实验 2-1　Python 爬取京东手机销售历史数据

建议课时：60 分钟

一、实验目的

- 熟练使用 scrapy 爬虫框架；
- 掌握通过关键词搜索爬取数据的方法；
- 熟练编写 python 代码实现数据爬取；

二、实验原理

网络爬虫，是一种按照一定的规则，自动的抓取万维网信息的程序或者脚本。通俗来说就是模拟用户在浏览器上的操作，从特定网站，自动提取对自己有价值的信息。主要通过查找域名对应的 IP 地址、向 IP 对应的服务器发送请求、服务器响应请求，发回网页内容、浏览器解析网页内容四个步骤来实现。

本实验时通过爬虫框架 Scrapy 爬取舆情数据，scrapy 是用纯 Python 实现一个为了爬取网站数据、提取结构性数据而编写的应用框架，用途非常广泛。

框架的力量，用户只需要定制开发几个模块就可以轻松的实现一个爬虫，用来抓取网页内容以及各种图片，非常之方便。

Scrapy 使用了 Twisted(其主要对手是 Tornado)异步网络框架来处理网络通讯，可以加快我们的下载速度，不用自己去实现异步框架，并且包含了各种中间件接口，可以灵活的完成各种需求。

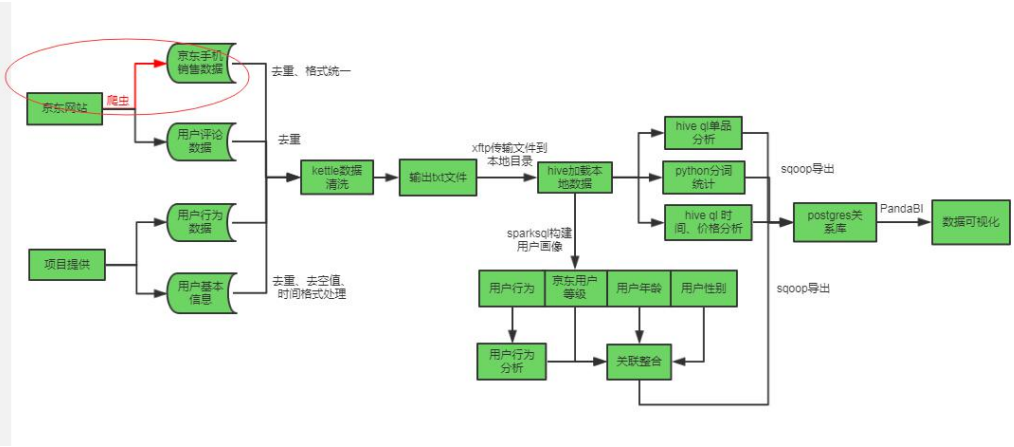
三、实验环境

Dsight 实验室中的 python3 环境、第三方包有 scrapy，re

Pycharm 、NotePad++、Sublime Text 等代码编辑工具

四、实验步骤

本节实验所做内容如下红色标注：

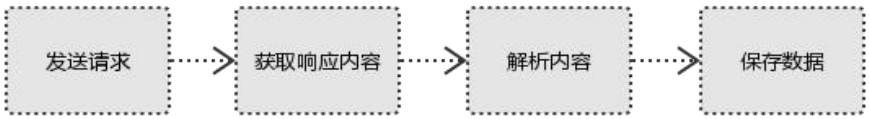


本节实验主要是通过爬虫实现京东手机销售数据的爬取。

1. 实验工具

PyCharm、Sublime Text 等编辑工具

2. 爬虫流程示意图



3. 爬虫步骤

本实验采用 scrapy 爬虫框架编写爬虫脚本，下面选取核心代码讲解爬取京东手机销售数据的爬取逻辑。具体步骤如下：

3.1 获取电商网站目标数据信息

电商网站上手机基本信息如下：

3.2 代码编写

(1) 搜索关键词

根据手机品牌作为搜索关键词：

手机品牌.csv

```
华为（HUAWEI）
小米（MI）
OPPO
三星（SAMSUNG）
魅族（MEIZU）
一加
vivo
诺基亚（NOKIA）
努比亚（nubia）
美图（meitu）
飞利浦（PHILIPS）
黑鲨
锤子（smartisan）
360
中兴（ZTE）
天语（K-Touch）
小辣椒
21KE
酷派（Coolpad）
黑莓（BlackBerry）
索尼（SONY）
纽曼（Newman）
```

