

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个文件夹下的 `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",那么第三步得到 `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

```

1  {
2
3      "productId": "691922894565534175",
4      "brandId": "1710615839",
5      "brandStoreSn": "10015162",
6      "categoryId": "391185",
7      "spuId": "793836422097518636",
8      "skuId": "796932625196765184",
9      "status": "0",
10     "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",
11     "brandShowName": "Mini Balabala",
12     "smallImage": "http://h2.appsing.com/a.appsing.com/upload/merchandise/pdcvis/605177/2021/0414/167/5e6a92d4-5c51-42a9-a176-a64e937693c2_420_531.jpg",
13     "squareImage": "http://h2.appsing.com/a.appsing.com/upload/merchandise/pdcvis/605177/2021/0414/92/787bdceb-7c8a-478c-af3d-2e5f0827dd95.jpg",
14     "logo": "http://a.vpimg3.com/upload/brandcool/0/LOGO/162979362865502716/primary.png",
15 >     "price": { ...
25     },
26 >     "attrs": [ ...
43     ],
44 >     "labels": [ ...
49     ],
50     "flags": 32
51 },
52
53 {
54     "productId": "6919346009250032260",
55     "brandId": "1710613892",
56     "brandStoreSn": "10014090",
57     "categoryId": "384419",
58     "spuId": "660698784983371817",
59     "skuId": "660698784983375941",
60     "status": "0",
61     "title": "【HelloKitty联名款】戴维贝拉儿童鞋运动鞋女童老爹鞋",
62     "brandShowName": "DAVE&BELLA",
63     "smallImage": "http://h2.appsing.com/a.appsing.com/upload/merchandise/pdcvis/112098/2021/0605/118/20a2a177-f104-4c95-bbbf-6f06d976c1c2_420_531.jpg",
64     "squareImage": "http://h2.appsing.com/a.appsing.com/upload/merchandise/pdcvis/112098/2021/0605/172/ca821c09-4883-41a3-b38f-68239c98102b.jpg",
65     "logo": "http://a.vpimg3.com/upload/brandcool/0/6d2f39302c4041c7a0bc9431953218a9/10014090/primary.png",
66 >     "price": { ...
75     },
76 >     "attrs": [ ...
93     ],
94 >     "labels": [ ...
99     ],
100    "flags": 32
101 },
102 {
103     "productId": "6919504424258879553",
104     "brandId": "1710616225",

```

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:

```

1  {
2
3      "productId": "691922894565534175",
4      "brandId": "1710615839",
5      "brandStoreSn": "10015162",
6      "categoryId": "391185",
7      "spuId": "793836422097518636",
8      "skuId": "796932625196765184",
9      "status": "0",
10     "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",
11     "brandShowName": "Mini Balabala",
12     "smallImage":
13     "http://h2.appsing.com/a.appsing.com/upload/merchandise/pdcvis/605177/2021/0414/167/5e6a
14     92d4-5c51-42a9-a176-a64e937693c2_420_531.jpg",
15     "squareImage":
16     "http://h2.appsing.com/a.appsing.com/upload/merchandise/pdcvis/605177/2021/0414/92/787bd
17     ceb-7c8a-478c-af3d-2e5f0827dd95.jpg",
18     "logo":
19     "http://a.vpimg3.com/upload/brandcool/0/LOGO/162979362865502716/primary.png",
20     "price": {
21         "priceType": "coupon",
22         "priceLabelType": "text",
23         "priceLabel": "特卖价",
24         "salePrice": "48.9",

```

```
19         "salePriceSuff": "",
20         "saleDiscount": "3.6折",
21         "marketPrice": "139",
22         "couponPrice": "48.9",
23         "mixPriceLabel": "3.6折"
24     },
25     "attrs": [
26         {
27             "name": "适用性别",
28             "value": "通用"
29         },
30         {
31             "name": "闭合方式",
32             "value": "魔术贴"
33         },
34         {
35             "name": "功能",
36             "value": "透气"
37         },
38         {
39             "name": "适用场景",
40             "value": "日常"
41         }
42     ],
43     "labels": [
44         {
45             "bizType": "coupon",
46             "value": "券¥15"
47         }
48     ],
49     "flags": 32
50 },
51
```

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.