1. **爬虫脚本目录结构说明**

Spider

丨—continued

丨 丨—log

丨 丨 丨—image\_log.txt 记录未下载成功的图片链接及所属爬虫商品id

丨 丨—comm\_use\_img\_download.py 商品图片下载脚本

丨 丨—translate\_desc.py 爬虫商品描述表中没有中文版的记录新增中文翻译

丨

丨—db 数据库相关脚本目录

丨 丨—daos.py 数据模型存库脚本

丨 丨—dbconnecter.py 数据库链接脚本

丨 丨—models.py 数据模型

丨

丨—tools 工具脚本目录

丨 丨—translate.py 调用Google翻译API实现脚本

丨

丨—netvisiter 网络访问工具目录

丨 丨—net\_openner.py 请求网络资源的实现，请按自己需求在里面配置cookie文件路径。

丨

丨—handler 存放商品爬取实现脚本的目录

丨 丨—base\_product\_spider.py 商品爬虫基础类，爬单个商品的公用代码

丨 丨—base\_url\_spider.py 商品链接爬虫基础类，爬商品链接列表的公用代码

丨 丨—farfetch\_product\_spider.py 爬取farfetch商品的实现脚本

丨 丨—netaporter\_product\_spider.py 爬取netaporter商品的实现脚本

丨 丨—jackwills\_product\_spider.py 爬取jackwills商品的实现脚本

丨 丨—reiss\_product\_spider.py 爬取reiss商品的实现脚本

丨 丨—tedbaker\_product\_spider.py 爬取tedbaker商品的实现脚本

丨 丨—frenchconnection\_product\_spider.py 爬取frenchconnection商品的实现脚本

丨

丨—farfetch

丨 丨—log

丨 丨 丨—farfetch\_log.txt 记录未爬取、未找到sku、未找到图片的爬虫商品

丨 丨—farfetch\_urls\_spider.py 商品列表商品链接爬取实现脚本

丨 丨—grab\_farfetch\_products.py 根据t\_spider\_product\_url表中记录执行商品爬取

丨

丨—frenchconnection

丨 丨—log

丨 丨 丨—frenchconnection\_log.txt 记录未爬取、未找到sku、未找到图片的爬虫商品

丨 丨—frenchconnection\_url\_spider.py 商品列表商品链接爬取实现脚本

丨 丨—grab\_frenchconnection\_product.py 根据t\_spider\_product\_url表中记录执行商品爬取

丨

丨

丨

丨

丨—jackwills

丨 丨—log

丨 丨 丨—jackwills\_log.txt 记录未爬取、未找到sku、未找到图片的爬虫商品

丨 丨—jackwills\_url\_spider.py 商品列表商品链接爬取实现脚本

丨 丨—grab\_jackwills\_product.py 根据t\_spider\_product\_url表中记录执行商品爬取

丨

丨—netaporter

丨 丨—log

丨 丨 丨—netaporter\_log.txt 记录未爬取、未找到sku、未找到图片的爬虫商品

丨 丨—netaporter\_url\_spider.py 商品列表商品链接爬取实现脚本

丨 丨—grab\_netaporter\_prorducts.py 根据t\_spider\_product\_url表中记录执行商品爬取

丨

丨—reiss

丨 丨—log

丨 丨 丨—reiss\_log.txt 记录未爬取、未找到sku、未找到图片的爬虫商品

丨 丨—reiss\_url\_spider.py 商品列表商品链接爬取实现脚本

丨 丨—grab\_reiss\_products.py 根据t\_spider\_product\_url表中记录执行商品爬取

丨

丨—tedbaker

丨 丨—log

丨 丨 丨—tedbaker\_log.txt 记录未爬取、未找到sku、未找到图片的爬虫商品

丨 丨—tedbaker\_url\_spider.py 商品列表商品链接爬取实现脚本

丨 丨—grab\_tedbaker\_products.py 根据t\_spider\_product\_url表中记录执行商品爬取

丨

丨—batch\_alltocover.py 将运营处理好的封面图存为商品目录形式的脚本

丨

丨—asos 目录下脚本可用，但已老旧

丨—luisaviaroma 站点已被封IP，目录下脚本已弃用

\* 未在本目录说明的脚本文件为老旧、测试或弃用脚本文件

1. **爬虫库表说明**

|  |  |
| --- | --- |
| **t\_spider\_product\_url（商品链接记录表）** | |
| url\_id | 自增主键 |
| source\_code | 该站点对该商品的唯一标识 |
| flag | 来源标识 |
| url | 商品链接 |
| spider\_flag | 是否已爬取（0、未爬取，1、已爬取） |
| create\_time | 记录生成时间 |
| modify\_time | 记录更新时间 |

|  |  |
| --- | --- |
| **t\_spider\_product（爬虫商品基础信息记录表）** | |
| spider\_product\_id | 自增主键 |
| designer | 商品品牌、设计师 |
| gender | 性别 |
| category | 所属分类 |
| cover | 封面图链接 |
| status | 爬虫商品记录状态（1、正常，5、存档） |
| flag | 来源标识 |
| color\_code | 颜色码 |
| designer\_code | 该商品品牌对其唯一标识 |
| resource\_code | 该站点对商品唯一标识 |
| source\_url | 商品源链接 |
| editor\_name | 记录人 |
| date\_added | 记录新增时间 |
| date\_modified | 记录更新时间 |
| selling\_status | 在售状态（1、正常，2、下架） |