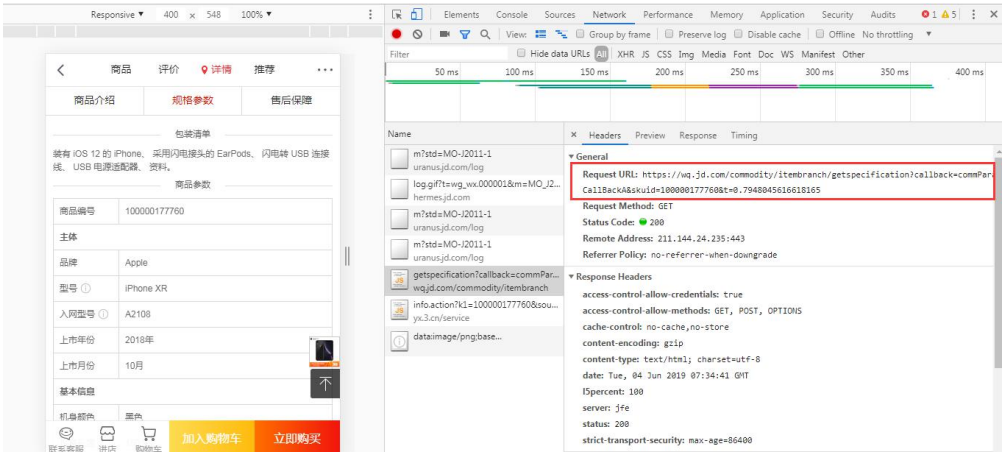
实现代码如下：

```
with open('./mobile_project/data/手机品牌.csv', 'r',    encoding='utf-8')          as f:
csv_reader = csv.reader(f)    # 通过 csv 按行读取
for brand in csv_reader:
        brand = brand[0]
        print('++++++++++crawling:{}'.format(brand))
        if brand.strip():
```

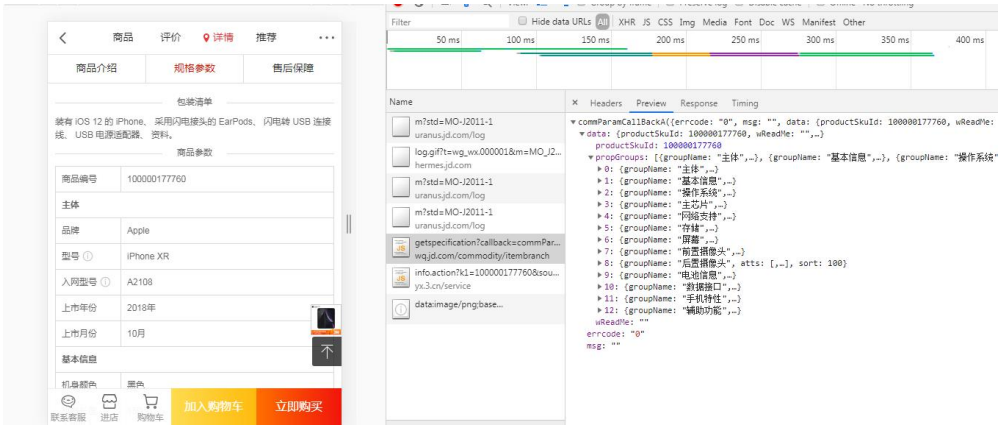
```
                brand = brand.strip() + ' 手机'
                yield Request(jd_search_url.format(kw=brand,
page=page), headers=self.headers,
                              meta={'kw': brand, 'page': page},
callback=self.parse_search_result)
```

