

<https://github.com/hyk-1997/python>

豆瓣爬虫代码:

```
from django.shortcuts import render,HttpResponse

# Create your views here.
import requests
from lxml import etree
from douban.models import Movie,Jd

def crawler(request):
    top250 = []
    for j in range(0, 226, 25):
        url = 'https://movie.douban.com/top250?start=' + str(j)
        res = requests.get(url)
        tree = etree.HTML(res.text)

        for i in range(1, 26):
            movie = []
            movie += tree.xpath('//*[@id="content"]/div/div[1]/ol/li[' + str(i) + ']/div/div[1]/em/text()')
            movie += tree.xpath('//*[@id="content"]/div/div[1]/ol/li[' + str(i) + ']/div/div[2]/div[1]/a/span[1]/text()')

            top250.append(movie)

        save_list = []
        for i in top250:
            save_list.append(Movie(uid=i[0], name=i[1]))
        if Movie.objects.count() == 0:
            Movie.objects.bulk_create(save_list)
            print(save_list)
        return HttpResponse('success')

def home(request):
    return render(request, 'Home.html')

def douban(request):
    movie = Movie.objects.all()
    data = {
        'movies': movie
    }
    return render(request, 'douban.html', context=data)
```

```
def jd(request):
    jd = Jd.objects.all()
    data = {
        'phone': jd
    }
    return render(request, 'jd.html', context=data)
```

京东爬虫代码:

```
# -*- coding: utf-8 -*-
import scrapy
from ..items import JdphoneItem

class JdSpider(scrapy.Spider):
    name = 'jd'
    allowed_domains = ['jd.com']
    start_urls = ['http://jd.com/']

    keyword = "手机"
    page = 1
    url =
'https://search.jd.com/Search?keyword=%s&enc=utf-8&qrst=1&rt=1&stop=1&v
t=2&wq=%s&cid2=653&cid3=655&page=%d&click=0'
    next_url =
'https://search.jd.com/s_new.php?keyword=%s&enc=utf-8&qrst=1&rt=1&stop=
1&vt=2&wq=%s&cid2=653&cid3=655&page=%d&scrolling=y&show_items=%s'

    def start_requests(self):
        yield scrapy.Request(self.url % (self.keyword, self.keyword,
self.page), callback=self.parse)

    def parse(self, response):
        """
        爬取每页的前三十个商品，数据直接展示在原网页中
        :param response:
        :return:
        """
        ids = []
        for li in response.xpath('//*[@id="J_goodsList"]/ul/li'):
            item = JdphoneItem()

            title = li.xpath('div/div/a/em/text()').extract() # 标题
```

```

        price = li.xpath('div/div/strong/i/text()').extract() # 价格
        id = li.xpath('@data-pid').extract() # id
        ids.append(''.join(id))

        url = li.xpath('div/div[@class="p-name
p-name-type-2"]/a/@href').extract() # 需要跟进的链接

        item['title'] = ''.join(title)
        item['price'] = ''.join(price)
        item['url'] = ''.join(url)

        if item['url'].startswith('//'):
            item['url'] = 'https:' + item['url']
        elif not item['url'].startswith('https:'):
            item['info'] = None
            yield item
            continue

        yield scrapy.Request(item['url'], callback=self.info_parse,
meta={"item": item})

        headers = {'referer': response.url}

        self.page += 1
        yield scrapy.Request(self.next_url % (self.keyword, self.keyword,
self.page, ''.join(ids)),
                            callback=self.next_parse, headers=headers)

    def next_parse(self, response):
        """
        爬取每页的后三十个商品，数据展示在一个特殊链接中：url+id(这个 id 是前三
        十个商品的 id)
        :param response:
        :return:
        """
        lis = response.xpath('//li[@class="gl-item"]')
        for li in response.xpath('//li[@class="gl-item"]'):
            item = JdphoneItem()
            title = li.xpath('div/div/a/em/text()').extract() # 标题
            price = li.xpath('div/div/strong/i/text()').extract() # 价格
            url = li.xpath('div/div[@class="p-name
p-name-type-2"]/a/@href').extract() # 需要跟进的链接

