护城项目

爬虫框架适配规范

版本号：V0.8

|  |  |  |  |
| --- | --- | --- | --- |
| 拟制: | 丁永强 | 日期： | 2016.6.1 |
| 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** |
| V0.8 | 2016.1.1 | 丁永强 |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

目录

[1 引言 4](#_Toc452568894)

[1.1 目的 4](#_Toc452568895)

[1.2 预期读者 4](#_Toc452568896)

[2 约定 4](#_Toc452568897)

[2.1 中间结果约定 4](#_Toc452568898)

[2.2 事件约定 4](#_Toc452568899)

[2.2.1 事件判定优先级次序 4](#_Toc452568900)

[3 子类约束 5](#_Toc452568901)

[3.1.1 setRowkey 5](#_Toc452568902)

[3.1.2 search\_company 5](#_Toc452568903)

[3.1.3 resultCollect 5](#_Toc452568904)

[4 爬虫验证 5](#_Toc452568905)

[4.1.1 翻页 5](#_Toc452568906)

[4.1.2 无此公司 5](#_Toc452568907)

[4.1.3 搜索列表中无有效链接 5](#_Toc452568908)

[4.1.4 url抓取 5](#_Toc452568909)

# 引言

## 目的

通用爬虫框架适配规范为开发人员提供开发的依据，以确保爬虫适配的高效和正确性。

## 预期读者

护城项目爬虫开发人员、产品经理、项目管理人员。

# 约定

## 中间结果约定

1. 验证码输出 yzm

2.页面输出页面内容及web对象 html,web

3.json输出页面内容及web对象 json,web

4.股东信息表头表值键值对 gdxx\_list

5.待搜索公司关键字 company\_key

6. 列表中的每一个公司名 search\_company

7.公司信息中间抓取结果以键值company存放在result\_dict和value\_dict

8.待搜索公司的company\_url

## 事件约定

### 事件判定优先级次序

优先级由高至低：

1.异常事件

2.web页面内容访问异常，如被反爬

3.断言失败事件

4.未到期望输出事件

# 子类约束

### setRowkey(self, map\_config=None)

self.page\_dict['rowkey\_dict'] = {'company\_name':u'name', 'company\_zch':u'999003'}

为外部提供生成rowkey的必备属性，父类提供默认实现，但子类需考虑复写

注意:不属于模块方法,需要注意try except

### setNonCompanyConfig(self, non\_company\_set)

初始化时调用，传递用来判断无此公司的关键字set集合

### search\_company

module.appendOutput(name=’search\_company’)

搜索列表中每一家公司的公司名,子类在解析搜索列表后需要生成，作为output放置到中间结果集

### resultCollect

回收搜索列表中的每一家公司的抓取结果，不可或缺，需要作为配置模块中的最后一个

### nb\_name

每个年度年报所属年度，例如2013、2014，不可或缺，页面内容保存时作为key的一部分使用

# 爬虫验证

### 翻页

需要验证股东、变更、备案、分支机构翻页情况是否正确, 可以从solr中搜索到恰当的用例

### 无此公司

需要验证

### 搜索列表中无有效链接

某些省份，如江西会出现搜索出来的一条或多条记录有名称无url的情况, 可以从solr中查询用例

### url抓取

通过url进行抓取，某些省份可能不支持，例如北京

### rowkey验证

需要验证外部生成rowkey所依赖的属性是否已经被成功设置