(2) 获取访问链接

通过电脑网页访问手机端商品详情页，查看商品详情请求的 api：



(3) 明确解析字段



(4) 解析搜索结果

解析商家信息：

```
if 'data' in json_data and 'searchm' in json_data['data'] and
json_data['data']['searchm']['Paragraph']:
for item in json_data['data']['searchm']['Paragraph']:
```

```
            has_next_page = True
            ret = {}
            content = item['Content']
            ret['name'] = content['warename']
            ret['custom_attr_list'] = content['CustomAttrList']
            ret['shop_name'] = item['shop_name']
            ret['comment_count'] = item['commentcount']
            ret['good_rate'] = item['good']
            ret['shop_id'] = item['shop_id']
            ret['id'] = item['wareid']
            ret['price'] = item['dredisprice']
            ret['url'] =
'https://item.jd.com/{}.html'.format(item['wareid'])
            ret['keyword'] = kw
            yield Request(jd_wine_info_url.format(skuid=ret['id']),
headers=self.headers,
                          meta=ret, callback=self.parse_product_info)
```

解析手机详细配置信息：

```
"""解析商品详细配置信息"""
ret = response.meta
matcher = product_info_ptn.findall(response.text)
if not matcher:
    print('**********get product info error')
    return
json_data = json.loads(matcher[0])
# 商品属性信息，这里直接将属性的中文作为key，方便理解！！！
prop_dict = {}
for prop_group in json_data['data']['propGroups']:
    for attr in prop_group['atts']:
```

```
                prop_dict[attr['attName']] =
    '|'.join(attr['vals'])
    ret['prop'] = prop_dict
    yield ret
```

(5) 循环爬取多页数据

(6) 爬取结果存储到 csv 文件

3.3 爬取数据

运行编写完成的 python 脚本，爬取目标数据

五、实验成果

本次实验完成后，需要得到以下结果：

- 京东手机商品数据爬虫代码编写；
- 爬取数据得到 csv 文件；

爬取结果示例：


```
{
    "name":"努比亚 nubia Z18 全面屏 3.0 极夜黑 8GB+128GB 全网通移
        动联通电信 4G 手机 双卡双待",
    "custom_attr_list":"6.0 英寸^8GB^128GB^2400 万+1600 万像素^骁龙
        845(SDM845)^800 万像素^2160*1080^8.55",
    "shop_name":"努比亚京东自营旗舰店",
    "comment_count":"13266",
    "good_rate":"97",
    "shop_id":"1000001961",
    "id":"100000047414",
    "price":"2549.00",
    "url":"https://item.jd.com/100000047414.html",
    "keyword":"努比亚（nubia） 手机",
    "prop":{
        "品牌":"努比亚（nubia）",
```

"型号":"Z18",
"入网型号":"NX606J",
"上市年份":"2018 年",
"上市月份":"9 月",
"机身颜色":"黑色",
"机身长度（mm）":"148.58",
"机身宽度（mm）":"72.54",
"机身厚度（mm）":"8.55",
"机身重量（g）":"172",
"输入方式":"触控",
"运营商标志或内容":"无",
"机身材质分类":"金属边框|玻璃后盖",
"屏占比":"91.8%",
"操作系统":"Android",
"操作系统版本":"nubia UI6.0",
"CPU 品牌":"骁龙（Snapdragon)",
"CPU 频率":"2.8GHz",
"CPU 核数":"八核",
"CPU 型号":"骁龙 845（SDM845）",
"双卡机类型":"双卡双待单通",
"最大支持 SIM 卡数量":"2 个",
"SIM 卡类型":"Nano SIM",
"4G 网络":"4G：移动（TD-LTE)|4G：联通(FDD-LTE)|4G：电信(FDD-LTE)|4G：联通(TD-LTE)|电信(TD-LTE)",
"3G/2G 网络":"3G：移动(TD-SCDMA)|3G：联通(WCDMA)|3G：电信(CDMA2000)|2G：移动联通(GSM)+电信(CDMA)",
"副 SIM 卡类型":"Nano SIM",
"副 SIM 卡 4G 网络":"4G：移动（TD-LTE)|4G：联通(FDD-LTE)|4G：电信(FDD-LTE)|不支持主副卡同时使用电信卡|4G：联通(TD-LTE)",

