爬虫文档说明

1. 爬虫文档结构

黑龙江、湖北、山西、青海统一使用黑龙江爬虫文件

湖南、福建统一使用总局爬虫文件

辽宁使用自身爬虫文件

每个省份爬取步骤一致，解析方式不同，将在下文做详细描述

单个爬虫文件有页面数据获取、数据表分析、单元测试三个类。

页面数据获取（class HeilongjiangClawer）：包括得到验证码图片；调用模型，获得破解结果,post破解结果；判断验证码破解是否成功；发送查询请求，提取企业ID号；获取目标页面数据等功能

数据表分析（class HeilongjiangParser）：包括分析页面数据；存储数据等功能

单元测试（class TestParser）：编写单元测试。

1. 开发环境

环境变量配置

import importlib

ENT\_CRAWLER\_SETTINGS = os.getenv('ENT\_CRAWLER\_SETTINGS')

if ENT\_CRAWLER\_SETTINGS:

settings = importlib.import\_module(ENT\_CRAWLER\_SETTINGS)

else:

import settings

爬虫基类

from crawler import Crawler

from crawler import Parser

from crawler import CrawlerUtils

1. 爬取流程

（图图图）

1. 功能及步骤描述
2. 得到验证码图片**功能描述**

访问验证码获取网址

存储验证码图片

存储路径：在settings中配置

1. 判断验证码破解是否成功**功能描述**

查找返回值或返回页面中关键字，如0，1或true，false

根据关键字判断验证码是否破解成功

1. 发送查询请求，提取企业ID号**功能描述**

将正确的验证码破解结果、企业注册号构造为请求参数

post请求参数，获得查询出的结果页面

从返回页面内容，企业链接中提取企业ID号

1. 获取目标页面数据**功能描述**

固定文件路径结合企业ID号获得目标页面地址

访问目标网址，获得页面数据

调用页面解析函数

1. 分析页面数据**功能描述**

查找页面所有表格

对表格分类处理

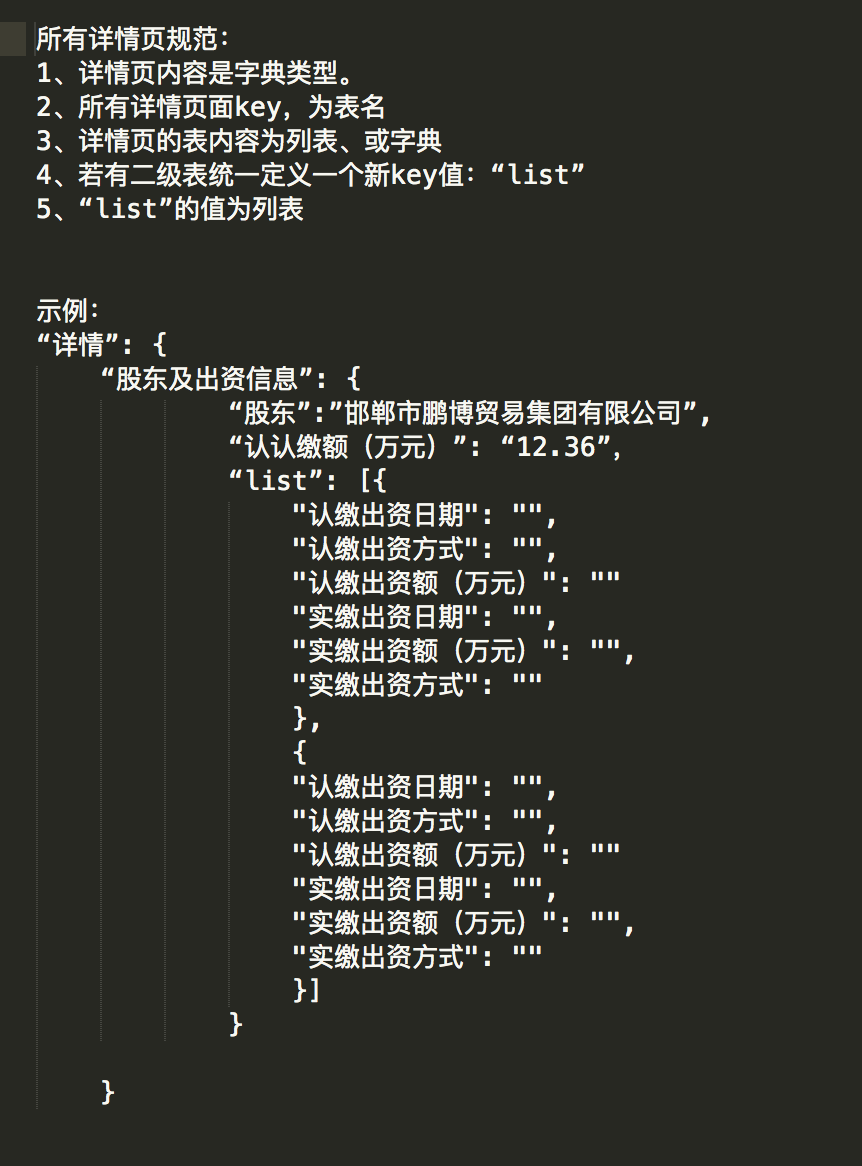
1. 存储数据**功能描述**

**每一家企业存储为一个字典对象，key值：企业注册号value值：页面爬取内容，格式如下图3.1**



**图3.1 企业数据存储**

**关于详情页面存储规范如下图3.2**



**图3.2 详情页存储规范**

最后将所有企业爬取结果转换为json格式存储

