



链滴

Scrapy | 豆瓣电影评分爬虫

作者: [fpdan](#)

原文链接: <https://ld246.com/article/1590763809675>

来源网站: [链滴](#)

许可协议: [署名-相同方式共享 4.0 国际 \(CC BY-SA 4.0\)](#)

本 demo 只做学习之用，请勿转载和商用，谢谢
y

观察链接

https://movie.douban.com/explore#!type=movie&tag=冷门佳片&sort=recommend&page_limit=20&page_start=0

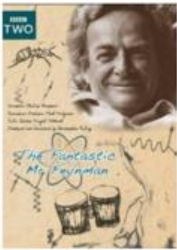
豆瓣电影 搜索电影、电视剧、综艺、影人

影讯&购票 选电影 电视剧 排行榜 分类 影评 2019年度榜单 2019书影音报告

选电影

热门 最新 经典 可播放 豆瓣高分 冷门佳片 华语 欧美 韩国 日本 动作
喜剧 爱情 科幻 悬疑 恐怖 治愈

☒ 按热度排序 ☐ 按时间排序 ☐ 按评价排序 ☐ 我没看过的 ☐ 可在线播放

 默片解说员 8.4	 2008年第29届北京奥运会 9.6	 伊凡王子与百变公主 8.3	 大地 8.7	 黄子华栋笃笑之金盆囍口 9.5
 少年泰坦出击大战少年泰坦 8.4	 神奇的费曼先生 9.2	 搬家 8.5	 福星小子2: 绮丽梦中人 8.8	 我们的故事2 8.1

创建爬虫

```
# 创建项目
$ scrapy startproject douban
# 创建爬虫
$ scrapy genspider doubanSpider https://www.douban.com/
```

doubanSpider.py

```
# -*- coding: utf-8 -*-
import scrapy
import json
from douban.items import DoubanItem

class DoubanSpider(scrapy.Spider):
    name = 'doubanSpider'
    allowed_domains = ['https://movie.douban.com/']
    start_urls = [
        "https://movie.douban.com/j/search_subjects?type=movie&tag=冷门佳片&sort=recommend&page_limit=20&page_start="
        + str(x) for x in range(1, 50, 1)]

    def parse(self, response):
        rs = json.loads(response.text)
        datas = rs.get("subjects")
        # items.py 中定义的参数
        item = DoubanItem()
        for data in datas:
            item['title'] = data.get('title')
            item['rate'] = data.get('rate')
            item['url'] = data.get('url')
            item['id'] = data.get('id')
            yield item
```

items.py

```
# -*- coding: utf-8 -*-
import scrapy

class DoubanItem(scrapy.Item):
    title = scrapy.Field()
    rate = scrapy.Field()
    url = scrapy.Field()
    id = scrapy.Field()
    cover = scrapy.Field()
```

pipelines.py

```
# -*- coding: utf-8 -*-

class DoubanPipeline:
    def process_item(self, item, spider):
        with open('text.txt', 'a') as f:
            f.write(item['id'] + "," + item['rate'] + "," + item['title'] + "," + item['url'] + "\n")
            f.close()
        return item
```

