实验 2-2　Python 爬取用户评论信息

建议课时：60 分钟

一、实验目的

- 熟练使用 scrapy 爬虫框架；
- 掌握 python 爬取数据的思路；
- 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。

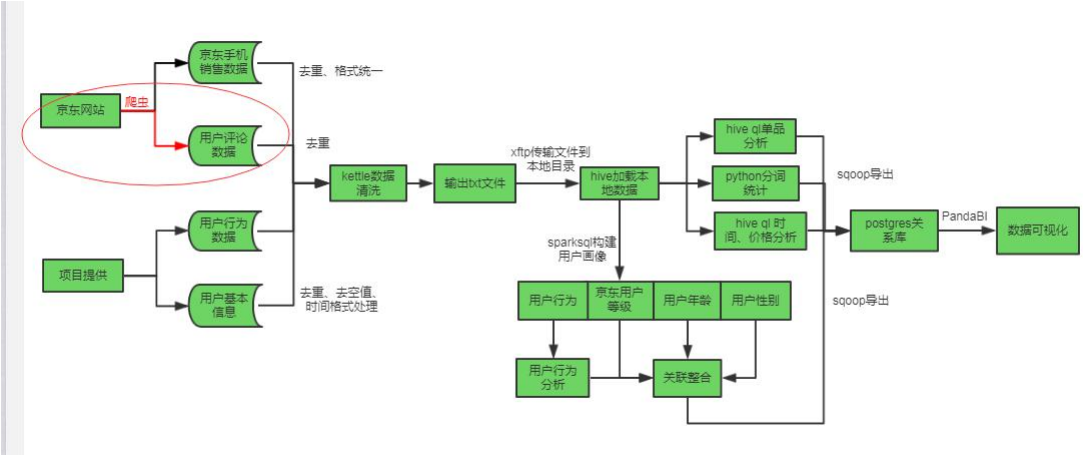
三、实验环境

Dsight 实验室中的 python3 环境

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：

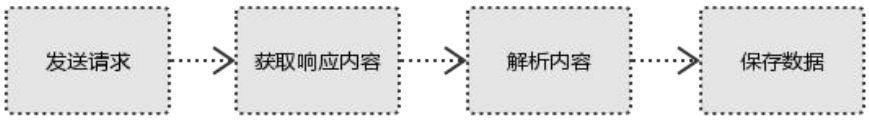


本节实验主要是通过爬虫实现京东用户评论信息的数据爬取。

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用户手机评论数据的爬取逻辑。具体步骤如下：

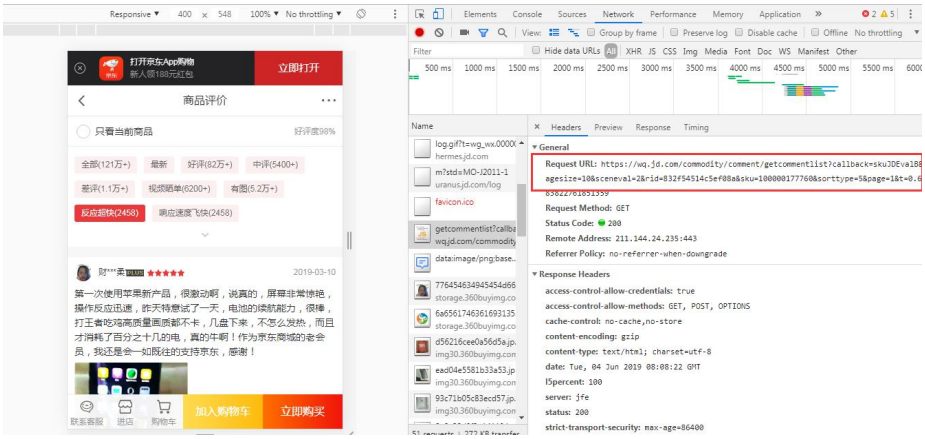
3.1 获取电商网站目标数据信息

电商网站上用户评论信息如下：

3.2 代码编写

(1) 获取用户评论信息的网页链接

通过电脑网页访问手机端用户评论页，查看评论详情请求的 api：



(2) 解析网页内容

```
json_data = json.decode(matcher[0])    # 必须使用`demjson`模块进行解析。原始json 模块解析会出错
for item in json_data['result']['comments']:
        ret = {}
        has_next_page = True
        ret['id'] = item['id']
        ret['keyword'] = product_id
        ret['content'] = item['content']
        ret['post_at'] = item['creationTime']
```

```
            ret['image_count'] = item['imageCount']
            ret['is_mobile'] = item['isMobile']
            ret['mobile_version'] = item['mobileVersion']
            ret['user_name'] = item['nickname']
            ret['product_color'] = item['productColor']
            ret['product_sales'] = '|'.join(item['productSales'])
            ret['product_size'] = item['productSize']
            ret['recommend'] = item['recommend']
            ret['reply_count'] = item['replyCount']
            ret['score'] = item['score']
            ret['title'] = item['title']
            ret['source'] = item['userClientShow']
            ret['user_level'] = item['userLevelName']
            yield ret
```