|  |  |
| --- | --- |
| **t\_spider\_product\_description（爬虫商品描述记录表）** | |
| spider\_product\_id | 爬虫商品ID |
| name | 商品名称 |
| description | 商品描述 |
| constitute | 成分与护理 |
| location | 产地 |
| category\_description | 分类描述 |
| language\_id | 该记录的语种（1、中文，2、英文） |
| editor\_name | 记录人 |
| date\_added | 记录新增时间 |
| date\_modified | 记录更新时间 |

|  |  |
| --- | --- |
| **t\_spider\_product\_iamges（爬虫商品图片记录表）** | |
| spider\_product\_id | 爬虫商品ID |
| images | 图片爬取完成后生成的本站图片链接（json数组字符串） |
| farfetch\_images | 商品图片来源站的图片地址集（以“,”隔开的字符串） |
| date\_added | 数据生成时间 |

|  |  |
| --- | --- |
| **t\_spider\_product\_sku（爬虫商品SKU记录表）** | |
| spider\_product\_sku\_id | 自增主键 |
| spider\_product\_id | 爬虫商品ID |
| size | 尺码 |
| price | 原价 |
| discount\_price | 折扣价 |
| currency | 币种 |
| editor\_name | 记录人 |
| date\_added | 记录新增时间 |
| date\_modified | 记录更新时间 |
| stock\_status | 在售状态（1、正常，0、售罄） |

|  |  |
| --- | --- |
| **t\_spider\_source（爬虫数据来源记录表）** | |
| spider\_source\_id | 自增主键 |
| flag | 来源标识 |
| name | 来源名称 |
| site\_name | 来源站点名称 |
| homepage | 来源站点链接 |
| allow\_list\_display | 0-不再商品列表显示 1-在商品列表显示 |

1. **数据爬取步骤说明**
2. 记录商品数据源

如果需要爬取的是一个新站点或需要独立记录的来源，需要先在t\_spider\_source表中新增一条记录。

1. 爬取需要爬的商品链接到t\_spider\_product\_url表

以farfetch为例，进入到爬虫脚本目录 spider>farfetch 编辑farfetch\_urls\_spider.py文件，

在 if \_\_name\_\_=="\_\_main\_\_": 方法下新增一行：

fus.grab\_product\_urls("015","https://www.farfetch.cn/it/sets/women/new-in-this-week-eu-women.aspx")

第一个参数为该来源的标识，第二个参数为该来源商品列表页链接，如果有多个商品列表页，则按此规则新增多行，新增完成后保存脚本并执行，在t\_spider\_product\_url表中将生成需要爬取的一个个商品链接。

1. 根据链接表中的记录爬取商品信息

商品列表链接爬取完成后，执行 grab\_farfetch\_products.py 脚本，此脚本将依次访问表中记录的待爬取的商品链接，然后解析返回结果，将解析出来的商品信息存库。

1. 爬取商品图片

待商品爬取完成后，进入 spider>continued 目录，执行comm\_use\_img\_download.py 脚本，下载爬取到的商品的图片，图片根目录名请使用 ptncp，文件路径自行修改脚本配置，脚本还会二外存一套商品封面图片，用于运营修改，存储目录可根据自己需要进行修改。

1. 翻译描述信息

部分站点只有英文站，需要将描述表的英文存一份中文版，可进入spider>continued目录，执行translate\_desc.py脚本，此脚本将会调用Google翻译，将英文版商品描述信息翻译成中文并存入库中，当前需要翻译的站点有reiss、tedbaker、jackwills、frenchconnection。

1. 处理封面图

在t\_spider\_source表中来源记录的allow\_list\_display字段为1的爬虫商品，需要对封面做标准化处理，将图片爬取时生成的封面图目录发给运营的相关负责人进行处理，待其处理完成发回后，使用spider下的batch\_alltocover.py（请更具自己的目录路径对脚本中的目录配置进行修改）脚本将封面图片列表复制成安商品目录保存的形式，然后上传到图片服务器。

1. 上传数据及图片到线上

待所有步骤完成后，通知运维负责人将数据及图片同步到线上。

注意：

当前脚本还有待完善，部分脚本执行时偶尔有报错出现。

遇到过的问题有：

1. 网络不稳定，访问超时，遇到此问题需检查源网站是否还能正常访问；
2. 网络访问卡死，脚本未报错也不继续执行，使用的python urllib库本身可能存在问题，只能强行中止脚本运行后重新运行脚本；
3. 对源站点数据结构判断不全面，导致数据不对或解析报错，请根据报错情况定位到具体解析错误的数据，重新编写解析代码；
4. 源站点数据结构变化，导致解析报错，如数组下标越界，空值方法调用等，请根据报错情况定位到具体解析错误的数据，重新编写解析代码。

一点心得：

编写爬虫的核心工作在于分析源网站的页面数据，定位关键数据的位置特性，大部分站点的关键数据是在页面的静态js代码中，一般商品详情页都有该站点对该商品的唯一标识，可根据该标识查找封装商品信息的代码部分，部分站点的关键信息是通过ajax请求加载，可以拦截并分析请求，由请求名及返回数据分析是否是有效数据，编写数据解析代码时，一定要多分析几个页面，尤其是类型不同的页面，如男士的和女士的，有折扣的和没折扣的，英文的和中文的，在架的和下架的等，找出这些页面关键数据的通性。

Python版本2.7

用到的第三方python 库：

urllib urllib2 cookielib multiprocessing pymysql BeautifulSoup execjs