

工商解析系统

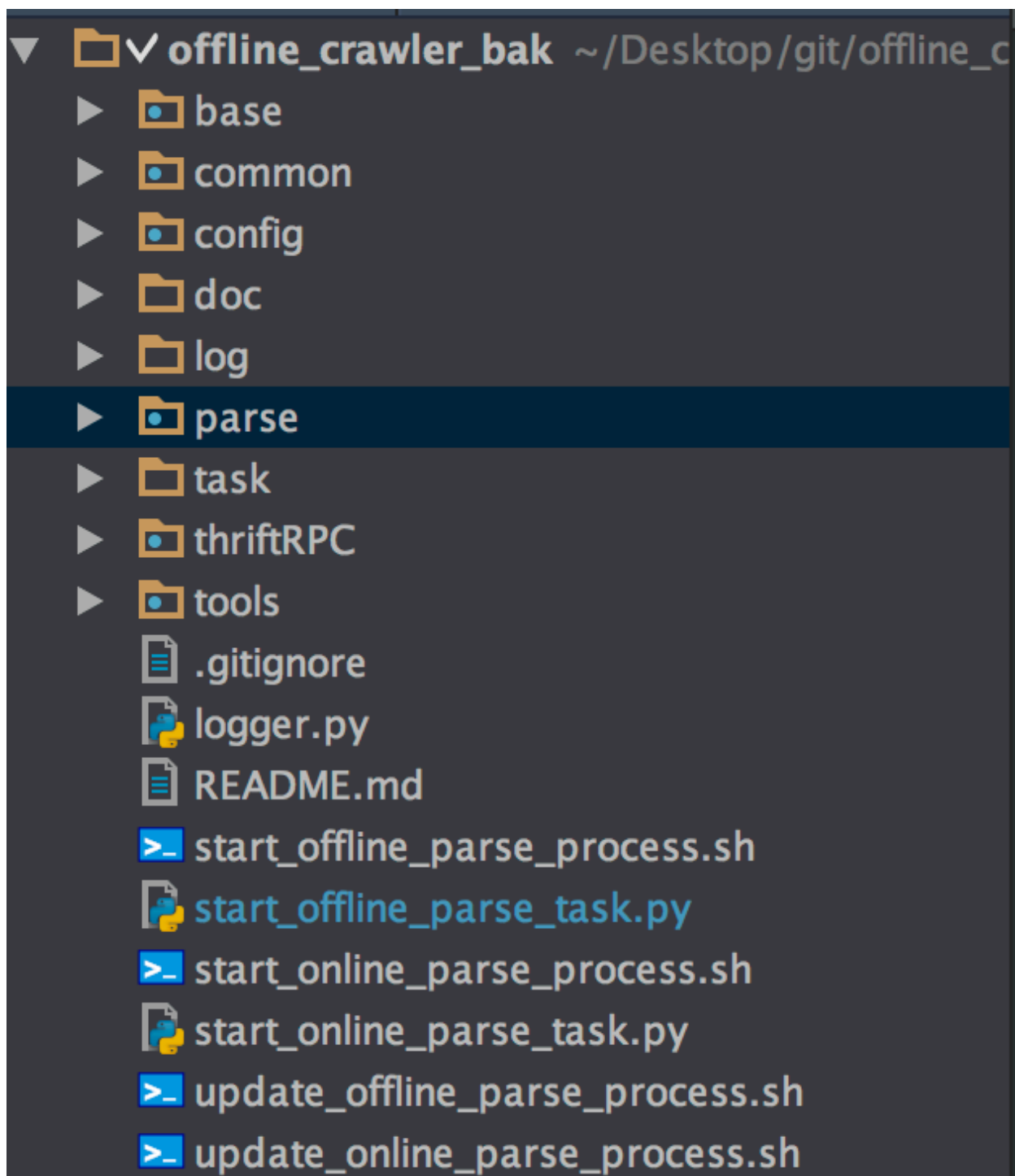
1.功能说明

对工商提取网页库进行解析

网页库位于mongodb crawl_data库 表格命名方式为: online_crawl省份new

案例:江苏省 --- online_crawl_jiangsu_new

2.程序文件目录说明



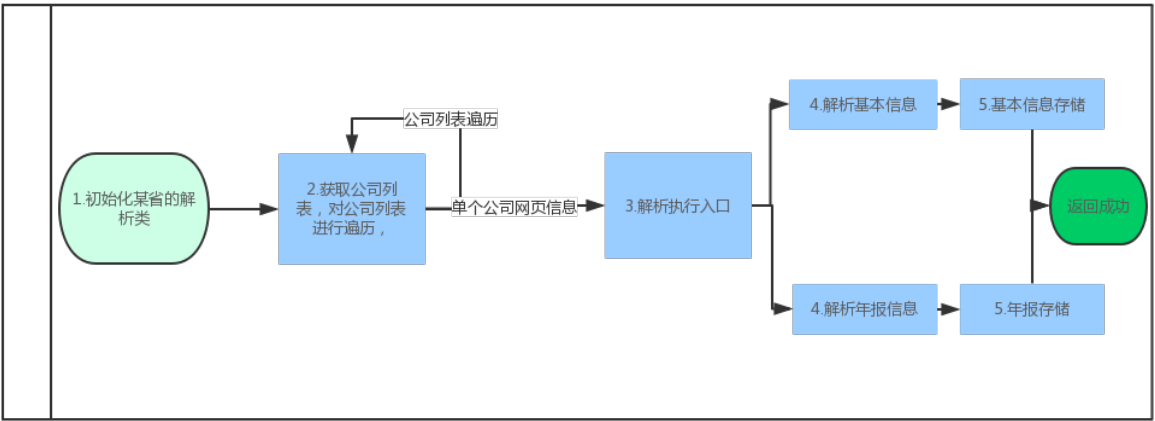
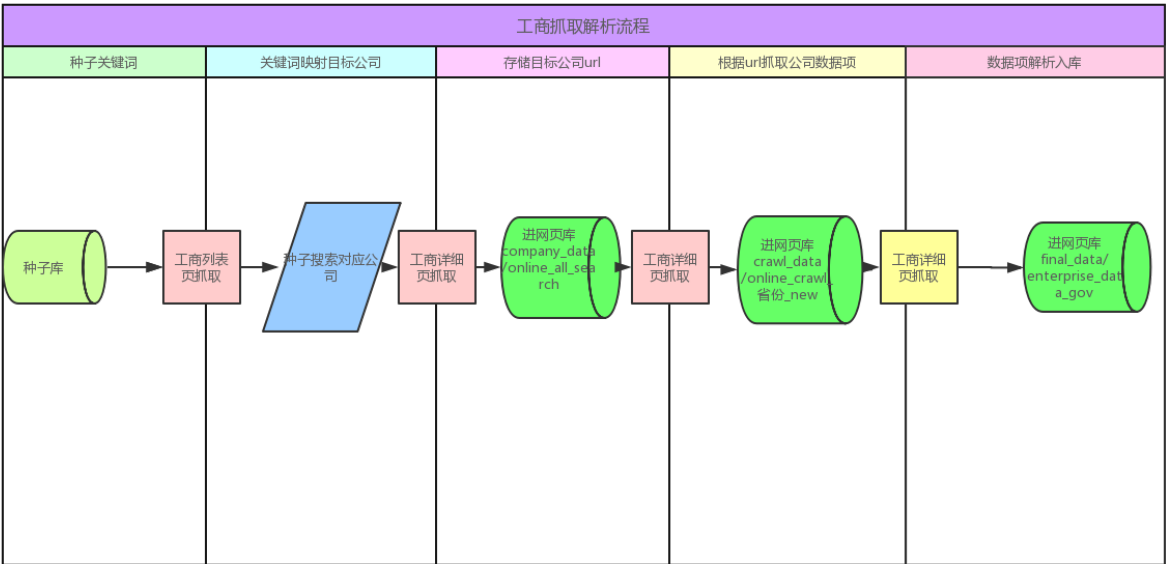
- /base —— 存放基类
- /common —— 程序工具共同类
- /config —— 配置文件
- /doc —— 说明文档
- /log —— 日志文档
- /parse —— 解析业务代码类
- /tools —— 工具脚本
- logger.py —— 日志类
- start_offline_parse_task.py —— 程序入口

案例1:江苏省具体业务代码:/parse/parse_jiangsu_worker.py

案例2:数据库密码配置:/config/conf.py

3.程序执行流程

工商抓取解析系统流程



使用说明ctrl +f 搜索代码

3-1.初始化解析类:start_offline_parse_task.py

```
self.worker_list[key] = create_crawl_object(value, key)
```

案例:解析江苏省 ,此时:

key为:'jiangsu' value{'class':'GsxtParseGuiZhouWorker','host':'gz.gsxt.gov.cn'.....}

3-2.获取公司列表, 并对其遍历 start_offline_parse_task.py

```
# 公司列表
#分两个数据库装数据, 老库(source_db)已满, 所以添加新库(source_db_new)
company_list_old=self.__get_iterator_company_list(self.source_db)
company_list_new=self.__get_iterator_company_list(self.source_db_new)

count = 0
# 网页库遍历
for item in company_list_old:
    count += 1
    self.worker_list[self.province].query_offline_task(item, CHOOSE_DB_OLD)
    #....

# 新版网页库 遍历
for item in company_list_new:
    count += 1
    self.worker_list[self.province].query_offline_task(item, CHOOSE_DB_NEW)
    #....
```