1. 单元测试**功能描述**
2. 部分函数描述及参数说明
3. parse\_ind\_comm\_pub\_pages(self, page)

解析工商公示信息-页面。

Page：工商公示信息页面html

在黑龙江和总局页面中，工商公示信息所有信息都通过一个http get返回；在辽宁页面中，工商公示信息中的每一个表格通过http get返回。

1. parse\_ent\_pub\_pages（self, page）

解析企业公示信息-页面。

Page：企业公示信息页面html

在黑龙江和总局页面中，企业公示信息所有信息都通过一个http get返回；在辽宁页面中，企业公示信息中的每一个表格通过http get返回。

1. crawl\_other\_dept\_pub\_pages（self, page）

其他部门公示信息-页面。

Page：其他部门公示信息页面html

在黑龙江和总局页面中，其他部门公示信息所有信息都通过一个http get返回；在辽宁页面中，其他部门公示信息中的每一个表格通过http get返回。

1. parse\_judical\_assist\_pub\_pages（self, page）

司法协助信息-页面。

Page：司法协助信息页面html

在黑龙江和总局页面中，司法协助信息所有信息都通过一个http get返回；在辽宁页面中，司法协助信息中的每一个表格通过http get返回。

1. parse\_table1（self，table）

描述：一类解析表，一个th对应一个td

参数：able:需要解析的表

返回值：table\_save:将th与td对应，得到解析好的内容，返回字典

思路：获取所有不跨列的th，获得所有td，将th与td一一对应

1. parse\_table2(table, table\_name)

描述：二类表格解析，就诶洗**表列名、内容、页数为三个同级的表格**

参数： table: 内容表

table\_name： 表的名字

返回值：

[table\_content, content\_tr]:有详情页时返回解析的表table\_content列表和每一行的内容content\_tr列表

[table\_ts, "None"]:若没有详情，返回解析的表table\_content列表

["", "None"]:若出现页面读取错误返回""和"None"

