南华大学
UNIVERSITY OF SOUTH CHINA

密级：普通

文档编号：20170630-SDD　　版次：2

分册名称：　　第 1 册/共 1 册

分布式爬虫

（20170630）

代码结构说明书

编　　写　：　杨奎 、刘宝玉

校　　对　：　王浩茂

审　　核　：　刘立

批　　准　：　刘立

生效日期　：　2017 年 6 月 30 日

南华大学计算机学院制

2017 年 6 月

文件修改控制

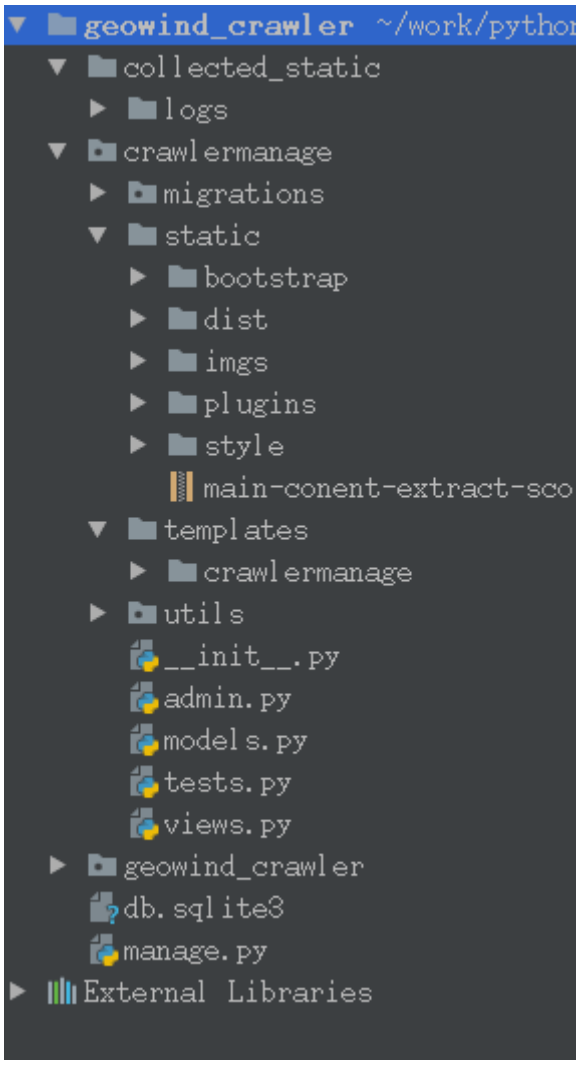
修改记录编号	修改状态	修改页码及条款	修改人	审核人	批准人	修改日期

目录

1. Web 应用代码结构说明

1.1. 整体结构

Web 应用主要采用 Python 编写，开发环境为 PyCharm，代码整体结构如下：



其中主要包括：

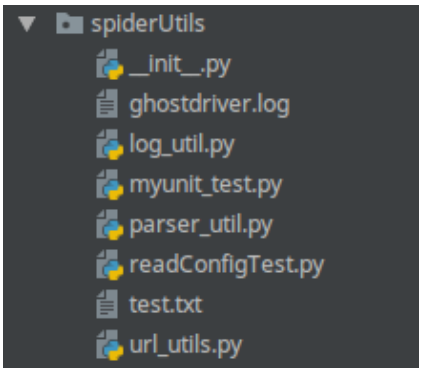
- collected_static 文件夹：其中包括 logs 文件夹，主要负责存储日志。
    - logs 文件夹：存储日志。
- crawlermanage 文件夹：其中包括 migrationythons、static、templates 文件夹。
    - static 文件夹：存储静态文件，包括 CSS 样式、各类组件 JS 脚本和