3-3.解析执行入口:parse_base_worker.py

query_offline_task() 内会 调用 query_company()

```
def query_company(self, item):
```

3-4.1解析基本信息:parse_base_worker.py

```
gs_flag = self.parse_gs_info(company, data_list, u_time,
                              _in_time, province, in_time,
                              base_info_url, item)
```

3-4.2解析年报基本信息:parse_base_worker.py

```
nb_flag = self.parse_nb_info(company, data_list, u_time,
                              _in_time, province, in_time,
                              base_info_url)
```

3-5.1存储基本信息:parse_base_worker.py

```
# 位于 parse_gs_info()中
# 存储解析信息
return self.__store_model(company, base_info_url, in_time, model)
```

3-5.2存储年报信息:parse_base_worker.py

```
#位于 parse_nb_info()中
# 年报需要单独存储
if not self.__store_annual_model(company, year, base_info_url, in_time,
nb_model):
```

4.程序运行案例

各省解析类(parse_jiangsu_worker.py) —— 继承 —— parse_base_worker.py —— 继承—— task_base_worker.py

各省网页结构不同，解析代码不同。各省的解析代码 均在/parse下各个解析器下 如:江苏省的 parse_jiangsu_worker.py

案例: 省份:江苏省 公司名:伊诗夏兰薇时装（南京）有限公司

4.1.入口参数修改

文件start_offline_parse_task.py

```
def main():
    config = 'config/offline_gsxt_parse.conf'
    province = 'jiangsu'
    max_count = 100000
```

province : 省份参数

config : 配置文件参数

config= 'config/offline_gsxt_parse.conf' # 以config/offline_gsxt_parse.conf为配置文件

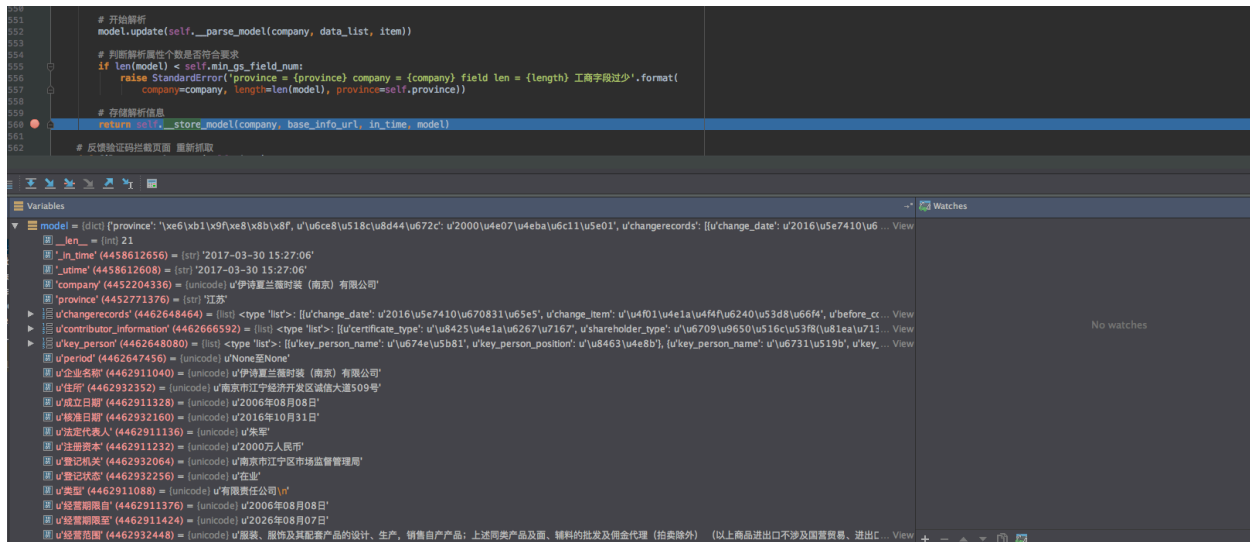
4.2.配置文件修改

```
#14
[jiangsu]
province=jiangsu
logfile=offline_parse_jiangsu
host=www.jsqsj.gov.cn:58888
clazz=GsxtparseJiangSuWorker
source_table=online_crawl_jiangsu_new
target_table=enterprise_data_gov
annual_table=annual_reports
crawl_flag=crawl_online
success_flag=50
source_select_param=('_id': '伊诗夏兰薇时装（南京）有限公司', 'crawl_online': {'$ne': 100})
is_nb_db_open=False
is_nb_mq_open=True
is_gs_db_open=False
is_gs_mq_open=True
gs_topic=49
gs_nb_topic=136
```