"4G+（CA）":"移动 4G+|联通 4G+|电信 4G+",
"高清语音通话（VOLTE）":"移动 VOLTE|电信 VOLTE",
"网络频率（2G/3G）":"2G：GSM 850/900/1800/1900|2G：CDMA 800|3G ： TD-SCDMA 1900/2000|3G ： WCDMA 850/900/1900/2100|3G ： CDMA2000|2G：GSM 900/1800|2G：GSM 900/1800/1900|3G：CDMA 800MHz 1X&EVDO|3G：WCDMA：850/900/1700/1900/2100MHz|TD-SCDMA1880/2010",
"是否支持同时使用联通卡":"支持双卡同时在线，并同时使用联通 4G 移动数据",
"ROM":"128GB",
"ROM 类型":"UFS",
"RAM":"8GB",
"RAM 类型":"LPDDR 4X",
"存储卡":"不支持",
"主屏幕尺寸（英寸）":"6.0 英寸",
"分辨率":"2160*1080",
"屏幕像素密度（ppi）":"403",
"屏幕材质类型":"LTPS",
"屏幕生产厂商":"JDI",
"亮度":"500(type)",
"对比度":"1500（type）",
"前置摄像头":"800 万像素",
"前摄光圈大小":"f/2.0",
"美颜技术":"支持",
"摄像头数量":"2 个",
"后置摄像头":"2400 万+1600 万像素",
"摄像头光圈大小":"其他",
"闪光灯":"双色温灯",
"副摄像头光圈大小":"其他",
"拍照特点":"防抖|美颜|连拍|微距|全景|滤镜|场景模式|HDR|PDAF|

```
        微信小视频|水印",
        "电池容量（mAh）":"3450",
        "电池类型":"锂电池",
        "电池是否可拆卸":"否",
        "充电器":"9V/2A",
        "数据传输接口":"WIFI|NFC|蓝牙|WiFi 热点|OTG 接口",
        "NFC/NFC 模式":"支持（点对点模式）|支持（读卡器模式）|支持（卡模式）|支持卡模拟",
        "耳机接口类型":"Type-C",
        "充电接口类型":"Type-C",
        "数据线":"USB2.0",
        "指纹识别":"支持",
        "语音识别":"支持",
        "GPS":"支持",
        "电子罗盘":"支持",
        "陀螺仪":"支持",
        "红外遥控":"不支持",
        "其他":"距离感应|呼吸灯|多麦降噪技术|光线感应",
        "常用功能":"录音|便签|重力感应"
    }
}
```


