唯品会json爬虫

1 注意事项

- 唯品会所有数据都记录在json格式里面,包括各种评论信息以及图片URL,因此不需要下载html
- 如有必要,请自行学习json格式,不难
- 完成作业几乎所有需要的信息都在这些打包的json里面了,我认为没有再做别的网站爬虫的必要

2 食用方法

- 需要安装`playwright`,请自行安装

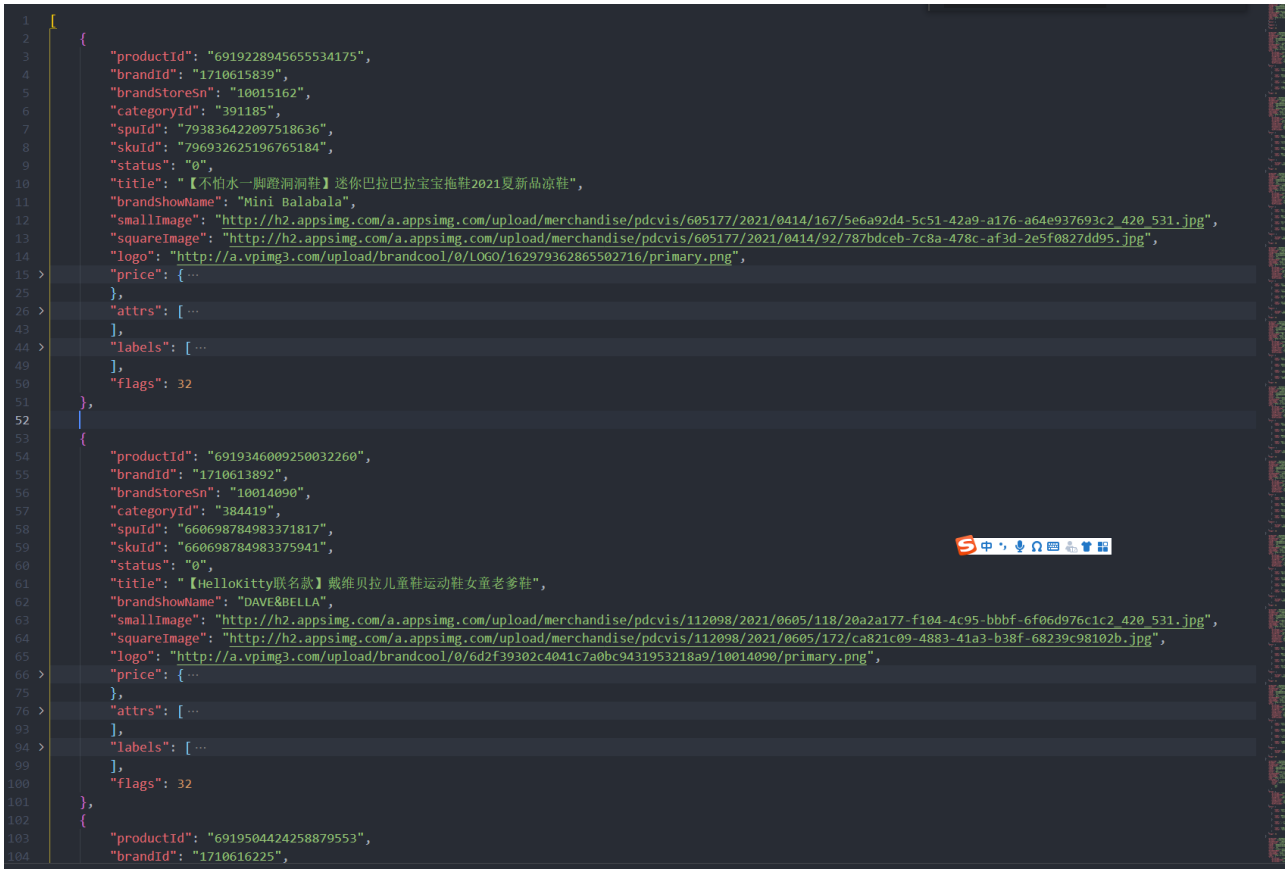
1. 首先爬取商品目录信息,请自行更改`crawler-index-json.py`中的`search_keyword`一项(如选择鞋)
2. 选择爬取页数`n`,一页120个商品,在crawler-index-json.py程序文件中更改
3. 爬取得到目录,所有的商品会以列表形式存储在同一文件夹下的`data-鞋.json`中,后面的爬虫需要依赖这个文件才能运行
4. 在其余crawler程序文件中更改`keyword`一词,使之与`data-{your-keyword}.json`中的`your-keyword`相匹配
5. 运行其余程序,得到相关信息,比如选择关键词"`鞋`",那么第三部得到`data-鞋.json`后,运行`crawler-comment-json.py`得到`comment-鞋.json`文件,其余同理.
6. 爬取评论时,默认每商品爬取100条.