部分图片等。

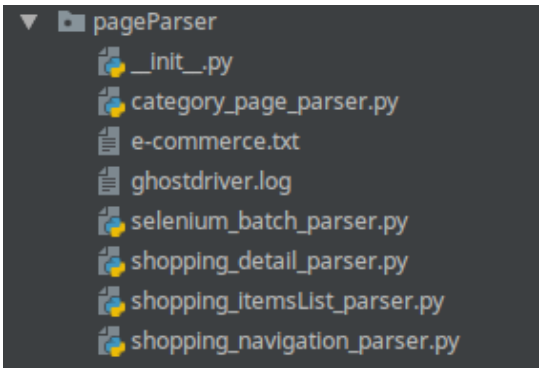
- templates 文件夹：
  - crawlemanage 文件夹：存储所有 html 页面。
- utils 文件夹：工具包。
- geowind_crawler 文件夹：配置文件。

1.2. 各目录结构一览

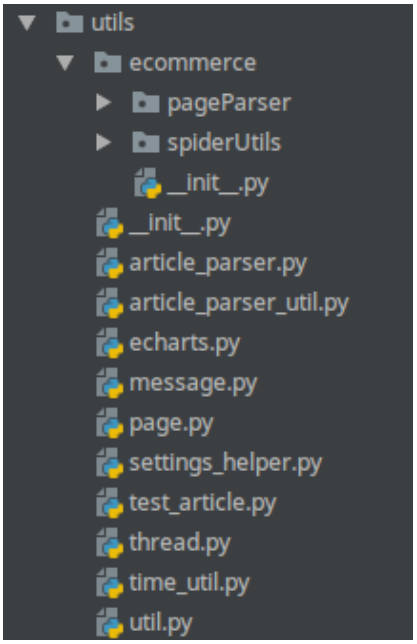
1) spiserUtils 文件夹



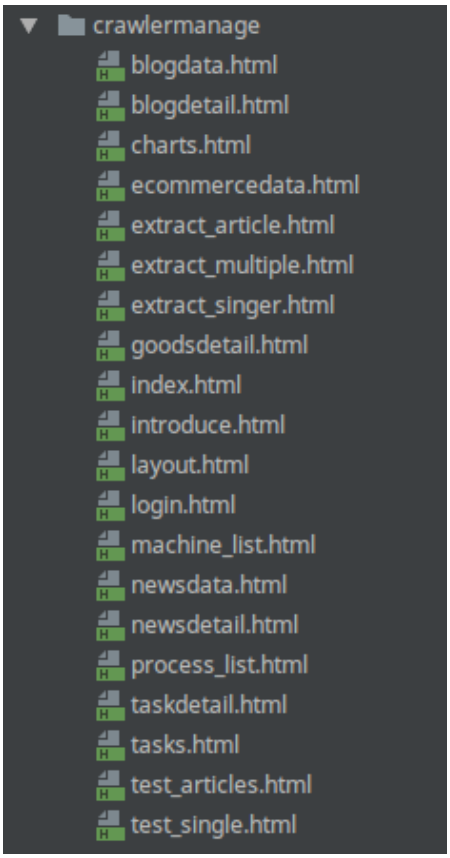
2) pagePaser 文件夹



3) utils 文件夹



4) crawlermanage 文件夹



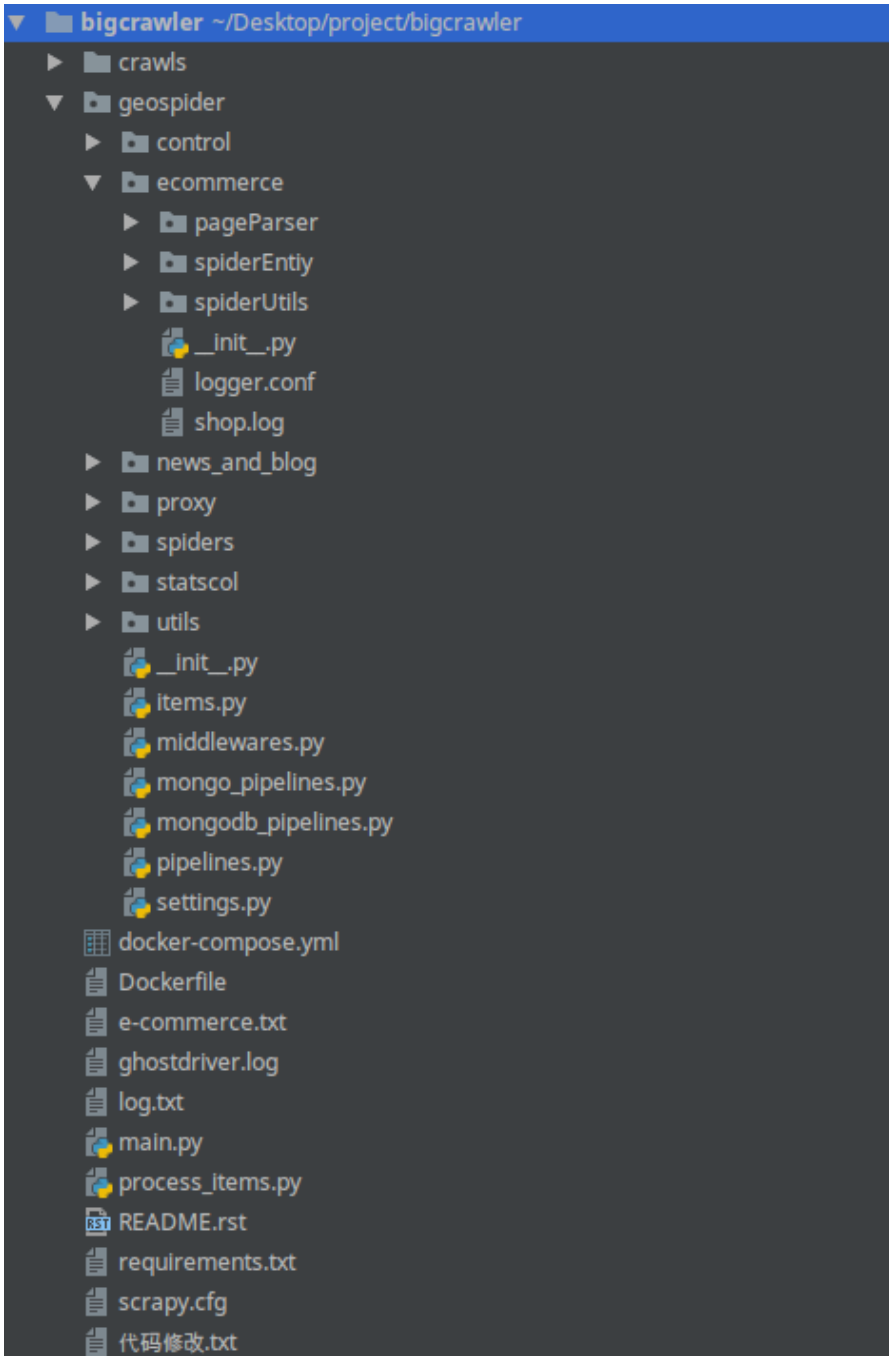
文件名	介绍
blogdata.html	博客列表页
blogdetail.html	博客详细页

charts.html	报表页
Ecommercedata.html	电商列表页
Extract_article.html	正文批量抽取页
