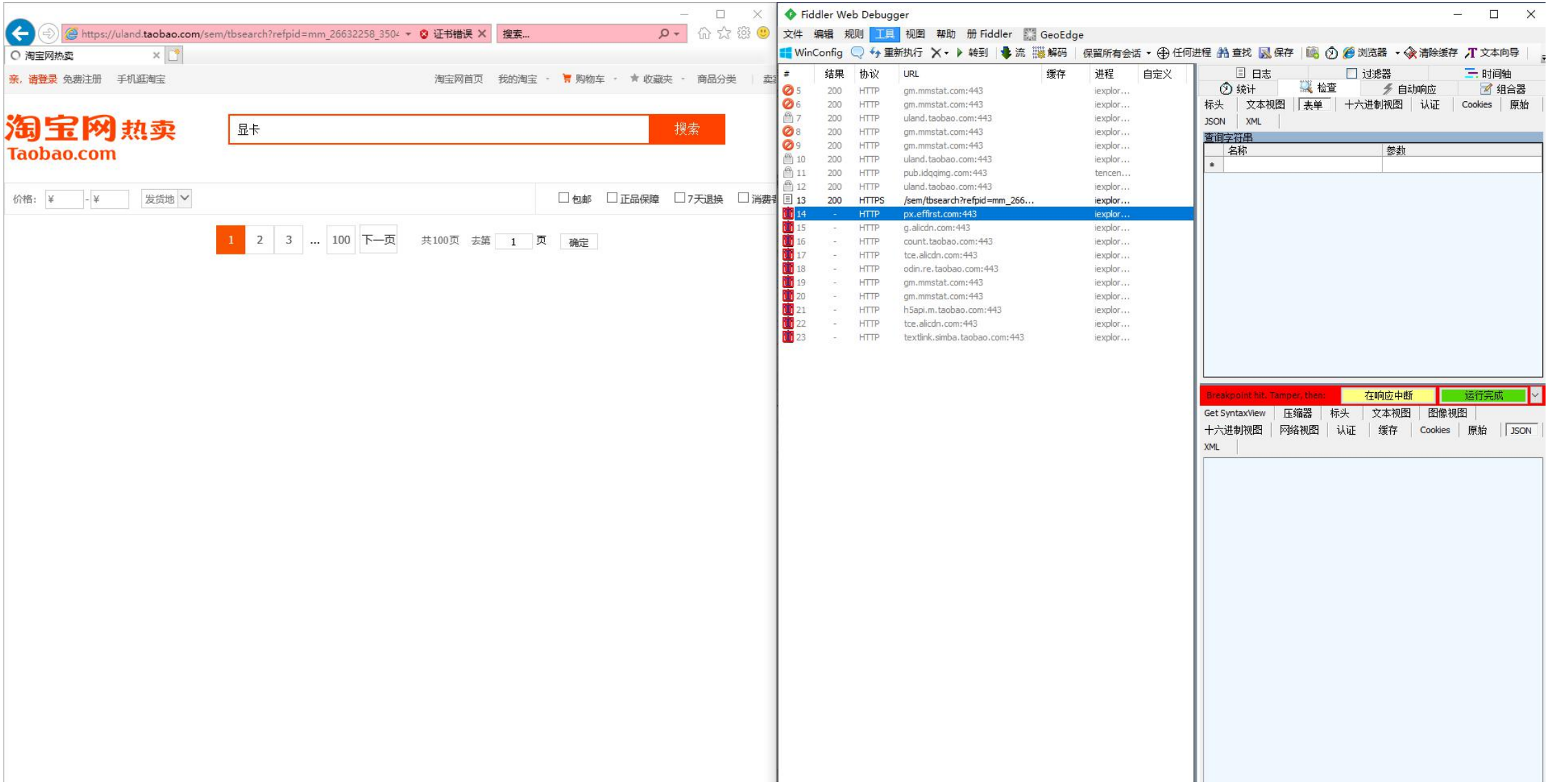
04

# 第四章　实现过程

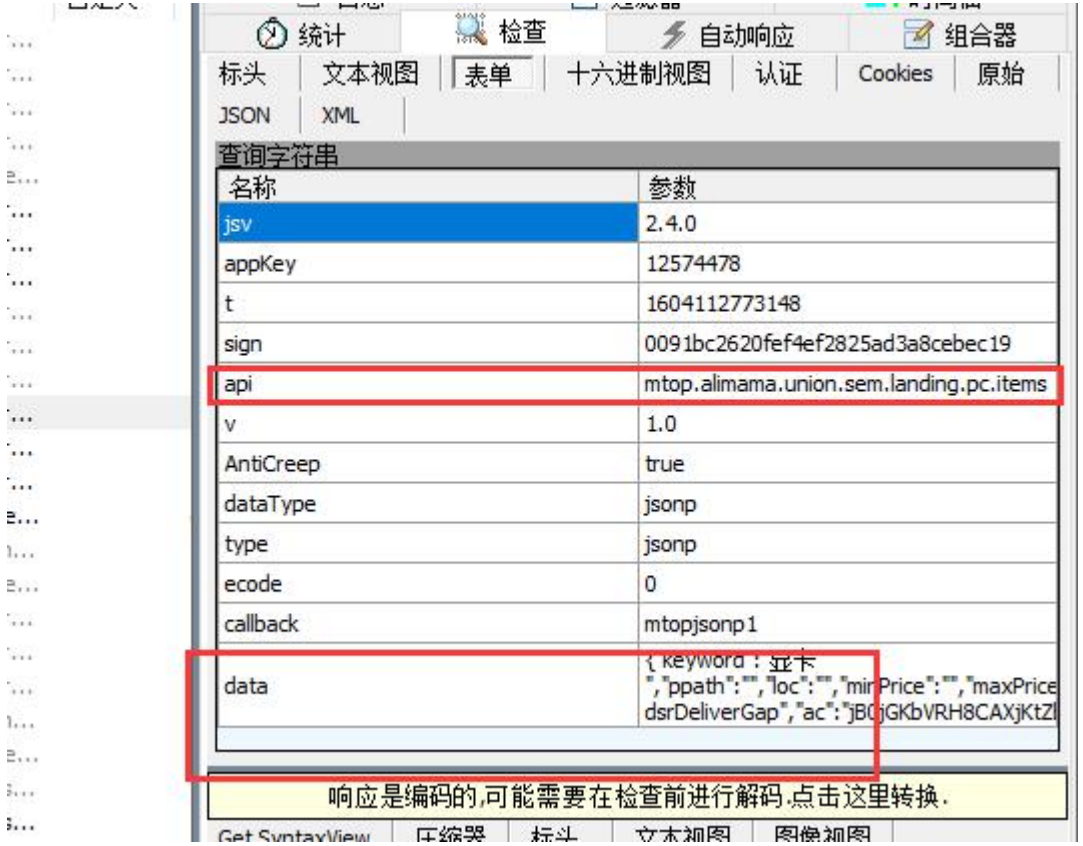
# 实现过程

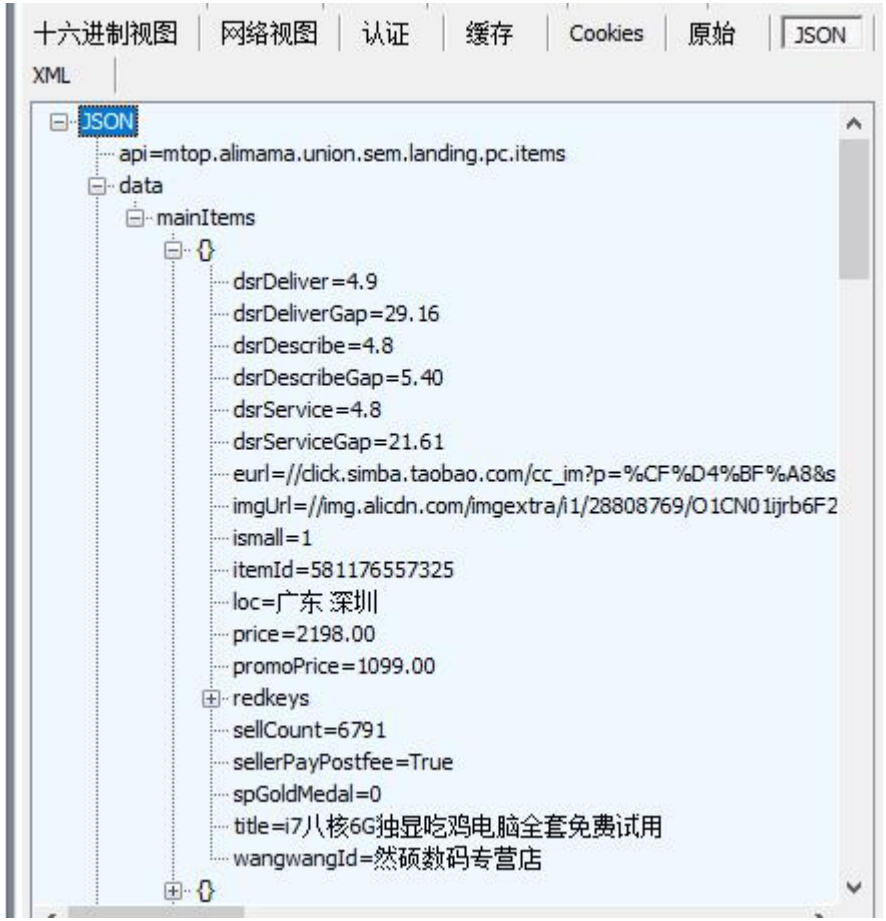


抓包判断请求那个接口可以获取到数据



抓包判断请求那个接口可以获取到数据



# 实现过程



## 主函数

```python
def main():
    # 输入要查询的产品
    n = input('输入要查询的产品种类数量')
    num = input('输入每种爬取数量')

    words = []
    # 循环录入商品名，每个商品一个线程
    for i in range(int(n)):
        words.append(input(f'请输入第{i + 1}种商品名'))

    # 队列结果
    que = Queue()

    threads = []
    # 每一个词一个线程
    for word in words:
        t = threading.Thread(target=getDataByName,args=(word,num,que))
        t.start()
        threads.append(t)
    # 等待全部结束
    for t in threads:
        t.join()
    # 查看结果
    result = []
    for t in threads:
        result += que.get()
    # 保存到csv中
    pd = pandas.DataFrame(columns=['title','wangwangId','loc','price'],data=result)
    pd.to_csv('list.csv',encoding='utf-8')
```

```python
根据商品名获取数据
ef getDataByName(name,num,que):

    # lock.acquire()
    # 首先拼接请求字符串
    url = 'https://odin.re.taobao.com/search_tbuad?keyword='+name+'&count='+num+'&pid=1'

    doc = requests.post(url)
    data = doc.text.encode().decode('unicode_escape')
    # 请求的数据转成json
    list = json.loads(data)

    # 获取有效部分
    sps = list.get('data').get('data1')
    arr = []
    for sp in sps:
        title = sp.get('title')
        wangwangId = sp.get('wangwangId')
        loc = sp.get('loc')
        price = sp.get('price')
        arr.append({
            'title':title,
            'wangwangId':wangwangId,
            'loc':loc,
            'price':price
        })
    # 结果发送到队列中
    que.put(arr)
```

