**Half.ebay 数据抓取相关说明文档**

|  |  |  |  |
| --- | --- | --- | --- |
| **版本** | **日期** | **说明** | **作者** |
| V1.0 | 2016.10.20 | Half.com数据抓取相关说明文档（初版） | 于永桓 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**目录**

[1 流程图 4](#_Toc305764321)

[1.1 Half.ebay amazon市场抓取asin流程图 4](#_Toc955678352)

[1.2 Half.ebay 根据asin抓取（更新）书籍信息流程图 5](#_Toc1150035752)

[1.2.1 获取书籍信息详细流程图（get\_book\_info函数） 6](#_Toc1901984399)

[1.3 Half.ebay 源市场抓取书籍信息流程图（已停用） 7](#_Toc22631119)

[2 Half.ebay 脚本说明 7](#_Toc709437595)

[2.1 Half.ebay amazon市场抓取asin 7](#_Toc706704152)

[2.1.1 get\_price\_url.py 7](#_Toc2052255682)

[2.1.2 get\_asins.py 7](#_Toc1773728156)

[2.2 Half.ebay 抓取书籍信息 8](#_Toc144488222)

[2.2.1 snatch\_half\_book\_info.py 8](#_Toc82752935)

[2.3 Half.ebay 更新书籍信息 8](#_Toc1957874494)

[2.3.1 update\_half\_book\_info.py 8](#_Toc935138050)

[2.3.2 auto\_update\_price.sh 8](#_Toc416268899)

[2.4 Half.ebay 源市场抓取书籍信息（已停用） 8](#_Toc116373713)

[2.4.1 start.py 8](#_Toc1298019517)

[2.4.2 select.py 8](#_Toc808081845)

[2.4.3 get\_book\_cate.py 8](#_Toc670319091)

[2.4.4 get\_book\_listing.py 8](#_Toc1077646143)

[2.4.5 get\_book\_html.py 8](#_Toc1371882248)

[2.4.6 get\_book\_info.py 8](#_Toc1814433121)

[2.4.7 classify\_book\_info.py 8](#_Toc498689826)

[2.4.8 config.py 9](#_Toc1889388543)

[2.5 其他相关文件 9](#_Toc1338320072)

[3 字段说明 9](#_Toc1491986672)

[3.1 half\_info.csv 9](#_Toc615447157)

[3.2 onshelf\_info.csv 10](#_Toc1895627808)

[3.2.1 onshelf\_info.csv（美国市场更新） 10](#_Toc220444552)

[3.2.2 onshelf\_info.csv（美国市场新上架） 10](#_Toc338408283)

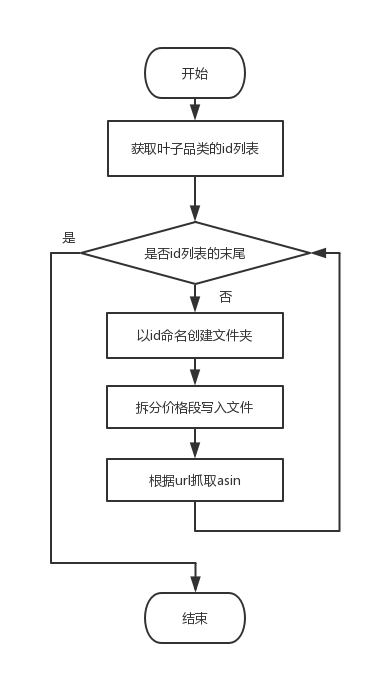
[3.2.3 onshelf\_info.csv(加拿大市场更新） 11](#_Toc1475827904)

[4 服务器 11](#_Toc892450800)

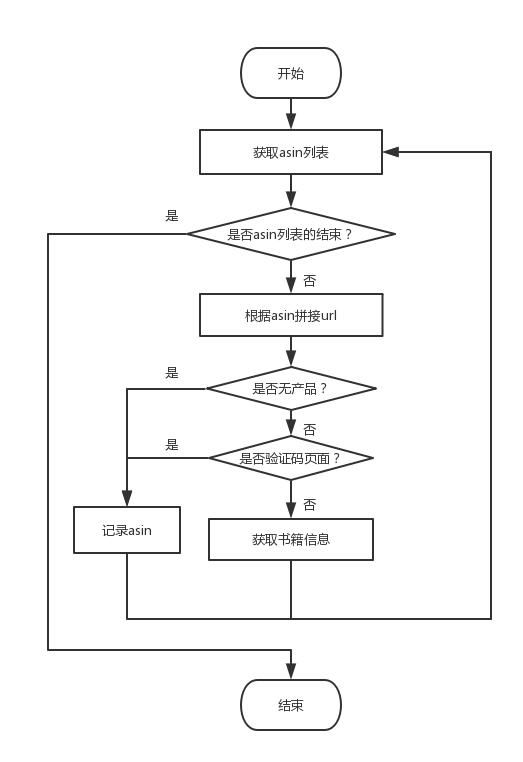
[4.1 half.com价格更新 11](#_Toc644172604)