Index.html	首页
Introduce.html	使用说明页
Layout.html	任务发布页
Login.html	登陆页
Machine_list.html	从机列表页
Newsdata.html	新闻列表页
Newsdetail.html	新闻详细页
Process_list.html	进程列表页
Taskdetail.html	任务详细页
Tasks.html	任务列表页
Test_articles.html	正文批量测试页
Test_single.html	正文单例测试页

2. 爬虫代码结构说明

2.1. 整体结构

爬虫代码主要采用 Python 编写，开发环境为 PyCharm，代码整体结构如下：



其中主要包括：

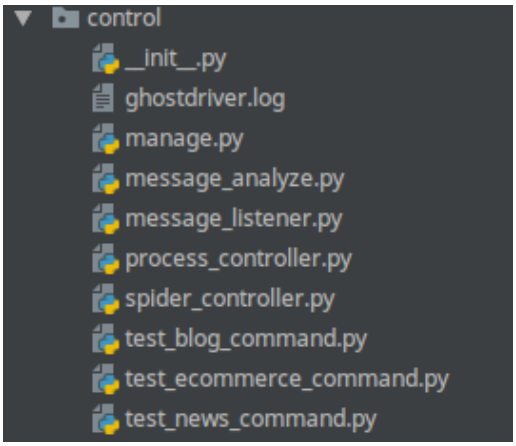
- geospider 文件夹：
  - control 文件夹：包括命令控制、进程控制、爬虫控制文件等。
  - ecommerce 文件夹：电商网站的逻辑部分和一些自定义工具类，主要包括导航栏提取、自动翻页解析、商品详情页面、店铺详情页面解析等模块。