(3) 循环获取多页评论数据

(4) 保存爬虫数据到 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：

- 京东用户手机评论数据的爬虫代码编写；
- 爬取数据得到 csv 文件；

爬取结果示例：

```
{
  "id":"11909216182",
  "keyword":"100000047414",
  "content":"电池超耐用玩王者荣耀不卡手机不发热，手机反应速
          度快",
```

"post_at":"2018-09-14 15:37:03",

"image_count":"0",

"is_mobile":true,

"mobile_version":"7.0.2",

"user_name":"jd_087534",

"product_color":"极夜黑",

"product_sales":"",

"product_size":"8GB+128GB",

"recommend":"true",

"reply_count":"2",

"score":"5",

"title":"",

"source":"来自京东 Android 客户端",

"user_level":"钻石会员"

}

保存爬取数据到 csv 文件：

评论ID	商品ID	评论正文	发布时间	用户名	用户等级	商品颜色	商品大小	回复数	评价分数	标题	来源
1E+10	1E+10	说的32寸[	########	M***i	银牌会员	F200 小号款(14-32寸		0	1		来自微信购物
1E+10	1E+10	东西很好者	########	x***d	银牌会员	遮阳板款		0	5		来自京东Android客户端
1.01E+10	1E+10	配件方便妖	########	飘***7	钻石会员	C型(适用于望远镜目镜		0	5		来自京东Android客户端
1E+10	1E+10	好好好好妖	########	q***B	铜牌会员	雨露蓝		0	5		来自京东Android客户端
1.02E+10	1E+10	一般般。。	########	j***8	银牌会员	珍珠白		0	3		来自微信购物
1E+10	1E+10	先试试，看	########	y***d	PLUS会员	雨露蓝		0	5		来自京东Android客户端
1.01E+10	1E+10	质量不错，	########	凌***行	钻石会员	7/6/6S通用4.7寸3D曲		0	5		来自京东iPhone客户端
1.02E+10	1E+10	非常不错	########	Asin潇洒	PLUS会员	2.5D弧边钢化膜		0	5		来自京东Android客户端
1.02E+10	1E+10	非常不错	########	Asin潇洒	PLUS会员	透明手机壳		0	5		来自京东Android客户端
1.02E+10	1E+10	还不错，这	########	s***9	PLUS会员			0	5		来自京东Android客户端
1.01E+10	1E+10	质量不错，	########	廉***巧	铜牌会员	细腻亲肤-宝石蓝+送钢		1	5		
1.01E+10	1E+10	还可以吧！	########	i***T	银牌会员	通用(壹博源III代)座		0	3		来自微信购物
1.02E+10	1E+10	很好，是正	########	奥***6	钻石会员	(5V/2A带线充电器)		1	5		来自京东Android客户端
1E+10	1E+10	手感很不错	########	j***e	铜牌会员	iphone7Plus鸭子		0	5		
1.02E+10	1E+10	好好好好妖	########	j***v	银牌会员	弧边0.26mm-高清钢化		0	5		来自微信购物
1E+10	1E+10	手机壳很漂	########	j***e	铜牌会员	iphone7Plus鸭子		0	5		
1.02E+10	1E+10	亮光	########	z***g	PLUS会员	透明壳		1	5		来自京东iPhone客户端
1.01E+10	1E+10	不错	########	j***l	银牌会员	MateS 2A数据线		0	5		来自京东Android客户端
1.01E+10	1E+10	价格很便宜	########	c***5	金牌会员	X6plus(5.7英寸)全屏		0	5		来自京东Android客户端
1.02E+10	1E+10	挺好	########	任***1	钻石会员	108MF中号(42-60寸)		0	5		
1.02E+10	1E+10	感觉一般呢	########	M***c	银牌会员	基础版-粉色		0	3		来自微信购物
1.02E+10	1E+10	【{}】	########	略***了	金牌会员	2.5D弧边钢化膜		0	5		来自京东iPhone客户端
1.01E+10	1E+10	可以可以。	########	j***9	铜牌会员	月光狼		0	5		来自手机QQ购物
1.01E+10	1E+10	货过差	########	j***j	银牌会员	钢化膜-前膜(0.3直边		0	1		来自京东Android客户端
1.01E+10	1E+10	很不错，很	########	o***Y	铜牌会员	透明手机壳		1	5		来自手机QQ购物
1.03E+10	1E+10	好	########	韩***9	钻石会员	黑色327MT	#NAME?	0	5		来自京东Android客户端