# 实现过程

杨立伦-淘宝

根据商品名

获取数据

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1160 | 1158 | VGA显卡竖放转换架 垂直支架 带延长线 PCI- | gamemax旗舰店 | 广东 东莞 | 229 |
| 1161 | 1159 | AMD锐龙R5 2600组装电脑 | 东邦神 | 广东 广州 | 1599 |
| 1162 | 1160 | 电脑台式机酷睿九代I5家用迷你整机办公主机 | thinkcentre旗舰店 | 浙江 温州 | 3199 |
| 1163 | 1161 | 九代i7八核GTX1660游戏型吃鸡主机 | 晶锐科技01 | 四川 成都 | 765.56 |
| 1164 | 1162 | 2020扬天V340酷睿十代i7联想笔记本电脑i7笔 | 联想龙企华商专卖店 | 北京 | 6499 |
| 1165 | 1163 | HP惠普战66台式机电脑WIN7 i3无线小主机箱 | 惠普悟吉塔专卖店 | 上海 | 2699 |
| 1166 | 1164 | Lenovo/联想 MIIX 700 - | 力圣通讯 | 广东 广州 | 1850 |
| 1167 | 1165 | 迷你主机i5i7办公家用客厅小电脑微... | 宝路数码 | 广东 深圳 | 1122.5 |
| 1168 | 1166 | 吃鸡显卡GTX950 2G GTX960 | 杨得培 | 广东 深圳 | 369 |
| 1169 | 1167 | 耕升GTX1650/1650S super显卡4G 电竞游戏独 | 峰贸数码专营店 | 江苏 徐州 | 1298 |
| 1170 | 1168 | 顺丰ROG华硕RTX3070 GAMING O8G猛禽显卡 | 嘉志硕数码专营店 | 重庆 | 5899 |
| 1171 | 1169 | 电脑主机i5 10400F/GTX1050ti/1650 super/1 | 万达凯旋数码专营店 | 广东 深圳 | 2999 |
| 1172 | 1170 | 吧台餐桌一体家用厨房吧台餐桌一体阳台吧台 | kangguiwen | 江苏 徐州 | 74.5 |
| 1173 | 1171 | 技嘉RTX2060 SUPER GAMING OC 8G独立显卡 G | 深圳七七数码专营店 | 广东 深圳 | 3799 |
| 1174 | 1172 | 台式电脑机箱挂梁支架子主机壁挂吊悬挂托盘 | on_light | 广东 中山 | 146 |
| 1175 | 1173 | 联想台式机电脑e75 e74 e95 e96电脑整机 | 驰雍数码专营店 | 上海 | 1750 |
| 1176 | 1174 | 华硕GTX650 1G 2G 台式独显游戏显卡LOL CF | yanning1980 | 广东 深圳 | 288 |
| 1177 | 1175 | 七彩虹GTX1660 SUPER 6G 战斧 游戏显卡 | 千机堂电脑专营店 | 湖北 武汉 | 2199 |
| 1178 | 1176 | 超薄一体机电脑十代酷睿I7高配独显全套 | tb23782393 | 广东 深圳 | 1098 |
| 1179 | 1177 | 戴尔台式电脑全套迷你主机游戏办公主机 | 戴尔岩阳专卖店 | 上海 | 3399 |
| 1180 | 1178 | 七彩虹GTX1650 4G D6 战斧 吃鸡游戏显卡 | 千机堂电脑专营店 | 湖北 武汉 | 1299 |
| 1181 | 1179 | PHANTEKS追风者PR22 支持RTX2080Ti/2070抗 | phanteks旗舰店 | 广东 深圳 | 199 |
| 1182 | 1180 | 现货全新技嘉3090 TURBO-24GAI智能运算图形 | 淘淘淘贝贝宝宝 | 山东 济南 | 14599 |
| 1183 | 1181 | 台式电脑主机高配组装电脑网吧游戏台式电脑 | 心照不宣8023 | 山东 烟台 | 990 |
| 1184 | 1182 | 非公版RTX2070S显卡2080S Ti定制全覆盖水冷 | kingvalen | 湖北 武汉 | 349 |
| 1185 | 1183 | 技嘉(GIGABYTE) GeForce 1650 WINDFORCE OC | hshushi1982 | 贵州 贵阳 | 1139 |
| 1186 | 1184 | I3 9100F/1060 6G大显存吃鸡LOL电竞网吧家 | 峰贸数码专营店 | 北京 | 1959 |
| 1187 | 1185 | gt 620游戏显卡dp高清dvi a卡 | 百年数码_2005 | 浙江 杭州 | 356 |
| 1188 | 1186 | 联力华硕ROG包豪斯O11联名机箱 台式 | lianli旗舰店 | 湖北 武汉 | 2099 |
| 1189 | 1187 | 铭瑄GTX1650S Super 终结者 | 铭瑄电脑硬件旗舰店 | 湖北 武汉 | 899 |
| 1190 | 1188 | 旌宇多屏显卡一拖四显卡4HDMI显卡GT | 第n街坊 | 北京 | 499 |
| 1191 | 1189 | 游戏高配置组装电脑主机吃鸡全套台式整机 | 御龙堂数码专营店 | 广东 深圳 | 1210 |
| 1192 | 1190 | 武汉本地电脑回收 台式机 笔记本 显示器 | v网科技 | 湖北 武汉 | 888 |
| 1193 | 1191 | 全新Quadro K2000 2GB专业视频编辑设计显卡 | 天上会飞的小鱼 | 广东 广州 | 370 |
| 1194 | 1192 | i5/i7八核吃鸡电脑主机GTA5游戏型 | a240401241 | 广东 深圳 | 2199 |
| 1195 | 1193 | 酷睿I5i7电脑主机台式英雄联盟游戏手游 | h5558607 | 广东 东莞 | 3396 |
| 1196 | 1194 | BC黑猫馆 高端纯白女神雅典娜性感蕾丝透视 | 道明1958 | 上海 | 68 |
| 1197 | 1195 | 联想i5台式机 电脑主机全套 M6603 | 联想索格专卖店 | 北京 | 3668 |
| 1198 | 1196 | 游戏电脑主机组装电脑吃鸡gta5网吧台式机 | 共好数码专营店 | 浙江 宁波 | 4999 |
| 1199 | 1197 | 多屏显卡4屏显卡一拖四分屏显卡炒股期货4 | 第n街坊 | 北京 | 269 |
| 1200 | 1198 | Intel酷睿I5四核台式电脑 全套DI | 字阳电子科技 | 浙江 杭州 | 599 |
| 1201 | 1199 | i5 10500六核独显迷你吃鸡游戏电脑 | 新希望数码商城 | 湖南 长沙 | 4699 |
| 1202 | | | | | |

# 实现过程



总设计：爬取火车站时刻表，每一个火车站一个线程，可以设置要爬取的车站

1.设置车站代号列表

2.根据代号创建响应的线程

3.爬取后追加到总结果中

4.输出到文件

```python
def job(station):
    global result
    resp = requests.get(f'http://shike.114piaowu.com/{station}/')
    resp.encoding = 'utf-8'
    doc = pq(resp.text)
    sk_lists = doc('div.sk_list')
    for list in sk_lists:
        list = pq(list)
        # 列车号
        code = list('li.sk01 a').text()
        # 出发时间
        start = pq(list('li.sk02 p')[0]).text()
        # 到达时间
        end = pq(list('li.sk02 p')[1]).text()
        temp = list('li.sk03 p').text().split(' ')
        # 始发站
        start_station = temp[0]
        # 终点站
        end_station = temp[1]
        type = list('li.sk04').text()
        result.append({
            'code':code,
            'start':start,
            'end':end,
            'start_station':start_station,
            'end_station':end_station,
            'type':type
        })
```