选取其中需要的字段输出到 csv 文件：

商品ID	商品名称	商品基本属	店铺名称	店铺ID	商品链接	价格	搜索关键词	评论数	好评率	品牌	型号	机身颜色	上市时间	操作系统
1E+10	酷派（Coo	5.5英寸^3	酷派京东自	1E+09	https://i	499	酷派（Coo	6321	97	酷派（Coo	酷玩7C	波尔多红	2018年8月	Android
1E+10	联想火辣椒	499流明以	联想火辣椒	1E+09	https://i	1599	联想（Len	157	99	联想火辣椒	T3			安卓
1E+10	华为 HUAW	6.3英寸^6	华为京东自	1E+09	https://i	1799	华为（HUA	39599	99	华为（HUA	麦芒7	铂光金	2018年9月	Android
1E+10	莫凡 华为	硅胶	莫凡京东自	1E+09	https://i	39	荣耀	12970	99	莫凡(mofi				
1E+10	【两片装】	前膜^高清	悦可京东自	1E+09	https://i	29.9	D.PHONE	286	94	悦可				
1E+10	锤子（smartisan）手		锤子科技京	1E+09	https://i	59	锤子（sma	32934	99	锤子				
1E+10	锤子（sma	Type-C接口	锤子科技京	1E+09	https://i	19	锤子（sma	20962	99	锤子				
1E+10	努比亚 nu	6.0英寸^8	努比亚京东	1E+09	https://i	2549	努比亚（n	13266	97	努比亚（n	Z18	黑色	2018年9月	Android
1E+10	摩托罗拉	4GB^64GB	摩托罗拉手	1E+09	https://i	1899	摩托罗拉	121	97	摩托罗拉（	XT1941-2	黑色	2018年8月	Android
1E+10	飞利浦 充	双向快充^	飞利浦手机	1E+09	https://i	149	飞利浦（P	18198	98	飞利浦	DLP8712C			
1E+10	【限量送券	前膜^高清	KOOLIFE京	1E+09	https://i	19.9	小米（MI）	28516	96	KOOLIFE				
1E+10	KOOLIFE	3前膜^高清	KOOLIFE京	1E+09	https://i	22.9	360手机	11715	96	KOOLIFE				
1E+10	天语（K-T	2.4英寸^3	天语手机京	1E+09	https://i	199	天语（K-T	859	98	天语（K-T	V6	黑色	2019年以前	以官网信息为准
1E+10	天语（K-T	2.4英寸^3	天语手机京	1E+09	https://i	199	天语（K-T	335	98	天语（K-T	V6	金色	2019年以前	以官网信息为准
1E+10	天语（K-T	2.4英寸^3	天语手机京	1E+09	https://i	199	天语（K-T	2040	99	天语（K-T	V6	金色	2019年以前	以官网信息为准
1E+10	荣耀8X Ma	4GB^64GB	荣耀京东自	1E+09	https://i	1399	华为（HUA	259817	99	华为（HUA	荣耀8X Ma	幻夜黑	2018年9月	Android
1E+10	飞利浦（P	2.0英寸^30	万像素^2(	0	https://i	198	飞利浦（P	6616	98	飞利浦（P	飞利浦E20	红色	2018年7月	其他
1E+10	亿色(ESR)	全屏膜^防	ESR京东自	1E+09	https://i	35.9	D.PHONE	130907	97	亿色（ESR）				
1E+10	摩托罗拉m	6.2英寸^6	摩托罗拉手	1E+09	https://i	1999	摩托罗拉	67	86	摩托罗拉（	XT1929-15	黑色	2018年8月	Android
1E+10	OPPO A5	6.2英寸^3	OPPO京东自	1E+09	https://i	999	OPPO	17356	98	OPPO	A5	幻镜粉	2018年7月	Android
1E+10	联想Lenovo	车载无线^	联想（Len	1E+09	https://i	169	联想（Len	2130	98	联想				
1E+10	Freeson 3	全屏膜^无	Freeson京	1E+09	https://i	29.9	D.PHONE	935	98	Freeson				
1E+10	品胜（PIS	前膜^防偷	品胜京东自	1E+09	https://i	26	D.PHONE	22938	96	品胜（PISEN）				
1E+10	华硕 ASUS	20.7英寸^	华硕京东自	1E+09	https://i	2049	华硕（ASU	4376	98	ASUS华硕	ME16AP			
1E+10	Freeson 3	全屏膜^无	Freeson京	1E+09	https://i	19.9	360手机	869	98	Freeson				
1E+10	斯泰克 苹	前膜^高清	斯泰克（S	1E+09	https://i	39	D.PHONE	200	98					