```

```

        item['title'] = ''.join(title)
        item['price'] = ''.join(price)
        item['url'] = ''.join(url)

        if item['url'].startswith('//'):
            item['url'] = 'https:' + item['url']
        elif not item['url'].startswith('https:'):
            item['info'] = None
            yield item
            continue

        yield scrapy.Request(item['url'], callback=self.info_parse,
meta={"item": item})

    if self.page <10:
        self.page += 1
        yield scrapy.Request(self.url % (self.keyword, self.keyword,
self.page), callback=self.parse)

    def info_parse(self, response):
        """
        链接跟进，爬取每件商品的详细信息,所有的信息都保存在 item 的一个子字段
info 中
        :param response:
        :return:
        """
        item = response.meta['item']

        yield item

```

- 1

肖申克的救赎
- 2

霸王别姬
- 3

这个杀手不太冷
- 4

阿甘正传
- 5

美丽人生
- 6

泰坦尼克号
- 7

千与千寻
- 8

辛德勒的名单
- 9

盗梦空间
- 10

忠犬八公的故事
- 11

机器人总动员
- 12

三傻大闹宝莱坞
- 13

海上钢琴师
- 14

放牛班的春天
- 15

楚门的世界
- 16

大话西游之大圣娶亲
- 17

星际穿越
- 18

龙猫
- 19

教父
- 20

KindEditor使用 - whitesky-root - 博客园

- 57

饮食男女
- 58

狮子王
- 59

美丽心灵
- 60

情书
- 61

钢琴家
- 62

本杰明·巴顿奇事
- 63

美国往事
- 64

看不见的客人
- 65

黑客帝国
- 66

小鞋子
- 67

西西里的美丽传说
- 68

大闹天宫
- 69

哈利·波特与魔法石
- 70

让子弹飞
- 71

拯救大兵瑞恩
- 72

致命魔术
- 73

七宗罪
- 74

被嫌弃的松子的一生
- 音乐之声

1	黑鲨游戏2 8GB+128GB 暗影黑 骁龙855 Magic Press立体操控 塔式全域制冷 全面屏 全网通4G 双卡双待	3499.0
2	小米9 4800万超广角三摄 8GB+128GB全息幻彩蓝 骁龙855 全网通4G 双卡双待 水滴全面屏拍照游戏智能	3299.0
3	小米 红米6A AI美颜 3GB+32GB 流沙金 全网通4G 双卡双待	649.0
4	vivo S1 6GB+128GB 冰湖蓝 2480万AI高清自拍 超广角后置三摄拍照 移动联通电信全网通4G	2298.0
5	vivo Z3 6GB+64GB 极光蓝 性能实力派 全面屏游戏 移动联通电信全网通4G	1598.0
6	Apple iPhone X (A1865) 64GB 深空灰色 移动联通电信4G	6199.0
7	Apple iPhone 7 (A1660) 128G 黑色 移动联通电信4G	3199.0
8	vivo X27 8GB+256GB大内存 雀羽蓝 4800万AI三摄全面屏拍照 移动联通电信全网通4G	3598.0
9	小米 红米Redmi 7 幻彩渐变AI双摄 3GB+32GB 梦幻蓝 全网通4G 双卡双待 水滴全面屏拍照游戏智能	799.0
10	小米8SE 全面屏智能游戏拍照 6GB+64GB 灰色 骁龙710处理器 全网通4G 双卡双待	1399.0
11	小米8青春版 镜面渐变AI双摄 6GB+64GB 梦幻蓝 骁龙 全网通4G 双卡双待 全面屏拍照游戏智能	1499.0