# 1 流程图

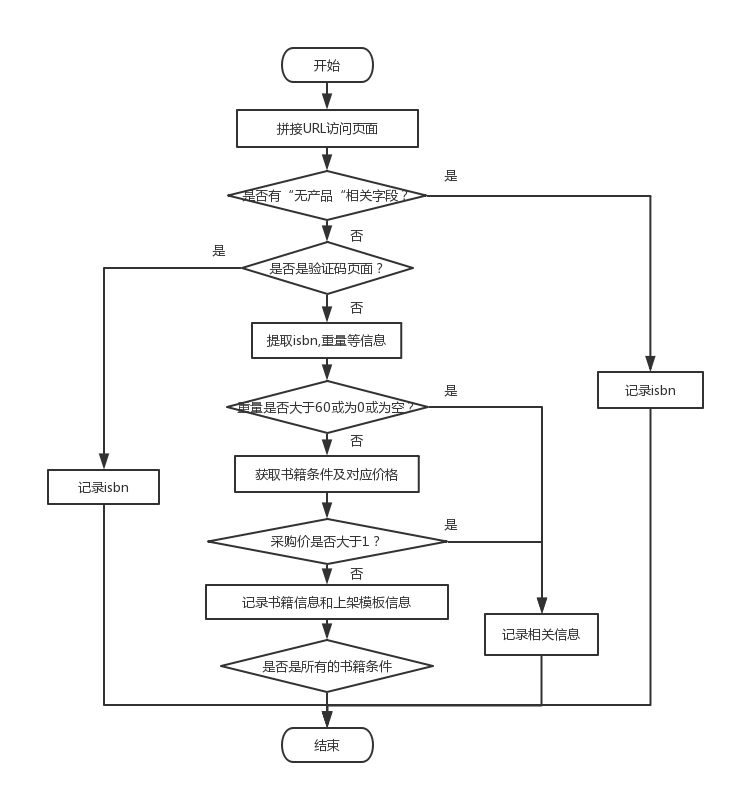
## 1.1 Half.ebay amazon市场抓取asin流程图



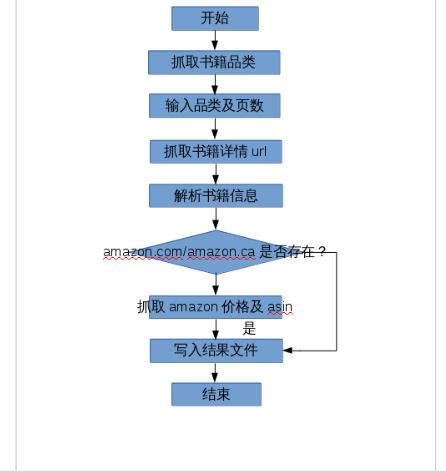
## 1.2 Half.ebay 根据asin抓取（更新）书籍信息流程图



### 1.2.1 获取书籍信息详细流程图（get\_book\_info函数）



## 1.3 Half.ebay 源市场抓取书籍信息流程图（已停用）



# 2 Half.ebay 脚本说明

## 2.1 Half.ebay amazon市场抓取asin

2.1.1 get\_price\_url.py

将品类按价格段拆分成不同的url 输入文件：result/categorys.txt 输出文件：result/\*/price\_url.txt

2.1.2 get\_asins.py

根据拆分的url 解析获取asin 输入文件：result/\*/price\_url.txt 输出文件：result/\*/asins\_more.txt

## 2.2 Half.ebay 抓取书籍信息

2.2.1 snatch\_half\_book\_info.py

根据asin抓取解析书籍信息 输入文件：snatch/asin/\*\*\*.csv 输出文件：new\_info.zip

## 2.3 Half.ebay 更新书籍信息

2.3.1 update\_half\_book\_info.py

根据isbn更新书籍信息 输入文件：update/isbn/\*\*\*.csv 输出文件：update\_info.zip

2.3.2 auto\_update\_price.sh

价格自动更新shell脚本

## 2.4 Half.ebay 源市场抓取书籍信息（已停用）

2.4.1 start.py

程序运行文件，对程序所需的基本信息输入配置 更改config.py

2.4.2 select.py

定义start.py所需的选择函数，被start.py调用 输入文件：cate\_url.txt

2.4.3 get\_book\_cate.py

抓取half.com的书籍品类及其URL,创建或更新书籍品类信息文件时运行 输出文件：cate\_url.txt

2.4.4 get\_book\_listing.py

抓取品类下书籍详情页面的URL 输入文件config.py 输出文件：book\_url.txt

2.4.5 get\_book\_html.py

抓取并保存书籍详情页面 输入文件：book\_url.txt 输出文件：book文件夹

2.4.6 get\_book\_info.py

解析抓取到的书籍页面提取所需信息 输入文件：book文件夹 输出文件：book\_info.txt

2.4.7 classify\_book\_info.py

读取解析到的half.com书籍信息逐条到Amazon抓取相关信息 输入文件：book\_info.txt 输出文件：us\_follow.csv us\_establish.csv ca\_follow.csv ca\_establish.csv

2.4.8 config.py

记录start配置选择

## 2.5 其他相关文件

1、 log文件夹：记录程序运行日志

2、 myutil文件夹：Amazon服务器代理

3、 result文件夹：存放程序运行结果

4、 snatch文件夹：存放书籍信息抓取相关文件

5、 update文件夹：存放价格自动更新相关文件

6、 filter\_new\_asin.py：简单去重脚本

