爬虫框架用户文档

笔记本：	spider		
创建时间：	2018/8/11 10:00	**更新时间：**	2018/8/11 16:51
作者：	zhuhanqing258@foxmail.com		
URL：	about:blank		

框架简介—>总体架构—>架构详解—>API

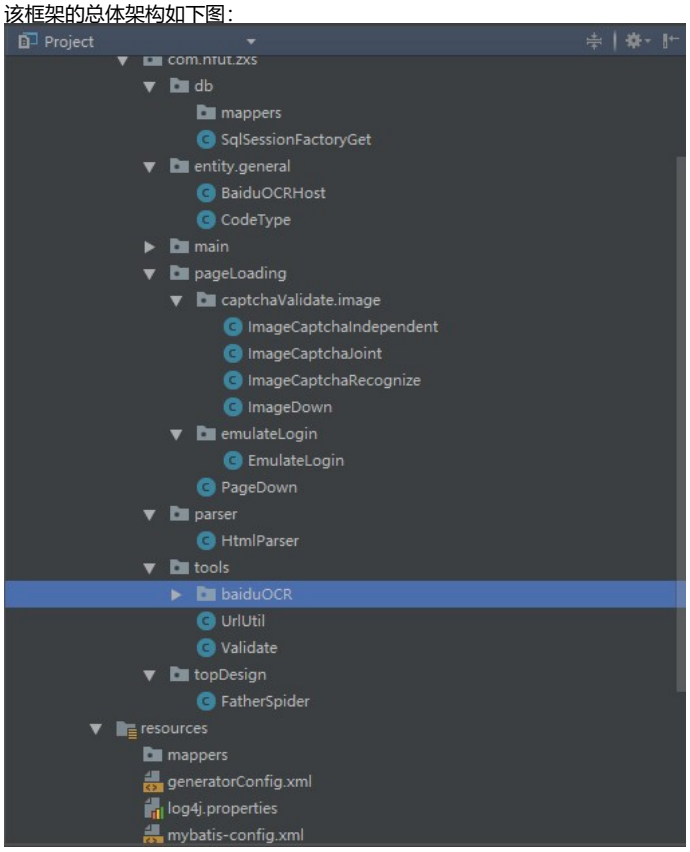
一、框架简介

该框架主要用于快速进行针对某一网站的爬虫架构，从而满足用户相应的信息需求。
该框架依托 maven 进行项目管理，以git作为版本控制工具，数据库及其操作采用mysql+mybatis 的方式，并以log4j作为日志输出工具。
该框架主要使用Jsoup框架对html页面进行解析，使用fastjson对json数据进行解析，使用httpclient工具进行网络数据获取。
参考文档如下：

- jsoup:
  - 中文文档：http://www.open-open.com/jsoup/
  - 英文文档：https://jsoup.org/
- fastjson：https://www.w3cschool.cn/fastjson/fastjson-api.html
- httpclient:
  - http://hc.apache.org/httpcomponents-client-ga/tutorial/html/index.html
  - http://hc.apache.org/httpclient-3.x/userguide.html
- mybatis:
  - 英文版：http://www.mybatis.org/mybatis-3/index.html
  - 中文版：http://www.mybatis.org/mybatis-3/zh/index.html
- git:https://www.liaoxuefeng.com/wiki/001373951630592960 6dd18361248578c67b8067c8c017b000
- log4j:https://blog.csdn.net/xiaoxiong_web/article/details/77932655