12	Apple iPhone XS Max (A2104) 256GB 深空灰色 移动联通电信4G 双卡双待	9699.0
13	小米 红米6 4GB+64GB 铂银灰 全网通4G 双卡双待	799.0
14	华为 HUAWEI Mate 20 麒麟980AI智能芯片全面屏超微距影像超大广角徕卡三摄6GB+128GB亮黑色全网通版双4G	3989.0
15	Apple iPhone 8 (A1863) 64GB 深空灰色 移动联通电信4G	3799.0
230	小米9se 幻彩蓝 6G+128G 全网通	2349.0
231	华为（HUAWEI） 华为畅享9 Plus 极光紫 全网通(6G+128G)	1818.0
232	一加6 8GB+256GB 墨岩黑 全面屏双摄游戏 全网通4G 双卡双待	3199.0
233	华为 HUAWEI P30 【分期免息现货当天发送6重好礼】麒麟980全面屏屏内指纹超感光徕卡三摄 天空之境 8+128GB全网通	4488.0
234	飞利浦 (PHILIPS) E279 极速灰 直板按键 移动联通2G 老人 老年机 老年功能机	198.0
235	华为（HUAWEI） 荣耀10青春版 渐变蓝 全网通(6G+128G)	1899.0
236	飞利浦 (PHILIPS) E212A 深银色 翻盖按键 长待机 移动联通2G 双卡双待 老人 学生备用功能机	288.0
237	Meitu 美图T9 浆果红 4GB+64GB 骁龙 全身美型 拍照 全面屏 脸部解锁 双卡双4G 全网通	1999.0
238	小米9 骁龙855 游戏 幻彩蓝 全网通 8G+128G	3299.0
239	华为p30麒麟980AI智能芯片全面屏屏内指纹超感光徕卡三摄 天空之境 全网通 （8G+128G）	4388.9
240	努比亚（nubia） V18 全面屏 长续航 4GB+64GB 曜石黑 移动联通电信全网通4G 双卡双待	999.0
241	华为（HUAWEI） 荣耀畅玩7 2GB+16GB 智能老人机 全网通 全面屏智能 金色	517.0
242	朵唯V33 18:9全面屏移动联通电信全网通双卡双待超薄4G老人女性学生智能 渐变蓝 2GB+16GB	379.0
243	【自营配送 豪礼套装】三星Galaxy A6s (SM-G6200)全面屏 性价比智能 撒浪黑 全网通(6G+64G)+3期免息套装	1428.0
244	OPPO R17 【低至2399元！6期免息+赠碎屏险+蓝牙耳机+私人定制】屏幕指纹 全网通 流光蓝（6G+128G） 官方标配	2399.0
245	华为（HUAWEI） 华为畅享9 幻夜黑 全网通(4G+64G)（分期免息0首付）	1128.0
246	华为（HUAWEI） 华为mate20 【6期免息送原装皮套+1年碎屏险】 极光色 (6G+64G全网通)	3499.0

id	uid	name
14	14	放牛班的春天
15	15	楚门的世界
16	16	大话西游之大圣娶亲
17	17	星际穿越
18	18	龙猫
19	19	教父
20	20	熔炉
21	21	无间道
22	22	疯狂动物城
23	23	当幸福来敲门
24	24	怦然心动
25	25	触不可及
26	26	蝙蝠侠：黑暗骑士
27	27	乱世佳人
28	28	活着
29	29	少年派的奇幻漂流
30	30	控方证人
31	31	天堂电影院
32	32	鬼子来了

id	name	price
131	华为 (HUAWEI) 5	1199
132	vivo 【新品上市】 X	3198
133	努比亚 nubia X 双屏	3399
134	小米 (MI) 红米6A	518
135	华为 (HUAWEI) 5	535
136	小米 红米Note7 Pro	1649.8
137	华为 HUAWEI 麦芒	1889
138	vivo S1 2480万AI	2298
139	联想Z5 6GB+64GB	998
140	努比亚 nubia Z18	2699
141	华为P30 Pro 超感	6788
142	vivo NEX 双屏版 AI	4998
143	OPPO 【6期免息】	2999
144	OPPO 【3期免息】	1399
145	努比亚 nubia Z18n	1599
146	诺基亚 NOKIA 811	499
147	飞利浦 (PHILIPS)	99
148	华为 (HUAWEI) 华	1178
149	OPPO Find X曲面	4999

+

-

✓

✕

↺

■