7、 get\_isbn\_from\_file.py：从.xls文件或者.xlsx文件提取isbn并转存为.csv格式

# 3 字段说明

## 3.1 half\_info.csv

|  |  |
| --- | --- |
| 字段 | 说明 |
| 'ISBN' | 10位ISBN |
| 'ISBN13' | 13位ISBN |
| 'weight' | 重量，单位OZ |
| 'condition' | 书籍状况 |
| 'price' | 最低价 |
| 'sec\_price' | 次低价 |

## 3.2 onshelf\_info.csv

|  |  |
| --- | --- |
| **condition** | **代码** |
| brand new | 11 |
| like new | 1 |
| very good | 2 |
| good | 3 |
| acceptable | 4 |

### 3.2.1 onshelf\_info.csv（美国市场更新）

|  |  |
| --- | --- |
| 字段 | 说明 |
| 'sku' | asin&"\_condition代码"&\_O\_MM |
| 'price' | 若存在次低价则次低价为采购价，否则采购价为最低价；（采购价+3.99）\*1.5与采购价+23.99比较取最大值 |
| 'minimum-seller-allowed-price' | 同上 |
| 'maximum-seller-allowed-price' | 8888800 |
| 'quantity' | 1 |
| 'leadtime-to-ship' | 5 |

### 3.2.2 onshelf\_info.csv（美国市场新上架）

|  |  |
| --- | --- |
| 字段 | 说明 |
| 'sku' | asin&"\_condition代码"&\_O\_MM |
| 'product-id' | asin |
| 'product-id-type' | 1 |
| 'price' | 若存在次低价则次低价为采购价，否则采购价为最低价；（采购价+3.99）\*1.5与采购价+23.99比较取最大值 |
| ‘minimum-seller-allowed-price' | 同上 |
| 'maximum-seller-allowed-price' | 8888800 |
| 'item-condition' | condition代码 |
| 'quantity' | 1 |
| 'will-ship-internationally' | 1 |
| 'expedited-shipping' | N |
| 'leadtime-to-ship' | 5 |

### 3.2.3 onshelf\_info.csv(加拿大市场更新）

|  |  |
| --- | --- |
| 字段 | 说明 |
| 'sku' | condition代码"&EVEN&asin&\_ |
| 'price' | 采购价+3.99，若存在次低价则次低价为采购价，否则采购价为最低价 |
| 'minimum-seller-allowed-price' | 同上 |
| 'maximum-seller-allowed-price' | 8888800 |
| 'quantity' | 1 |
| 'leadtime-to-ship' | 5 |

# 4 服务器

4.1 half.com价格更新  
part1 : zhaolei@54.237.244.147  
part2 : zhaolei@54.198.136.195  
part3 : zhaolei@54.84.216.69  
part4 : zhaolei@52.90.26.58  
part5 : zhaolei@54.205.179.66