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1220 | 1218 | G81 | 13:24 | 17:47 | 武汉 | 贵阳北 | 高速 |
| 1221 | 1219 | G2044 | 11:41 | 14:00 | 武汉 | 郑州东 | 高速 |
| 1222 | 1220 | D5997 | 8:48 | 13:25 | 汉口 | 利川 | 动车 |
| 1223 | 1221 | G71 | 13:08 | 18:03 | 武汉 | 福田 | 高速 |
| 1224 | 1222 | D3003 | 12:08 | 14:26 | 汉口 | 襄阳东 | 动车 |
| 1225 | 1223 | G505 | 20:59 | 22:32 | 武汉 | 长沙南 | 高速 |
| 1226 | 1224 | G1316 | 11:16 | 13:32 | 武汉 | 宜昌东 | 高速 |
| 1227 | 1225 | D296 | 18:11 | 21:14 | 汉口 | 郑州 | 动车 |
| 1228 | 1226 | G553 | 16:30 | 20:45 | 武汉 | 广州南 | 高速 |
| 1229 | 1227 | D5951 | 19:42 | 20:35 | 武汉 | 大冶北 | 动车 |
| 1230 | 1228 | G533 | 17:23 | 18:41 | 武汉 | 长沙南 | 高速 |
| 1231 | 1229 | G491 | 18:10 | 20:27 | 武汉 | 南昌西 | 高速 |
| 1232 | 1230 | G510 | 8:10 | 13:39 | 武汉 | 北京西 | 高速 |
| 1233 | 1231 | D3004 | 17:25 | 22:41 | 汉口 | 上海虹桥 | 动车 |
| 1234 | 1232 | D627 | 7:31 | 14:02 | 汉口 | 重庆北 | 动车 |
| 1235 | 1233 | G2043 | 18:16 | 20:49 | 武汉 | 南昌西 | 高速 |
| 1236 | 1234 | D5242 | 9:10 | 11:38 | 武昌 | 襄阳东 | 动车 |
| 1237 | 1235 | G424 | 16:39 | 20:51 | 武汉 | 石家庄 | 高速 |
| 1238 | 1236 | Z98 | 5:00 | 15:30 | 武昌 | 北京西 | 直特 |
| 1239 | 1237 | G1135 | 18:08 | 22:23 | 武汉 | 广州南 | 高速 |
| 1240 | 1238 | D296 | 17:27 | 21:14 | 武汉 | 郑州 | 动车 |
| 1241 | 1239 | D5245 | 15:10 | 17:40 | 汉口 | 襄阳 | 动车 |
| 1242 | 1240 | D5998 | 18:53 | 19:17 | 汉口 | 武汉 | 动车 |
| 1243 | 1241 | G520 | 7:00 | 11:26 | 汉口 | 北京西 | 高速 |
| 1244 | 1242 | G1289 | 20:35 | 22:08 | 武汉 | 长沙南 | 高速 |
| 1245 | 1243 | G1315 | 17:28 | 21:43 | 武汉 | 广州南 | 高速 |
| 1246 | 1244 | G588 | 8:30 | 12:42 | 武汉 | 北京西 | 高速 |
| 1247 | 1245 | G1318 | 16:51 | 21:43 | 汉口 | 广州南 | 高速 |
| 1248 | 1246 | G1317 | 11:37 | 13:32 | 汉口 | 宜昌东 | 高速 |
| 1249 | 1247 | G844 | 16:51 | 23:51 | 武汉 | 兰州西 | 高速 |
| 1250 | 1248 | G79 | 14:20 | 18:46 | 武汉 | 福田 | 高速 |
| 1251 | 1249 | D5954 | 19:15 | 20:35 | 汉口 | 大冶北 | 动车 |
| 1252 | 1250 | D7081 | 6:03 | 7:52 | 汉口 | 宜昌东 | 动车 |
| 1253 | 1251 | G1143 | 9:16 | 13:55 | 武汉 | 广州南 | 高速 |
| 1254 | 1252 | G528 | 17:59 | 23:20 | 武汉 | 北京西 | 高速 |
| 1255 | 1253 | K799 | 11:52 | 7:29 | 武昌 | 汕头 | 快速 |
| 1256 | 1254 | D5996 | 8:22 | 13:25 | 武汉 | 利川 | 动车 |
| 1257 | 1255 | G558 | 18:14 | 23:34 | 武汉 | 北京西 | 高速 |
| 1258 | 1256 | D5989 | 7:14 | 11:16 | 汉口 | 恩施 | 动车 |
| 1259 | 1257 | G426 | 20:35 | 23:12 | 武汉 | 郑州东 | 高速 |
| 1260 | 1258 | G1523 | 16:07 | 21:19 | 武汉 | 贵阳北 | 高速 |
| 1261 | 1259 | G542 | 14:05 | 16:21 | 武汉 | 郑州东 | 高速 |

# 实现过程



总设计：爬取新冠疫情地图数据，获取各个地区的新增，现有，累计，治愈，以及死亡患者的数据

1.获取页面数据
2.对数据进行关键提取
3.封装成自己需要的格式
4.打包成csv文件

```python
def main():
    driver = webdriver.Chrome("K:\上课\python\爬虫\chromedriver86.0.4240.22.exe")  # chromedriver所在路径
    driver.get('https://voice.baidu.com/act/newpneumonia/newpneumonia')
    html = pq(driver.find_element_by_css_selector('html').get_attribute('outerHTML'))
    trs = html('#nationTable tbody tr')
    arr = []
    for tr in trs:
        tr = pq(tr)
        tds = pq(tr('td'))
        # 地区
        area = pq(tds[0]).text()
        # 新增
        news = pq(tds[1]).text()
        # 现有
        nows = pq(tds[2]).text()
        # 累计
        sum = pq(tds[3]).text()
        # 治愈
        save = pq(tds[4]).text()
        # 死亡
        died = pq(tds[5]).text()
        arr.append({
            'area': area,
            'news': news,
            'nows': nows,
            'sum': sum,
            'save': save,
            'died': died
        })
    saveCSV('yiqing.csv', arr)
```

| | area | news | nows | sum | save | died |
|---|---|---|---|---|---|---|
| 0 | 香港 | 7 | 133 | 5320 | 5082 | 105 |
| 1 | 上海 | 8 | 95 | 1176 | 1074 | 7 |
| 2 | 新疆 | 6 | 51 | 953 | 899 | 3 |
| 3 | 台湾 | 1 | 33 | 554 | 514 | 7 |
| 4 | 四川 | 4 | 31 | 743 | 709 | 3 |
| 5 | 广东 | 3 | 29 | 1919 | 1882 | 8 |
| 6 | 陕西 | 0 | 28 | 454 | 423 | 3 |
| 7 | 福建 | 0 | 25 | 436 | 410 | 1 |
| 8 | 内蒙古 | 1 | 19 | 288 | 268 | 1 |
| 9 | 天津 | 5 | 16 | 270 | 251 | 3 |
| 10 | 河北 | 1 | 8 | 373 | 359 | 6 |
| 11 | 浙江 | 3 | 6 | 1286 | 1279 | 1 |
| 12 | 山东 | 0 | 6 | 847 | 834 | 7 |
| 13 | 江苏 | 0 | 6 | 672 | 666 | 0 |
| 14 | 辽宁 | 0 | 5 | 283 | 276 | 2 |
| 15 | 山西 | 0 | 5 | 212 | 207 | 0 |
| 16 | 河南 | 0 | 4 | 1284 | 1258 | 22 |
| 17 | 北京 | 0 | 4 | 942 | 929 | 9 |
| 18 | 重庆 | 0 | 4 | 589 | 579 | 6 |
| 19 | 云南 | 1 | 4 | 213 | 207 | 2 |
| 20 | 湖南 | 1 | 1 | 1020 | 1015 | 4 |
| 21 | 黑龙江 | 0 | 1 | 949 | 935 | 13 |

# 实现过程

姜裕-中关村ZOL

总设计：爬取热门手机商品信息，以及价格
1.中关村报价网
2.设置参数，要爬取的页数（每页48台手机）
3.每一页一个线程爬取，往总结果中追加

```python
    global result
    resp = requests.get(f'http://detail.zol.com.cn/cell_phone_index/subcate57_0_list_1_0_1_2_0_{page}.h
    doc = pq(resp.text)
    lis = doc('#J_PicMode li')
    arr = []
    for li in lis:
        li = pq(li)
        name = li('h3').text()
        if name == '' or name == None:
            continue

        price = li('b.price-type').text()
        score = li('span.score').text()
        arr.append({
            'name': name,
            'price': price,
            'score': score
        })
    result += arr

def main():
    threads = []

    # 三个页面，三个线程
    for i in range(3):
        t = threading.Thread(target=job,args=(i+1,))
        t.start()
        threads.append(t)

    # 等待全部结束
    for t in threads:
        t.join()
```

