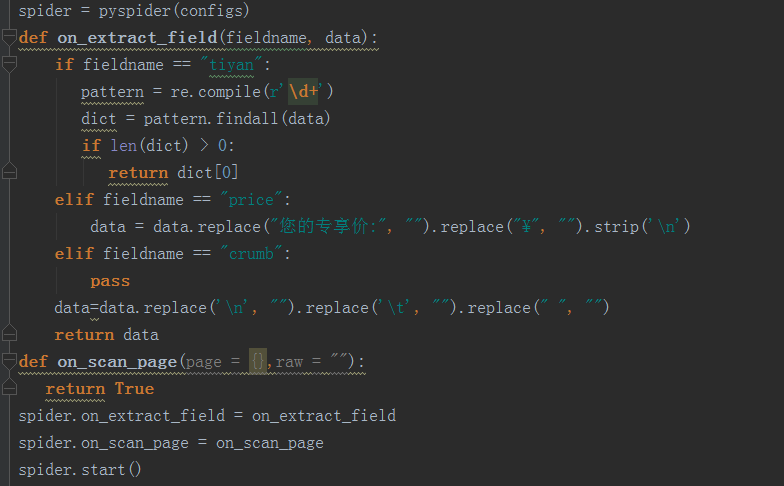
**第一个demo**

爬虫采用Python编写，下面以也买酒网站为例，来看一下爬虫长什么样子。





爬虫的整体框架就是这样，首先定义了一个configs字典，里面设置了待爬网站的一些信息，然后通过调用spider = pyspider(configs)和spider.start()；来配置并启动爬虫

# configs详解------之成员

Configs字典可以定义下面这些元素

**name**

定义当前爬虫名称

**String类型，可选设置**

举个例子：

“name”:”也买酒”

**log\_show**

是否显示日志

为true时显示调试信息

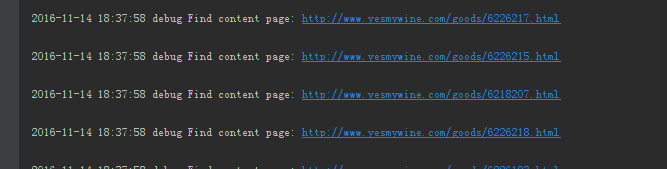
为false时显示爬取面板

**布尔类型 可选设置**

**log\_show默认值为false，即显示爬取面板**

例子1：

“log\_show”:True

****

**tasknum:**

同时工作的爬虫任务数

需要配合redis保存采集任务数据，供进程间共享使用

整形 可选设置

**tasknum默认为1,即单进程任务爬取**

举个例子：

开启5个进程爬取网页：

“tasknum”:5

proxy

代理服务器

如果爬取的网站根据IP做了反爬虫，可以设置此项

**字典类型 可选设置**

例子1：

普通代理

‘proxy’:{“http”:”http://host:port”}

例子2：

验证代理：

"proxy": {

"http": "http://H29R2R5TD0H6A7LD:F56A84DCCF2A82CF@proxy.abuyun.com:9010", "https": <http://H29R2R5TD0H6A7LD:F56A84DCCF2A82CF@proxy.abuyun.com:9010>

}

**注意：如果对方根据IP做了反爬虫技术，你可能需要到**[阿布云代理](http://www.abuyun.com/)**申请代理通道 或者 第三方免费代理IP，然后在这里填写代理信息**

**Interval**

**爬虫爬取每个网页的时间间隔**

**单位：秒**

**整形 可选设置**

**举个例子：**

**设置爬取时间间隔为5秒**

**“interval”:5**

**timeout**

爬虫爬取每个页面的超时时间

单位：秒

**整形 可选设置**

timeout默认值为10秒

举个例子：

“timeout”:10

**max\_try**

爬虫爬取每个网页失败后尝试次数

网络不好可能导致爬虫在超时时间内抓取失败，可以设置此项允许爬虫重复爬取

**整形 可选设置**

max\_try默认值为0

举个例子：

“max\_try”:5 #重复爬取5次

**user\_agents**

爬虫爬取网页所使用的随机浏览器类型

爬虫爬取网页时随机使用其中一种浏览器类型，用于破解防采集

**列表类型 可选设置**

举个例子：

"user\_agents":["Mozilla/5.0 (Macintosh; Intel Mac OS X 10\_11\_6) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/52.0.2743.116 Safari/537.36"," Mozilla/5.0 (iPhone; CPU iPhone OS 9\_3\_3 like Mac OS X) AppleWebKit/601.1.46 (KHTML, like Gecko) Version/9.0 Mobile/13G34 Safari/601.1"],

**export**

爬虫爬取数据导出

type:导出类型txt、sql、db

file:导出txt、sql文件地址

db:导出db、sql数据表名

# configs详解------之field

field定义一个抽取项，一个field可以定义下面这些东西

**name**

给此项数据设置个变量名

变量名中不能包含.

**String类型 不能为空**

举个例子：

给field起个名字叫title

{“name”:”title”,”selector”:”div.content”}

**selector**

定义抽取规则，目前只支持beautifulsoup，解析器使用html.parser

**String类型 不能为空**

举个例子：

{“selector”:”div.content”}

**required**

定义该field值是否必须，默认false

赋值为true的话，如果该field没有抽取到内容，该field对应的整条数据都将被抛弃

**布尔类型**

举个例子：

{"name": "title", "selector": "div.content", "required": True}

# configs详解------之requests

### requests方法

**requests.set\_proxies**

设置需要添加的代理，用于破解防采集

**requests.set\_referer**

一般在on\_start回调函数中调用，设置请求来路URL

@param $referer 需添加的来路URL，用于破解防采集

举个例子：

def on\_start(r):

r.add\_header("Referer", "http://list.yesmywine.com/z3/")

return r