特别注意 setting.py 文件

```
# Crawl responsibly by identifying yourself (and your website) on the user-agent
# USER_AGENT 需要，不开启会一直 403 拒绝访问
```

```

USER_AGENT = 'douban (+http://www.yourdomain.com)'

# Obey robots.txt rules
# 忽略 robots.txt 文件
ROBOTSTXT_OBEY = False

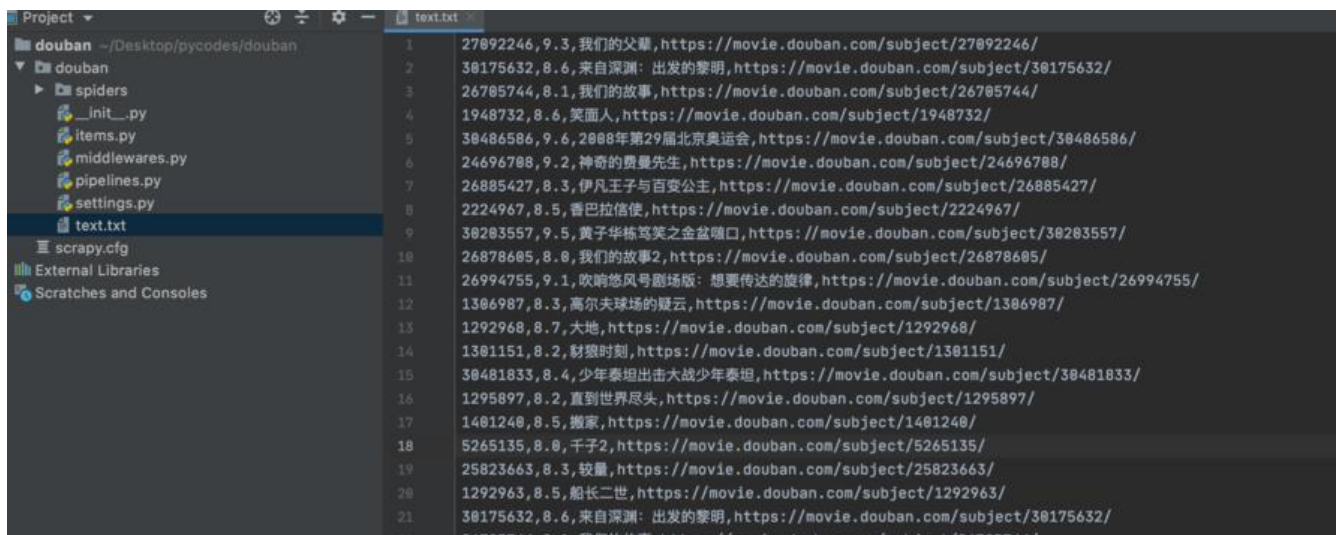
# 做一个有道德的人 设置一下下载延迟
DOWNLOAD_DELAY = 3

# Configure item pipelines
# See https://docs.scrapy.org/en/latest/topics/item-pipeline.html
# ITEM_PIPELINES 记得开启一下
ITEM_PIPELINES = {
    'douban.pipelines.DoubanPipeline': 300,
}

```

运行爬虫

\$ scrapy crawl doubanSpider



数据入库

pipelines.py

```

# -*- coding: utf-8 -*-

# Define your item pipelines here
#
# Don't forget to add your pipeline to the ITEM_PIPELINES setting
# See: https://docs.scrapy.org/en/latest/topics/item-pipeline.html

# 安装 Mysqlclient
# pip3 install Mysqlclient
# In -s /usr/local/mysql/bin/mysql_config /usr/local/bin/mysql_config
# 参考 https://www.jianshu.com/p/6411c14ce3f1

# scrapy crawl doubanSpider

```

```
# 运行爬虫的时候可能 MySQLdb 遇到个坑
# ImportError: dlopen(/Users/xxx/.local/share/virtualenvs/MyDjango-c9TXLMy3/lib/python3.6
site-packages/MySQLdb/_mysql.cpython-36m-darwin.so, 2): Library not loaded: libcrypto.1.0.0
dylib
# 参考这里 https://www.cnblogs.com/Peter2014/p/10937563.html
import MySQLdb

class DoubanPipeline:
    def __init__(self):
        self.conn = MySQLdb.connect('localhost', 'root', '123456', 'python_spider', charset="utf8",
use_unicode=True)
        self.cursor = self.conn.cursor()

    def process_item(self, item, spider):
        # 写文件
        # with open('text.txt', 'a') as f:
        #     f.write(item['id'] + "," + item['rate'] + "," + item['title'] + "," + item['url'] + "\n")
        #     f.close()
        # return item

        # 入库
        insert_sql = "insert into douban(movie_id, title, rate, url)VALUES (%s, %s, %s, %s)"
        self.cursor.execute(insert_sql, (item['id'], item['rate'], item['title'], item['url']))
        self.conn.commit()
        print('正在插入数据...')
        return item

    def close_spider(self, spider):
        # 接收结束信号
        self.conn.close()
        print('完成数据插入...')
```

项目地址

https://github.com/GlacierBo/pycodes/tree/master/douban_spider