思路:1.读取内容表同级上一张表，获得列名

2.读取内容表同级下一张表，获得页数

2.1若没有页数，直接通过get\_td解析表格，返回需要保存的表

2.2若只有一页，直接通过get\_td解析表格，返回需要保存的表

2.3若有多页，依次发送请求，通过get\_td解析表格，返回需要保存的表

**黑龙江：分析页面数据详细说明**

1. **通过企业id号获取工商公示信息页面，得到html页面数据**
   1. **通过BeautifulSoup代码库转为可解析的树型结构**
   2. **对表格进行分类，分四类处理**
      1. **一类：解析一个th对应一个td的表格**
      2. **二类：解析表列名、内容、页数为三个同级的表格**
      3. **三类：解析列名、内容、页数在同一张表格**
      4. **四类：解析有二级表头的表格（coarse\_page\_table）**
   3. **“基本信息”、“清算信息”以一类处理**
   4. **“变更信息”、“分支机构信息”、“股权出质登记信息”、“行政处罚信息”、“经营异常信息”、“严重违法信息”、“抽查检查信息”、“股东信息”和“动产抵押登记信息”、“主要人员信息”以上10 张表以二类处理**
   5. **其中“股东信息”和“动产抵押登记信息”信息有详情页，分别通过get\_shareholder\_detail和get\_movable\_property\_reg\_detail函数分别处理**
   6. **其中“主要人员信息”由于别列名重复，在二类表中做特殊处理**
2. **通过企业id号获取企业公示信息页面，得到html页面数据**
   1. **通过BeautifulSoup代码库转为可解析的树型结构**
   2. **对表格进行分类，分四类处理**
   3. **“股权变更信息”、“变更信息”、“行政许可信息”、“知识产权出质登记信息”、“行政处罚信息”以三类处理**
   4. **“股东及出资信息”有二级表头以四类处理**
   5. **“企业年报”有详情页**
      1. **通过链接提取目标网页id**
      2. **获取详情页内容，转为可解析的树型结构**
      3. **详情也中的“企业资产状况信息”、“企业基本信息”、“基本信息”表信息以一类处理**
      4. **“网站或网店信息”、 “股东及出资信息”、 “行政许可情况”、 “对外投资信息”、“对外提供保证担保信息”、 “股权变更信息”、 “修改记录”以上7张表以三类处理**
      5. **为父表“企业年报”添加详情字段，存粗详情页内容**
3. **通过企业id号获取其他部门公示信息，得到html页面数据**
   1. **通过BeautifulSoup代码库转为可解析的树型结构**
   2. **“行政许可信息”、“行政处罚信息”以一类表处理**
4. **通过企业id号获取司法协助信息页面，得到html页面数据**
   1. **通过BeautifulSoup代码库转为可解析的树型结构**
   2. **“股东变更信息”、“股权冻结信息”以一类表处理**

**辽宁：分析页面数据详细说明**

1. **通过成功破解的返回页**

**获得enttype、pripid、entname、optstate、regno参数**

1. **使用参数构造请求url**
2. **每一张表都是一个页面，对每张表格都需要发起一个请求**
3. **对表格进行分类，分三类处理**
   1. **一类：解析一个th对应一个td的表格**
   2. **二类：解析列名、内容、页数在同一张表格**
4. **“股东信息”**
   1. **读取网页**
   2. **使用正则表达式截取tzr\_paging变量并做json格式转换**
   3. **字段与表格列名对应**
   4. **使用参数pripid，和enttype得到详情页数据内容**
   5. **将字段与表格列名对应**
5. **“主要人员”**
   1. **读取网页表格内容**
   2. **正则表达式截取zyry\_nz\_paging变量并做json格式转换**
   3. **字段与表格列名对应**
6. **“清算信息”**
   1. **读取网页表格内容**
   2. **使用二类解析方式，字段与表格列名对应**
7. **“变更信息”**
   1. **读取网页表格内容**
   2. **正则表达式截取paging变量并做json格式转换**
   3. **字段与表格列名对应**
8. **“股权出质登记信息”**
   1. **读取网页表格内容**
   2. **正则表达式截取gqczdjPaging变量并做json格式转换**
   3. **字段与表格列名对应**