H26

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 103 | 101 | 小米10青春版（8GB/128GB/全网通/5G版） 50倍潜望式变焦，AI魔法分身，双模5G | 2199 | 8.1 | |
| 104 | 102 | 荣耀10（全网通） 变色极光玻璃，前置隐形湿手指纹解锁，2400万AI摄影，全面屏 | 1488 | 8 | |
| 105 | 103 | 三星Galaxy Z Fold2（12GB/512GB/全网通/5G版） 自适应分屏模式，双芯智能电池，双重预览 | 16999 | 9.1 | |
| 106 | 104 | vivo Y5s（6GB/128GB/全网通） AI智慧三摄，18W双引擎闪充，游戏魔盒 | 1498 | 8.5 | |
| 107 | 105 | OPPO A52（8GB/128GB/全网通） 18W疾速快充，星阵AI四摄，5000mAh电池 | 1399 | 9 | |
| 108 | 106 | realme Q2（6GB/128GB/全网通/5G版） 双模5G，天玑800U，120Hz畅速屏，30W闪充 | 1299 | | |
| 109 | 107 | 荣耀X10 Max（6GB/128GB/全网通/5G版） RGBW阳光屏，5000mAh大电池，双对称扬声器，7.09英寸大屏 | 2089 | 7.2 | |
| 110 | 108 | OPPO Reno Ace（8GB/128GB/全网通） Supervooc超级闪充2.0，90HZ电竞屏，骁龙855+ | 3299 | 9.2 | |
| 111 | 109 | vivo Y70s（8GB/128GB/全网通/5G版） 三星 Exynos 880，4500毫安电池，侧边指纹，双模5G | 2098 | 8.6 | |
| 112 | 110 | 三星Galaxy Note 20（8GB/256GB/全网通/5G版） 骁龙865+，S Pen，专业视频拍摄 | 7399 | | |
| 113 | 111 | 华为nova 7 Pro（8GB/256GB/5G版/全网通） 前置3200万追焦双摄 ，50倍潜望式变焦四摄，麒麟985 | 4099 | 8.8 | |
| 114 | 112 | 荣耀30青春版（6GB/64GB/全网通/5G版） 90Hz刷新率，4000mAh大电池，AI极速抓拍 | 1699 | 8.7 | |
| 115 | 113 | 小米10（12GB/256GB/全网通/5G版/至尊纪念版） 120倍变焦，120W闪充，120Hz刷新率，240Hz采样率 | 5999 | 9.2 | |
| 116 | 114 | iQOO 3（6GB/128GB/全网通/5G版） 零感网络切换，Multi-Turbo 3.0，iQOO电竞模式 | 3388 | 8.7 | |
| 117 | 115 | OPPO Reno4 SE（8GB/256GB/全网通/5G版） 光芒人像三摄，OLED超清护眼屏，一体化双模 5G | 2799 | 9 | |
| 118 | 116 | 苹果iPhone 12 mini（4GB/128GB/全网通/5G版） 超瓷晶面板，A14仿生，双模5G | 5999 | 9.3 | |
| 119 | 117 | 荣耀Play4 Pro（8GB/128GB/全网通/5G） 麒麟990，40W超级快充，4000万超感光暗拍 | 2899 | 8.8 | |
| 120 | 118 | OPPO Reno3 Pro（8GB/128GB/全网通/5G版） VOOC 闪充 4.0，视频超级双防抖，7.7mm 超轻薄机身，骁龙765G | 3499 | 9.3 | |
| 121 | 119 | 华为畅享20（4GB/128GB/全网通/5G版） 6.6英寸影音大屏，5000mAh大电池，AI三摄闪拍 | 1699 | | |
| 122 | 120 | 苹果iPhone 11 Pro（4GB/256GB/全网通） HDR10，后置三摄，神经网络引擎，独立取景器 | 9999 | 8.6 | |
| 123 | 121 | 小米10T（全网通/5G版） 骁龙865，后置AI三摄，144Hz刷新率 | 即将上市 | | |
| 124 | 122 | 华为Mate30 Pro（8GB/128GB/全网通） 双4000万徕卡电影四摄，超曲面OLED环幕屏 | 5399 | 8.5 | |
| 125 | 123 | OPPO Find X2（8GB/256GB/全网通/5G版） 120Hz超感屏，3K分辨率，10亿色显示 | 4999 | 9.2 | |
| 126 | 124 | vivo Y51s（6GB/128GB/全网通） 双模5G，4500mAh电池，双引擎闪充 | 1798 | | |
| 127 | 125 | 苹果iPhone 11 Pro Max（4GB/256GB/全网通） HDR10，后置三摄，神经网络引擎，独立取景器 | 10899 | 8 | |
| 128 | 126 | 荣耀V20（6GB RAM/全网通） 液冷散热，GPU Cloud，AI双频GPS，YOYO智慧生命体，3D美体塑型，TOF立体深感摄像头 | 2399 | 9.3 | |
| 129 | 127 | 华为P40 Pro+（8GB/512GB/全网通/5G版） 徕卡五摄，双长焦镜头，麒麟990 5G | 8888 | 8.7 | |
| 130 | 128 | 华为P20 Pro（6GB RAM/全网通） 4000万徕卡三摄，AI摄影，异形全面屏，指纹识别 | 2525 | 8.3 | |
| 131 | 129 | 荣耀20（8GB/256GB/全网通） 4800万超广角AI四摄，3200W美颜自拍，NFC，麒麟980 | 2499 | 8.3 | |
| 132 | 130 | 华为P40（8GB/128GB/全网通/5G版） 徕卡三摄，麒麟990 5G，WiFi 6 | 4488 | 8.6 | |
| 133 | 131 | 小米9（6GB RAM/全网通） 4800万像素三摄，水滴全面屏，全息幻彩玻璃机身，第五代极速屏下指纹 | 2300 | 9 | |
| 134 | 132 | OPPO Reno3（8GB/128GB/全网通/5G版） 天玑1000L，6400万像素，系统级暗夜模式 | 2699 | 9.2 | |
| 135 | 133 | 苹果iPhone 12 mini（4GB/256GB/全网通/5G版） 超瓷晶面板，A14仿生，双模5G | 6799 | 9.3 | |
| 136 | 134 | 三星Galaxy S20 Ultra（12GB/256GB/全网通） 1.08亿像素后置主摄，8k视频录制，100倍变焦 | 8999 | 8.7 | |
| 137 | 135 | vivo Y3（4GB/128GB/全网通） 5000mAh大电池，后置指纹识别，后置三摄 | 1198 | 8.4 | |
| 138 | 136 | 华为nova 6 SE（8GB/128GB/全网通） 全场景AI四摄，人像超级夜景2.0，GPU Turbo，40W超级快充 | 1799 | 6.7 | |
| 139 | 137 | Redmi 10X（4GB/128GB/全网通） 后置4800万超清四摄，小孔全面屏，9W反向充电 | 999 | 5.4 | |
| 140 | 138 | 三星Galaxy Note 20 Ultra（12GB/512GB/全网通/5G版） S Pen，疾速游戏体验，三星笔记，120Hz自适应屏幕 | 9999 | 8.6 | |
| 141 | 139 | 小米10（8GB/256GB/全网通/5G版/至尊纪念版） 120倍变焦，120W闪充，120Hz刷新率，240Hz采样率 | 5599 | 9.2 | |
| 142 | 140 | 苹果iPhone SE 2（3GB/128GB/全网通） 小屏手机，A13仿生芯片，触控ID | 3799 | 7.9 | |
| 143 | 141 | Redmi 9（4GB/64GB/全网通） 联发科G80，后置AI四摄，5020mAh大电池 | 799 | 6.3 | |
| 144 | 142 | 华为Mate30 Pro（8GB/128GB/全网通/5G版/琥珀版） 双4000万徕卡电影四摄，超曲面OLED环幕屏 | 5899 | 8.5 | |
| 145 | 143 | 荣耀20青春版（4GB/64GB/全网通） 4800万后置三摄，OLED 屏幕，屏幕指纹解锁，AOD息屏显示 | 1199 | 8.3 | |