- Pageparser 文件夹：电商网站解析的主要模块包，包括上述主要模块。
- SpiderEntiy 文件夹：存放电商解析中用到的实体类、对象文件等等
- SpiderUtils 文件夹：自定义工具模块，包括 URL 规范化、解析器提取等

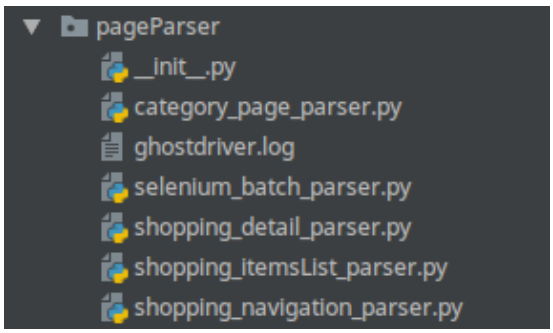
- news_and_blog 文件夹：新闻和博客数据结构化算法设计文件。
- proxy 文件夹：代理模块文件。
- spiders 文件夹：爬虫文件。包括新闻类爬虫、博客、类爬虫和电商类爬虫。
- statscol 文件夹：Graphite 配置文件。
- utils 文件夹：工具包。

2.2. 各目录结构一览

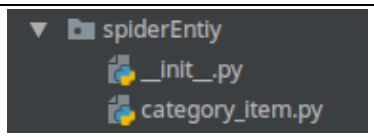
1）control 文件夹



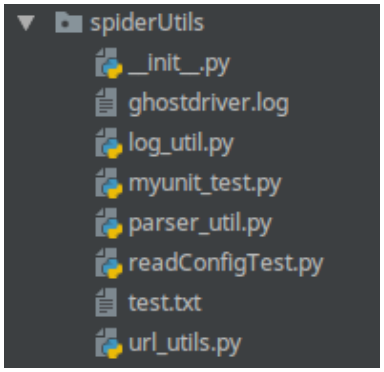
2）pageParser 文件夹



3）spiderEntiy 文件夹

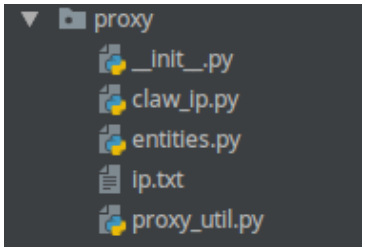


4）spiderUtils 文件夹

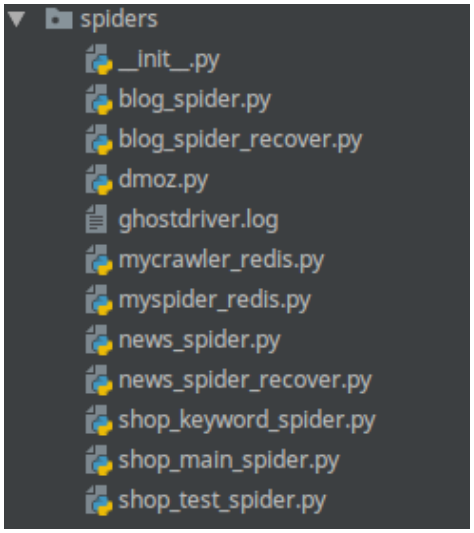


5）news_and_blog 文件夹

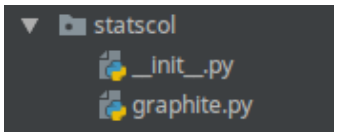
6）proxy 文件夹



7）spiders 文件夹



8）statscol 文件夹



9）util 文件夹