**电商爬虫描述文档**

Functional description document

# Version History

|  |  |  |  |
| --- | --- | --- | --- |
| **Date** | **Version** | **Author** | **Description** |
| 2017-10-20 | V1.0 | Larry.li | 1.0 version |
| 2017-10-23 | V1.1 | Larry.li | Update 3.1 How to Design |

目录

[Version History 2](#_Toc496607323)

[1. Background 4](#_Toc496607324)

[2. Function Description 4](#_Toc496607325)

[2.1 目标电商网站 4](#_Toc496607326)

[2.2 基本功能 4](#_Toc496607327)

[2.3 爬虫结果 4](#_Toc496607328)

[3. How to Design 5](#_Toc496607329)

[3.1类设计 5](#_Toc496607330)

[3.2使用HTML解析引擎——Jumony.Core 5](#_Toc496607331)

[3.3分析京东/苏宁的网页结构 6](#_Toc496607332)

[3.4问题 6](#_Toc496607333)

# Background

电商网站一般都有一个特点就是有点乱，信息量很多。这边有一些广告，商品中间又夹杂着一些推荐的商品。

有时候，我们只想简单的看清我们搜索商品的简单列表，包含一些商品基本信息。此时，我们就需要一个爬虫小工具来爬取电商上的商品信息，供我们查看。

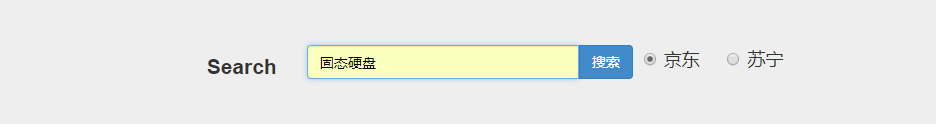
# Function Description

## 2.1 目标电商网站

基本确定京东商城和苏宁易购：

1. 刚好公司发了京东的礼品卡，首选京东
2. 苏宁易购进行扩展
3. 其他如淘宝之类的技术上不容易实现（淘宝防止爬取商品做得比较严格）

## 2.2 基本功能



1. 搜索框

输入商品关键字，点击搜索按钮进行搜索

1. 单选（选择京东or苏宁）

目前做了两个网站的爬虫工具。单选选择并使用1搜索框进行搜索。

1. 分页条



## 2.3 爬虫结果

列表显示搜索结果，包含商品名称、价格、评价数、商品名称、是否京东自营、链接。

苏宁易购的搜索结果类似，也是使用列表显示。字段有所不同。



# How to Design

## 3.1类设计

Product.cs

基本的Product类（包含name字段和price）

JDProduct.cs

继承Product类（新增京东特有字段shop，IsSelf，commit，link）

SNProduct.cs

继承Product类（新增苏宁特有字段shop，commit，link）

## 3.2使用HTML解析引擎——Jumony.Core

<package id="Jumony.Core" version="3.0.18" targetFramework="net452" />

介绍：

<https://github.com/Ivony/Jumony>

Jumony的一切基础建立在服务器端的HTML DOM之上。在服务器端将HTML（文件或动态网页技术的输出）按照客户端浏览器的处理方式解析为HTML DOM。操纵和处理HTML DOM，就像我们在客户端用JavaScript干的那些事情一样，不同的是，Jumony可以使你依托强大的.NET Framework，来解决以前用脚本和服务器端技术都难以解决的事情。

使用：

进行解析HTML，生成实体类。京东or苏宁。

class="p-shop"中data-selfware="1"代表自营，0代表非自营。

## 3.3分析京东/苏宁的网页结构

1）首先是他们搜索的URL分析：

苏宁搜索规则：https://search.suning.com/固态硬盘/&cp=n-1

关键字加页数（页数为n-1）

京东搜索规则：

https://search.jd.com/Search?keyword=固态硬盘&enc=utf-8&page=2n-1

关键字加&enc=utf-8加&page=页数（页数为2n-1）

2）分别分析他们的商品结构

京东：所有的商品都在这节点里面



苏宁：所有的商品都在这节点里面



然后分每个字段，找出他html结构中对应的节点。生成商品实体。

## 3.4问题

1）研究京东页面时。一开始发现HttpWebRequest请求过来的响应总是京东首页。而不是我们要的搜索出来的搜索页面。

后来发现我们需要再请求时request.UserAgent填上浏览器。来伪装成浏览器请求。不然对方服务器总不给返回我们需要的页面。

2）京东每一页只能请求到前29条的商品。后面30条需要你访问网页去用前端的动作触发他的懒加载。目前还没研究出这个问题。所以每页商品列表只有前面的29条数据.

3）苏宁由于他的价格是在图片里。所以这个字段也是显示不出。

4）基本上用最简单的东西写这个爬虫。实现了一点点效果，有待进一步的研究。