# 实现过程



总设计：爬取美妆品牌商品
面膜，化妆水，防晒隔离，唇膏
每个品牌一个线程

```python
# 获取指定页数的化妆品数据
def getDataByTotalPage(totalpage):
    total = []
    # 有多少页就循环多少次
    for i in range(totalpage):
        # 请求响应
        resp = requests.get(spanUrl(1444, i+1))
        html = pq(resp.text)
        # 根据选择器获取商品的li
        lis = html('div.dList>ul li')
        arr = []
        # 从当前商品li中获取数据
        for li in lis:
            li = pq(li)
            name = li('span.sTit').text()
            score = pq(li('span.sNub em')[0]).text()
            price = pq(li('span.sPay em')[0]).text()
            urls = pq(li('span.sTit a')[0]).attr('href')
            arr.append({
                'name': name,
                'score': score,
                'price': price,
                'urls': urls
            })
        total += arr
```

| | name | score | price | urls |
|---|---|---|---|---|
| 0 | 香奈儿 五号之水 | 9.2 | ￥1,060 | //cosme.pclady.com.cn/product/137335.html |
| 1 | 香奈儿 智慧紧肤精华液 | 9.1 | ￥1,340 | //cosme.pclady.com.cn/product/127774.html |
| 2 | 香奈儿 山茶花保湿面霜 | 9.1 | ￥720 | //cosme.pclady.com.cn/product/165703.html |
| 3 | 香奈儿 保湿隔离修饰乳 SPF30+ | 9 | ￥620 | //cosme.pclady.com.cn/product/93485.html |
| 4 | 香奈儿 青春光彩柔润粉饼SPF10（明星产品） | 9 | ￥680 | //cosme.pclady.com.cn/product/103347.html |
| 5 | 香奈儿 魅力香水润体霜 | 8.8 | ￥985 | //cosme.pclady.com.cn/product/47089.html |
| 6 | 香奈儿 双效完美粉饼 | 8.7 | ￥520 | //cosme.pclady.com.cn/product/27102.html |
| 7 | 香奈儿 炫亮魅力唇膏 | 8.7 | ￥320 | //cosme.pclady.com.cn/product/20433.html |
| 8 | 香奈儿 邂逅清新淡香水（明星产品） | 8.7 | ￥850 | //cosme.pclady.com.cn/product/27979.html |
| 9 | 香奈儿 双效眼部卸妆液 | 8.6 | ￥800 | //cosme.pclady.com.cn/product/26100.html |
| 10 | 香奈儿 四色眼影 | 8.6 | ￥560 | //cosme.pclady.com.cn/product/26856.html |
| 11 | 香奈儿 炫密睫毛膏 | 8.6 | ￥340 | //cosme.pclady.com.cn/product/102579.html |
| 12 | 香奈儿 可可小姐香水系列喷式香水 | 8.6 | ￥1,520 | //cosme.pclady.com.cn/product/23880.html |
| 13 | 香奈儿 蓝色肌底精华 | 8.5 | ￥890 | //cosme.pclady.com.cn/product/141944.html |
| 14 | 香奈儿 轻盈完美蜜粉 | 8.5 | ￥550 | //cosme.pclady.com.cn/product/25158.html |
| 15 | 香奈儿 防水眼线笔 | 8.4 | ￥235 | //cosme.pclady.com.cn/product/20642.html |
| 16 | 香奈儿 蔚蓝男士淡香水 | 8.4 | ￥860 | //cosme.pclady.com.cn/product/45104.html |
| 17 | 香奈儿 腮红霜 | 8.4 | ￥400 | //cosme.pclady.com.cn/product/101364.html |
| 18 | 香奈儿 水之吻唇蜜 | 8.4 | ￥270 | //cosme.pclady.com.cn/product/27966.html |
| 19 | 香奈儿 可可小姐唇膏（水亮系列） | 8.4 | ￥295 | //cosme.pclady.com.cn/product/72652.html |
| 20 | 香奈儿 青春光彩保湿粉饼SPF10 | 8.3 | ￥665 | //cosme.pclady.com.cn/product/57073.html |
| 21 | 香奈儿 腮红（拜占庭系列） | 8.3 | ￥450 | //cosme.pclady.com.cn/product/98265.html |
| 22 | 香奈儿 智慧紧肤按摩霜（香奈儿熨斗面膜） | 8.2 | ￥815 | //cosme.pclady.com.cn/product/126244.html |
| 23 | 香奈儿 山茶花保湿微精华露 | 8.2 | ￥810 | //cosme.pclady.com.cn/product/121455.html |
| 24 | 香奈儿 持久液体眼线笔 | 8.2 | ￥350 | //cosme.pclady.com.cn/product/21227.html |
| 25 | 香奈儿 COCO可可摩登小姐淡香精 | 8.2 | ￥990 | //cosme.pclady.com.cn/product/23016.html |
| 26 | 香奈儿 奢华精萃滋润乳霜 | 8.2 | ￥3,300 | //cosme.pclady.com.cn/product/110256.html |
| 27 | 香奈儿 可可小姐唇露 | 8.2 | ￥300 | //cosme.pclady.com.cn/product/142714.html |
| 28 | 香奈儿 智慧紧肤精华水 | 8.2 | ￥750 | //cosme.pclady.com.cn/product/126448.html |
| 29 | 香奈儿 邂逅淡香水 | 8.1 | ￥1,880 | //cosme.pclady.com.cn/product/47128.html |
| 30 | 香奈儿 山茶花保湿精华水（滋润型） | 8.1 | ￥570 | //cosme.pclady.com.cn/product/94530.html |
| 31 | 香奈儿 可可小姐女士香水（少女版） | 8.1 | ￥680 | //cosme.pclady.com.cn/product/23119.html |
| 32 | 香奈儿 智慧紧肤乳霜 | 8.1 | ￥1,305 | //cosme.pclady.com.cn/product/107490.html |
| 33 | 香奈儿 指甲油 | 8.1 | ￥215 | //cosme.pclady.com.cn/product/38134.html |
| 34 | 香奈儿 N°5香水系列低调奢华版香水 | 8 | ￥1,060 | //cosme.pclady.com.cn/product/23070.html |
| 35 | 香奈儿 炫亮魅力印记唇釉 | 8 | ￥320 | //cosme.pclady.com.cn/product/138624.html |
| 36 | 香奈儿 立体臻魅睫毛膏 | 8 | ￥320 | //cosme.pclady.com.cn/product/81500.html |
| 37 | 香奈儿 （倾城之魅）淡香水 | 8 | ￥880 | //cosme.pclady.com.cn/product/23175.html |
| 38 | 香奈儿 嘉柏丽尔香水 | 8 | ￥860 | //cosme.pclady.com.cn/product/146748.html |
| 39 | 香奈儿 美白保湿晚霜 | 8 | ￥925 | //cosme.pclady.com.cn/product/13351.html |