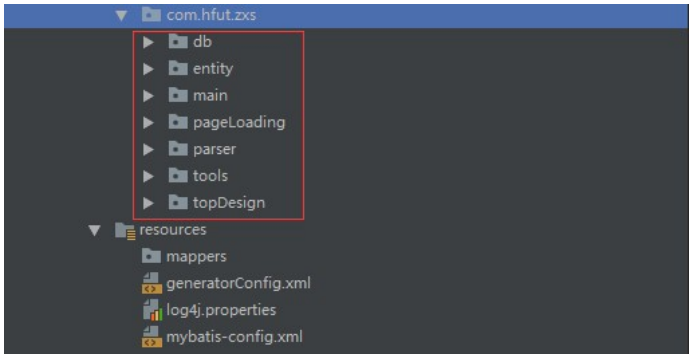
二、总体架构

该框架的总体架构如下图：



三、架构详解

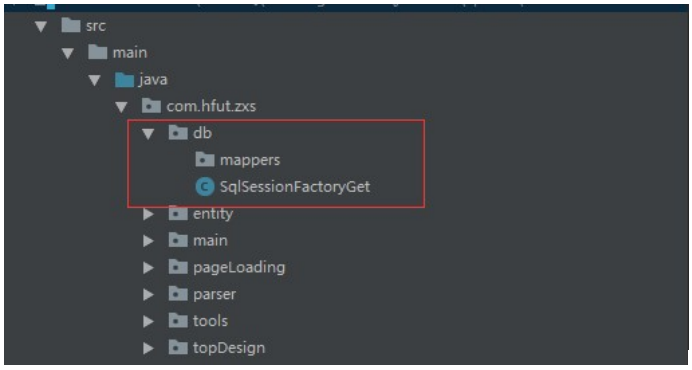
（一）该框架的顶层包主要包括以下几个部分，如图：



下面将从上到下依次讲解各包的组成及其作用：

1. **package: db**

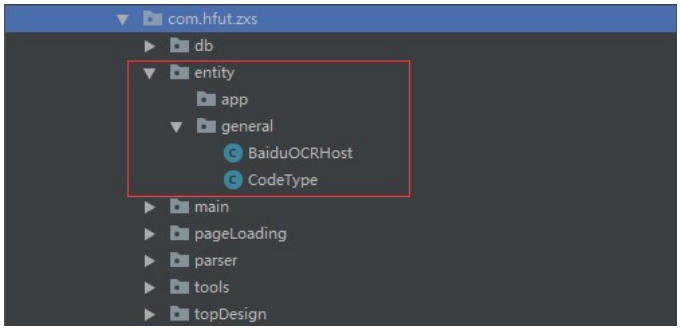
- db包主要包含以下两部分，如图：



- 其中：
  - **package:mappers** ： 用来存放对相应数据库实体类的操作接口，这些接口可自己手动创建，也可以使用 **mybatis** 的逆向工程进行自动生成。maven项目中配置mybatis
  - **class:SqlsessionFactoryGet:** 用来快捷获取使用mybatis所必须的 **SqlSessionFactory** 以及负责 **SqlSession** 的commit和close。
- mybatis 文档如下：
  - 英文版：http://www.mybatis.org/mybatis-3/index.html
  - 中文版：http://www.mybatis.org/mybatis-3/zh/index.html

2. **package: entity**

- entity包主要包含以下两部分，如图：



- 其中：
  - **package:app** ： 用来存放不同应用需要的实体类。
  - **package:general** ： 用来存放公用的特定用途的类。如 **class: BaiduOCRHost** 存储了百度OCR工具的一些借口，方便调用，如图：

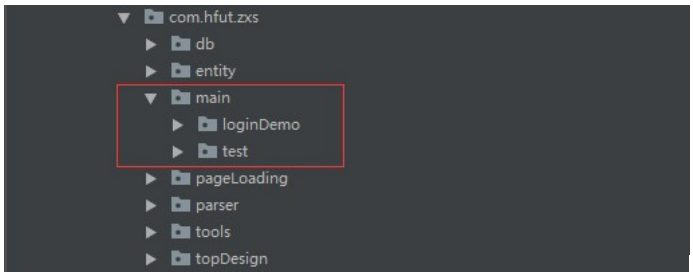
BaiduOCRHost.java

```
1   /.../
8
9   package com.hfut.zxs.entity.general;
10
11  /**...*/
20  public class BaiduOCRHost {
21
22      //通用文字识别
23      public static final String GENERAL_BASIC="https://aip.baidubce.com/rest/2.0/ocr/v1/general_basic";
24      //通用文字识别(高精度版)
25      public static final String ACCURATE_BASIC="https://aip.baidubce.com/rest/2.0/ocr/v1/accurate_basic";
26
27
28  }
```

- 百度OCR工具文档：https://ai.baidu.com/docs#/OCR-API/top

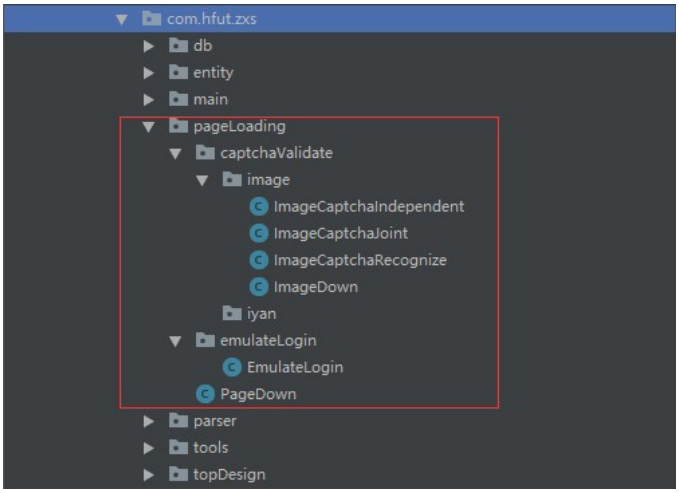
3. **package: main**

- main 包主要用于存放各程序应用的主程序，可按不同的应用创建不同的子包，如图:



4. **package: pageLoading**

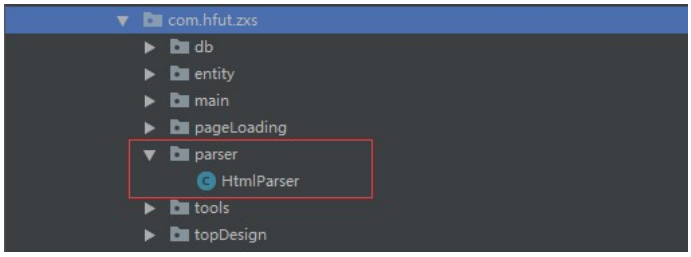
- pageLoading 包主要用于存放获取页面可能涉及到的一些额外操作的工具类。如某些网页的获取可能需要用户登录，则需要模拟登录类；而有些网站的登录需要验证码，则需要进行验证码识别验证的工具类。当然还可能涉及到一些其他的操作，可以在此进行扩展。其结构，如图:



- 其中:
  - **package:captchaValidate**： 用来存放不同类型验证码的获取及识别的工具类。如子包 image 下的 文字验证码的工具类。带有验证码的登录模拟--图片验证码（一）
  - **package:EmulateLogin**： 用来存放进行模拟登陆的工具类。
  - **class:PageDown**： 进行网页下载的工具类。

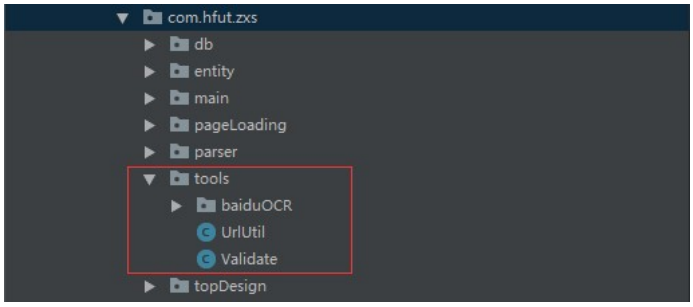
5. **package: parser**

- parser 包主要用于存放解析HTML页面、JSON数据等的工具类，类中的方法与特定的应用相关，如图:



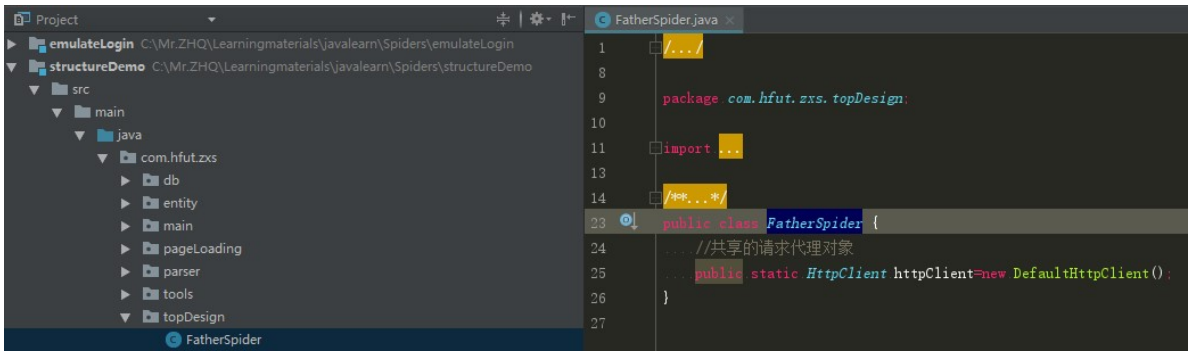
6. **package: tools**

- tools 包主要用于存放一些公用的工具类，如URL转换、字符串校验、百度OCR等，如图：



7. **package: topDesign**

- topDesign 包 主要包含一个类class:FatherSpider,该类基本是很多页面操作类的父类，提供一个公用的httpclient对象，如图：



四、API

类名	包名	功能	属性	方法
SqlSessionFactoryGet	db	获取SqlSessionFactory及SqlSession的提交关闭等		getSqlSessionFactory() : SqlSessionFactory closeSqlSession(SqlSession session) : void
BaiduOCRHost	entity.general	提供百度OCR工具接口	General_BASIC:String ACCURATE_BASIC:String	
CodeType	entity.general	编码格式	GBK: String UTF8: String	
ImageCaptchaIndependent	pageLoading.captchaValidate.image	独立验证码校验		captchaValidate(String url):boolean
ImageCaptchaJoint	pageLoading.captchaValidate.image	获取特定应用的验证码		getXXXCaptcha(String url):Map<String,String>
ImageCaptchaRecognize	pageLoading.captchaValidate.image	进行验证码识别		captchaRecognize(String captchaGetURL, String imageURL,String OCRHost)：String

ImageDown	pageLoading.captchaValidate.image	将图片保存至本地		saveImage(byte[] data,String fileURL):boolean
EmulateLogin	pageLoading.emulateLogin	模拟登陆		login(Map<String,String> loginInfo Map<String,Object> packageLoginInfor(Map<String,Str loginInfor)：StringEntity jsonSubmit(Map<String,String> loginInfor)：StringEntity formSubmit(Map<String,String> loginInfor)：StringEntity
PageDown	pageLoading	网页下载		downPage(HttpUriRequest request String
HtmlParser	parser	html页面解析		
JsonParser	parser	json数据解析		
BaiduOCR	tools.baiduOCR	ocr文字识别		recognizeText(String picURL,String OCRHost)：String
UrlUtil	tools	URL编码、解码		getURLEncoderString(String str,Str ENCODE)：String getURLDecoderString(String str,Str ENCODE)：String
Validate	tools	各种校验。。。		validateNull(Object object)：boolea
FatherSpider	topDesign	多类的父类，使各子类间共享同一httpclient对象	httpClient：HttpClient	