3 唯品会json格式概览

3.1 data-{your-keyword}.json

`data-{your-keyword}.json`

这个文件里包含了那些搜索页中会出现的信息,已经比较详细了



图中是一个`list`,`list`中每个元素是一个商品,注意图中显示全了两个商品(2-51行和53-100行)

单个商品的json代码如下:

```
    {
        "productId": "6919228945655534175",
        "brandId": "1710615839",
        "brandStoreSn": "10015162",
        "categoryId": "391185",
        "spuId": "793836422097518636",
        "skuId": "796932625196765184",
        "status": "0",
        "title": "【不怕水一脚蹬洞洞鞋】迷你巴拉巴拉宝宝拖鞋2021夏新品凉鞋",
        "brandShowName": "Mini Balabala",
        "smallImage":
"http://h2.appsimg.com/a.appsimg.com/upload/merchandise/pdcvis/605177/2021/0414/167/5e6a
92d4-5c51-42a9-a176-a64e937693c2_420_531.jpg",
        "squareImage":
"http://h2.appsimg.com/a.appsimg.com/upload/merchandise/pdcvis/605177/2021/0414/92/787bd
ceb-7c8a-478c-af3d-2e5f0827dd95.jpg",
        "logo":
"http://a.vpimg3.com/upload/brandcool/0/LOGO/162979362865502716/primary.png",
        "price": {
            "priceType": "coupon",
            "priceLabelType": "text",
            "priceLabel": "特卖价",
            "salePrice": "48.9",
```

```
            "salePriceSuff": "",
            "saleDiscount": "3.6折",
            "marketPrice": "139",
            "couponPrice": "48.9",
            "mixPriceLabel": "3.6折"
        },
        "attrs": [
            {
                "name": "适用性别",
                "value": "通用"
            },
            {
                "name": "闭合方式",
                "value": "魔术贴"
            },
            {
                "name": "功能",
                "value": "透气"
            },
            {
                "name": "适用场景",
                "value": "日常"
            }
        ],
        "labels": [
            {
                "bizType": "coupon",
                "value": "券¥15"
            }
        ],
        "flags": 32
    },

```


值得注意的字段有

- `productId`:标识了商品,**非常重要**
- `brandId`: 品牌
- `spuId`: 爬取评论时需要
- `attrs`:属性关键词
- `squareImage`:详细信息里第一张大图的url
- `logo`:不解释
- `price`:不解释

3.2 `detail-{your-keyword}.json`

`detail-{your-keyword}.json`包含了更多地详细信息

很大呢❤

文件是以dict的形式存储的,key就是`productId`!!!

每个key下面有两个dict,分别是product(代表商品)和brand(品牌信息),前者比较重要.

注意字段:

- "detailImages"
- "title": "【不怕水一脚蹬洞洞鞋】迷你巴拉巴拉宝宝拖鞋2021夏新品凉鞋",
- "vipshopPrice": "66",
- "marketPrice": "139",
- "agio": "4.7折",
- "brandStoreName": "Mini Balabala",
- brandStory
- "supportServices":
- longTitle":
- afterSaleServices
- props:更多的属性

3.3 comment-(your-keyword).json

评论文件,**文件是以dict的形式存储的,key就是`productId`!!!**

评论在reputation/content字段下面.