# 实现过程



总设计：爬取天极网产品库商品数据
每个页面一个线程，多个线程对一个队列进行添加

```python
def job(url):
    global arr
    resp = requests.get(url)
    doc = pq(resp.text)
    lists = doc('div.list')

    for list in lists:
        list = pq(list)
        name = list('h2 a').text()
        price = list('h3 a').text()
        heart = list('ul li:nth-child(2)').text()
        arr.append({
            'name':name,
            'price':price,
            'heart':heart
        })
```

| | name | price | heart |
|---|---|---|---|
| 0 | NVIDIA Quadro P620显卡 | ￥1,199 | CUDA核心：512个 |
| 1 | 七彩虹iGame GeForce RTX 2080 SUPER Vulcan OC | ￥6,099 | 核心代号：TU104 |
| 2 | 技嘉Radeon RX 5600 XT GAMING OC 6G | ￥2,399 | 核心代号：Navi 10 XLE |
| 3 | 技嘉Radeon RX 5500 XT GAMING OC 8GB | ￥1,699 | 核心代号：NAVI 14 |
| 4 | 华硕ROG-STRIX-RX 5600 XT-O6G-GAMING | ￥2,699 | 核心代号：Navi 10 XLE |
| 5 | 索泰RTX 2080 SUPER-8GD6 X-GAMING OC V2 | ￥5,399 | 核心频率：1650/1830MHz |
| 6 | 盈通RX 5600 XT 6G D6 游戏高手 OC | ￥2,299 | 核心代号：Navi 10 |
| 7 | 讯景RX 5600 XT 6GB 战狼版 | ￥2,399 | 核心代号：Navi 10 |
| 8 | 华擎RX 5600 XT Challenger D 6G OC | ￥2,399 | 核心代号：Navi 10 |
| 9 | 昂达GTX 1660 神盾6GD5 | ￥1,499 | 核心代号：TU116 |
| 10 | 盈通RX 5700 8G D6 游戏高手OC | ￥2,499 | 核心代号：NAVI XL |
| 11 | 华擎RX 5600 XT Phantom Gaming D3 6G OC | ￥2,699 | 核心代号：Navi 10 |
| 12 | 华硕DUAL-RTX 2060-6G-MINI | 即将上市 | 核心代号：TU106 |
| 13 | 华硕TUF3-RX 5600 XT-O6G-EVO-GAMING | ￥2,499 | 核心代号：Navi 10 XLE |
| 14 | 蓝宝石RX 590 GME 8G D5 超白金极光特别版 | ￥1,299 | 核心代号：Polaris |
| 15 | 七彩虹iGame GeForce RTX 2060 Vulcan X OC V3 | ￥3,188 | 核心代号：TU106-200A |
| 16 | 盈通RX 550-4G LP D5 | ￥549 | 核心代号：Polaris 11 |
| 17 | 小影霸GT730 2G D3 | ￥349 | 核心代号：GK208 |
| 18 | 蓝宝石RX 5700 XT 8G D6 超白金OC | ￥3,499 | 核心代号：NAVI XT |
| 19 | 耕升RTX 2060 Super 追风 | ￥3,099 | 核心代号：TU106-410 |
| 20 | 七彩虹网驰 GeForce GTX 1660 SUPER 电竞 6G | 暂无报价 | 核心代号：TU116 |
| 21 | 耕升GTX 1660 SUPER 炫光OC | ￥1,599 | 核心代号：TU116-300 |
| 22 | NVIDIA Quadro P2000显卡 | ￥2,790 | CUDA核心：1024个 |
| 23 | 七彩虹战斧 GeForce GTX 1650 SUPER 4G | ￥1,199 | 核心代号：TU116 |
| 24 | 小影霸R7 430 4G | ￥449 | 核心频率：780MHz |
| 25 | 蓝宝石Radeon RX 5700 XT 8G D6 | ￥3,099 | 核心代号：NAVI XT |
| 26 | 技嘉GTX 1650 MINI ITX OC 4G | ￥1,299 | 核心代号：TU117 |
| 27 | 翔升GT730 2GD5 战旗 单槽全高 | ￥359 | 核心代号：GK208 |
| 28 | 蓝宝石RX 5500 XT 8G D6 白金版 OC | ￥1,499 | 核心代号：NAVI 14 |
| 29 | NVIDIA Tesla V100 16GB | ￥66,000 | CUDA核心：5120个 |
| 30 | 蓝宝石RX 5700 XT 8G D6 白金版OC | ￥3,249 | 核心代号：NAVI XT |
| 31 | 微星半高式刀版GTX 1650 | 即将上市 | 核心代号：TU117 |
| 32 | 华擎RX 590 GME Phantom Gaming 8G OC | ￥1,399 | 核心代号：Polaris |
| 33 | 技嘉Radeon RX 5600 XT WINDFORCE OC 6G | ￥2,399 | 核心代号：Navi 10 XLE |
| 34 | 技嘉Radeon RX 5700 XT GAMING OC 8G | ￥3,099 | 核心代号：NAVI XT |
| 35 | 华硕DUAL-RTX 2070-8G-MINI | 即将上市 | 核心代号：TU106 |
| 36 | 迪兰RX 5700 XT 8G X战魔 | ￥3,349 | 核心代号：NAVI 10 |
| 37 | 讯景RX 5600 XT 6GB 海外三风扇版 | ￥2,599 | 核心代号：Navi 10 |
| 38 | NVIDIA Quadro M2000 | ￥2,599 | CUDA核心：768个 |
| 39 | 蓝宝石RX 5600 XT 6G D6 白金版 OC | ￥2,399 | 核心代号：Navi 10 XLE |
| 40 | 讯景RX 590 GME 黑狼版 | ￥1,199 | 核心代号：Polaris |
| 41 | EVGA RTX 2080Ti KINGPIN GAMING 11G | ￥20,000 | 核心代号：TU102 |
| 42 | 小影霸RX 550-4G D5 | ￥439 | 核心代号：Polaris 12 |

# 实现过程

王龙均 - 禅大师

总设计：爬取app排行榜
1.伪装headler获取源码
2.通过pyquery解析源码
3.提取有效数据部分
4.组合成新的数据格式
5.存入csv

```python
def main():
    headers = {
        'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/
    }
    doc = pq(requests.get(url='https://www.chandashi.com/ranking/index.html',headers=headers).text)
    lists = doc('#ranking-result>div')
    arr = []
    for item in lists:
        item = pq(item)
        temp = item('p.title a').text().split('.')
        paiming = temp[0]
        name = temp[1]
        info = item('div.info-box').text()
        arr.append({
            'paiming': paiming,
            'name': name,
            'info': info
        })

    saveCSV('appTOP100.csv',arr)
```

