**Crawler API Manual**

**0.1**

|  |  |  |  |
| --- | --- | --- | --- |
| 拟制: | 刘龙均、吴勇 | 日期： | 2016/5/27 |
| 审核: |  | 日期： |  |
| 批准: |  | 日期： |  |

**文档修订记录**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **修订版本** | 日期 | **作者** | 修改内容描述 | **备注** |
| 0.1 | 2016/5/27 | 刘龙均 | API手册初始化 |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

目录

[CrawlerBase 1](#_Toc452491440)

[\_\_init\_\_ 1](#_Toc452491441)

[visitHomePage 1](#_Toc452491442)

[visitValidateCode 2](#_Toc452491443)

[visitSearchList 2](#_Toc452491444)

[visitTopInfo 2](#_Toc452491445)

[visitJbxx 3](#_Toc452491446)

[visitGdxx 3](#_Toc452491447)

[ModuleManager 4](#_Toc452491448)

[CrawlerControl 4](#_Toc452491449)

# CrawlerBase

## \_\_init\_\_

**参数**：self, pinyin, config\_dict, check\_dict, callback

pinyin：省份拼音

config\_dict：模块配置字典

check\_dict：模块检查字典

callback：外部回调函数

**使用场景**：通常是在子类的构造函数中显示调用

**示例**：

|  |
| --- |
| class Crawlerxxx(CrawlerBase): config\_dict = dict()  config\_dict[CrawlerRunMode.COMPANY\_KEY] = [self.initHomePase, self.initValidateCode, self.initSereachList]  config\_dict[CrawlerRunMode.COMPANY\_URL] = []  check\_dict = dict()  check\_dict['html\_check\_dict'] = {u'过于频繁':WebAccessType.TOO\_OFTEN, u'非正常访问':WebAccessType.ACCESS\_VIOLATION }  CrawlerBase.\_\_init\_\_(self, pinyin, config\_dict, check\_dict, callbackFromOuterControl)  … |

## visitHomePage

**参数**：module  
module：Module对象

**使用场景**：访问指定省份的搜索首页时可调用该函数

**示例**：TODO

## visitValidateCode

**参数**：module  
module：Module对象

**输出值**：yzm，img\_path

**使用场景**：各省份进行验证码验证时调用该函数

**示例**：TODO

## visitSearchList

**参数**：module  
module：Module对象

**使用场景**：访问公司搜索列表

**示例**：TODO

## visitTopInfo

**参数**：module  
module：Module对象

**输出值**：html

**使用场景**：访问页面TOP信息

**示例**：TODO

## visitJbxx

**参数**：module  
module：Module对象

**输出值**：company

**使用场景**：访问基本信息页面

**示例**：TODO

## visitGdxx

**参数**：module  
module：Module对象

**输出值**：gdxx

**使用场景**：访问基本信息页面

**示例**：TODO

# ModuleManager

# CrawlerControl