# <JDItemSpider>开发文档

版本修订日志

|  |  |  |
| --- | --- | --- |
| 修订时间 | 修订人 | 修订内容 |
| 2016年10月19日 | 杨恬 | 创建 |
|  |  |  |
|  |  |  |
|  |  |  |

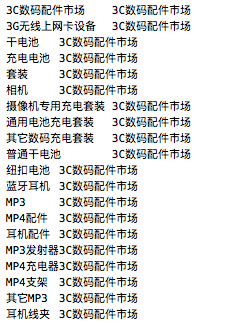
## 1．1 JDItemSpider 京东商品数量抓取程序

1. 程序编号： JDItemSpider
2. 程序名称：京东商品数量抓取程序
3. 程序说明：本程序为爬取京东相关商品数量的程序，目前是为判断商品相关的词语的词权重功能研发，主要研发思路：根据商品名称找到京东的搜索结果页，抓取其上的商品数量，包含的程序模块如下：

模块A：数据抓取模块

1. 程序功能可视化说明：

对于如下输入格式的文件：



程序先找出其中拥有多个nature的商品，记录在一个临时文件中。在处理时，对于单个nature的商品，搜索时不加nature；多个nature的商品，搜索时加上nature。访问如下的搜索网页：



抓取其中的商品数量，输出如下格式的文件：



1. 程序输入与输出说明：

输入格式: 商品名称 nature

输出格式: 商品名称 nature 商品数量

1. 处理过程流程图：

../../jd_p.pdf

1. 开放接口（若被调用，需要提供开放的接口说明）：

暂无

1. 引入的包及接口：

scrapy: 运用其中的Selector方法解析html和定位元素

1. 程序源码目录结构说明：

.

├── \_\_configure\_\_

│   ├── ConfigureParser.py

│   ├── \_\_init\_\_.py

│   └── configure.xml

├── \_\_doc\_\_

│   └── JDItemSpider文档.docx

├── \_\_file\_\_

│   ├── du.txt

│   ├── jd.txt

│   └── jd\_finished.txt

├── \_\_init\_\_.py

├── \_\_spider\_\_

│   ├── \_\_init\_\_.py

│   └── spider.py

├── \_\_util\_\_

│   ├── \_\_init\_\_.py

│   ├── double.sh

│   ├── library.py

├── main.py

└── run.sh

\_\_configure\_\_: 包含配置文件configure.xml以及配置文件解析类ConfigureParser

\_\_doc\_\_: 包含程序文档

\_\_file\_\_: 包含输入输出文件以及临时文件du.txt

\_\_spider\_\_: 包含爬虫代码

\_\_util\_\_: 包含常用的函数以及判别多个nature商品的double.sh脚本

main.py: 程序入口

run.sh：启动脚本

1. 程序启动方式说明

改变启动脚本权限: chmod +x run.sh

执行脚本: ./run.sh

1. 程序配置文件及说明

配置文件: configure.xml

input\_file: 输入文件

output\_file: 输出文件

xpath：数量所在的xpath

1. 部署程序打包方式：

直接放置于服务器

1. 程序部署说明：

修改配置文件configure.xml，添加输入输出文件路径，执行run.sh脚本。

1. 程序潜在的问题，推荐解决方案及其他特殊说明

貌似可能会被封？三万数量五个线程并没有被封过，也加了retry模块进行错误重试，若出现被封的情况需要设置等待时间。

1. 版本号和CHANGES

JDItemSpider 1.0