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | | paiming | name | info | | |
| 2 | 0 | 1 | 网易UU手游加速器 | 工具榜1名 | | |
| 3 | 1 | 2 | 淘宝特价版 | 购物榜1名 | | |
| 4 | 2 | 3 | 微信 | 社交榜1名 | | |
| 5 | 3 | 4 | 迅游手游加速器 | 工具榜2名 | | |
| 6 | 4 | 5 | 抖音 | 摄影与录像榜1名 | | |
| 7 | 5 | 6 | 支付宝 | 生活榜1名 | | |
| 8 | 6 | 7 | 手机淘宝 | 购物榜2名 | | |
| 9 | 7 | 8 | QQ | 社交榜2名 | | |
| 10 | 8 | 9 | 拼多多 | 购物榜3名 | | |
| 11 | 9 | 10 | biubiu加速器 | 工具榜3名 | | |
| 12 | 10 | 11 | 小红书 | 社交榜3名 | | |
| 13 | 11 | 12 | 奇游手游加速器 | 工具榜4名 | | |
| 14 | 12 | 13 | 交管12123 | 生活榜2名 | | |
| 15 | 13 | 14 | 剪映 | 摄影与录像榜2名 | | |
| 16 | 14 | 15 | 美团 | 生活榜3名 | | |
| 17 | 15 | 16 | 京东 | 购物榜4名 | | |
| 18 | 16 | 17 | 百度 | 工具榜5名 | | |
| 19 | 17 | 18 | 得物 | 体育榜1名 | | |
| 20 | 18 | 19 | 网易云音乐 | 音乐榜1名 | | |
| 21 | 19 | 20 | 花小猪打车 | 旅游榜1名 | | |
| 22 | 20 | 21 | 腾讯视频 | 娱乐榜1名 | | |
| 23 | 21 | 22 | 高德地图 | 导航榜1名 | | |
| 24 | 22 | 23 | 京东金融 | 财务榜1名 | | |
| 25 | 23 | 24 | 快手 | 摄影与录像榜3名 | | |
| 26 | 24 | 25 | 爱奇艺 | 娱乐榜2名 | | |
| 27 | 25 | 26 | 梦想养成计划 | 游戏榜1名 | | |
| 28 | 26 | 27 | 王者荣耀 | 游戏榜2名 | | |
| 29 | 27 | 28 | 钉钉 | 商务榜1名 | | |
| 30 | 28 | 29 | 淘宝直播 | 购物榜5名 | | |
| 31 | 29 | 30 | QQ音乐 | 音乐榜2名 | | |
| 32 | 30 | 31 | 酷狗音乐 | 音乐榜3名 | | |
| 33 | 31 | 32 | 快手极速版 | 摄影与录像榜4名 | | |
| 34 | 32 | 33 | 太古·妖皇诀 | 游戏榜3名 | | |
| 35 | 33 | 34 | 哔哩哔哩 | 娱乐榜3名 | | |
| 36 | 34 | 35 | 南瓜电影 | 娱乐榜4名 | | |
| 37 | 35 | 36 | 农行掌上银行 | 财务榜2名 | | |
| 38 | 36 | 37 | QQ邮箱 | 工具榜6名 | | |
| 39 | 37 | 38 | 抖音极速版 | 摄影与录像榜5名 | | |
| 40 | 38 | 39 | 云闪付 | 财务榜3名 | | |
| 41 | 39 | 40 | 滴滴出行 | 旅游榜2名 | | |
| 42 | 40 | 41 | 闲鱼 | 购物榜6名 | | |
| 43 | 41 | 42 | 企业微信 | 商务榜2名 | | |
| 44 | 42 | 43 | QQ浏览器 | 工具榜7名 | | |

appTOP100

# 实现过程



总设计：爬取NBA球队比赛分数信息

1.虎扑体育作为数据

2.模拟访问虎扑

3.获取数据

4.截取有效数据

5.整理打包成csv

```python
def main():
    driver = webdriver.Chrome("K:\上课\python\爬虫\chromedriver86.0.4240.22.exe")  # chromedriver所在路径
    driver.get('https://nba.hupu.com/standings')
    html = pq(driver.find_element_by_css_selector('html').get_attribute('outerHTML'))
    trs = html('table.players_table tr')
    arr = []
    for tr in trs:
        tr = pq(tr)
        if tr.attr('class') != None:
            continue
        tds = tr('td')
        team_name = pq(tds[1]).text()
        win = pq(tds[2]).text()
        lose = pq(tds[3]).text()
        arr.append({
            'team_name':team_name,
            'win':win,
            'lose':lose
        })
    saveCSV('langiu.csv',arr)

if __name__ == '__main__':
    main()
```

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | | team_name | win | lose | | | |
| 2 | 0 | 雄鹿 | 53 | 12 | | | |
| 3 | 1 | 猛龙 | 46 | 18 | | | |
| 4 | 2 | 凯尔特人 | 43 | 21 | | | |
| 5 | 3 | 热火 | 41 | 24 | | | |
| 6 | 4 | 步行者 | 39 | 26 | | | |
| 7 | 5 | 76人 | 39 | 26 | | | |
| 8 | 6 | 篮网 | 30 | 34 | | | |
| 9 | 7 | 魔术 | 30 | 35 | | | |
| 10 | 8 | 奇才 | 24 | 40 | | | |
| 11 | 9 | 黄蜂 | 23 | 42 | | | |
| 12 | 10 | 公牛 | 22 | 43 | | | |
| 13 | 11 | 尼克斯 | 21 | 45 | | | |
| 14 | 12 | 活塞 | 20 | 46 | | | |
| 15 | 13 | 老鹰 | 20 | 47 | | | |
| 16 | 14 | 骑士 | 19 | 46 | | | |
| 17 | 15 | 湖人 | 49 | 14 | | | |
| 18 | 16 | 快船 | 44 | 20 | | | |
| 19 | 17 | 掘金 | 43 | 22 | | | |
| 20 | 18 | 爵士 | 41 | 23 | | | |
| 21 | 19 | 雷霆 | 40 | 24 | | | |
| 22 | 20 | 火箭 | 40 | 24 | | | |
| 23 | 21 | 独行侠 | 40 | 27 | | | |
| 24 | 22 | 灰熊 | 32 | 33 | | | |
| 25 | 23 | 开拓者 | 29 | 37 | | | |
| 26 | 24 | 鹈鹕 | 28 | 36 | | | |
| 27 | 25 | 国王 | 28 | 36 | | | |
| 28 | 26 | 马刺 | 27 | 36 | | | |
| 29 | 27 | 太阳 | 26 | 39 | | | |
| 30 | 28 | 森林狼 | 19 | 45 | | | |
| 31 | 29 | 勇士 | 15 | 50 | | | |

05

# 第五章 总结

# 总结

要爬取有效数据，首先得能获取页面信息

可以根据实际情况选择不同的方式：

1. 不复杂的页面，不使用ajax异步加载的，可以使用request直接请求数据
2. 需要执行多个接口才能加载完页面的，可以使用webdriver爬取
3. 都行不通可以尝试FD抓包手动判断接口，找规律

学习Python的这短时间以来，觉得Python还是比较简单，容易上手的，就基本语法而言，但是有些高级特性掌握起来还是有些难度，需要时间去消化。

Python给我最大的印象就是语法简洁，就像写伪代码一样，很多其他语言要用很多行才能实现的操作Python可能几行就搞定了

其次就是python运行的方便，不需要编译，并且打包很方便，安装第三方库也很方便，自带pip，在以后的学习中我还会继续学习python~

# 谢谢观看

THANK YOU FOR WATCHING

汇 报 人 ： 杨 立 伦