文件config/offline_gsxt_parse.conf

前往[jiangsu]的代码段

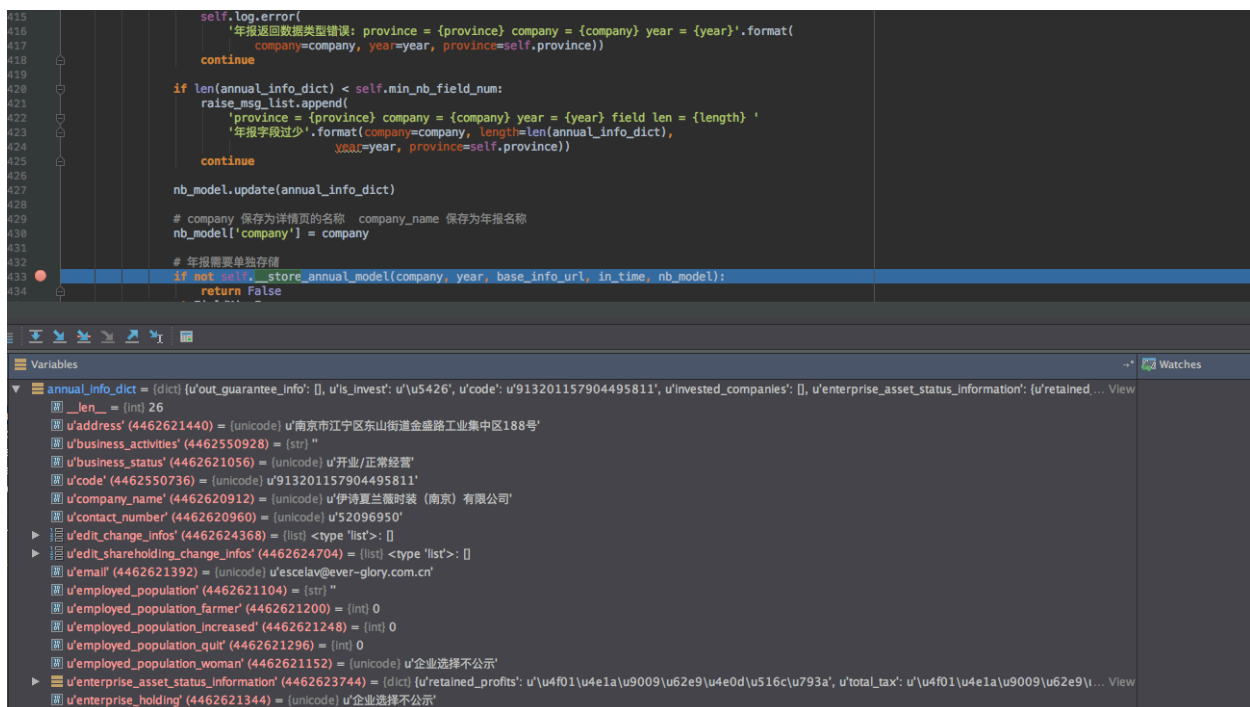
```
return self. store model(company, base info url, in time, model)打断点
```



4.5查看解析年报信息入库前一刻

断点位置在文件parse_base_worker.py

```
if not self.__store_annual_model(company, year, base_info_url, in_time, nb_model):
```



4.6数据库

前往mongodb,final_data库/enterprise_data_gov 查看数据

Copy of mongoDB 172.16.215.2:40042 final_data

db.getCollection('enterprise_data_gov').find({'company':'伊诗夏兰薇时装（南京）有限公司'})

enterprise_data_gov 0.816 sec. 0 50

Key	Value	Type
(1) ObjectId("583437befae0c...	{ 27 fields }	Object
_id	ObjectId("583437befae0c8cbb581f3...	ObjectId
changerecords	[4 elements]	Array
period	2006-08-08至2026-08-07	String
_src	[1 element]	Array
registered_date	2006-08-08 00:00:00	String
_record_id	0d219c8f6dae0438d620170b27a0b...	String
unified_social_credit_code	913201157904495811	String
enterprise_type	有限责任公司	String
registered_capital	20000000.00	String
_in_time	2016-11-22 20:19:08	String
company	伊诗夏兰薇时装（南京）有限公司	String
_utime	2017-02-23 13:04:37	String
shareholder_information	[2 elements]	Array
province	江苏	String
registered_code	320100400035118	String
hezhun_date	2016-10-31 00:00:00	String
business_scope	服装、服饰及其配套产品的设计、生产...	String
src_registered_capital	2000万元人民币	String
address	南京市江宁经济开发区诚信大道509号	String
key_person	[5 elements]	Array
registered_address	南京市江宁区市场监督管理局	String
business_status	在业	String

如上图,数据已入库.

字段说明

mongolddb,crawl_data库/online_crawl省份new 字段说明

字段	对照官网字段
base_info	基本信息
branch_info	分支机构
change_info	变更信息
shareholder_info	股东及出资信息
contributive_info	股东及出资信息
annual_info	企业年报信息
_id	公司名字
province	省份
crawl_online	提取状态

