操

作

流

程

|  |  |  |  |
| --- | --- | --- | --- |
| 版本 | 作者 | 版本描述 | 完成日期 |
| 1.0 | Jerry | 基本实现对星魂黄页数据抓取及去重 | 2016.12.26 |

使用介绍

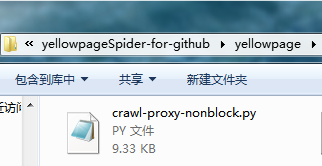
1.安装数据库browser文件



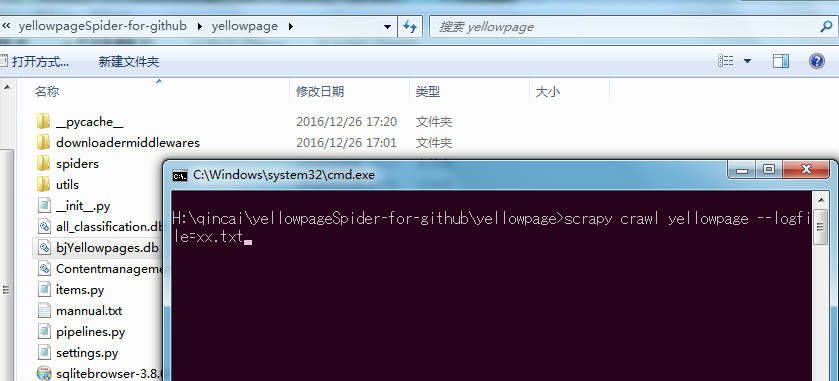
3.用sqlite打开星魂网分类目录页面数据库，罗列了35个一级分类一级下属二级分类、三级分类的url地址，其将作为spider爬虫的start\_urls

 (all\_classificatin.db可由生成

2.命令行中启动代理proxy爬取进程dianping\utils\crawl-proxy-nonblock.py



3.重开一个cmd正式启动爬虫进程scrapy crawl yellowpage，并同时加上日志



4.用sqlitebroser打开根文件夹下的数据库文件bjYellowpages.db，此文件即为爬取到的星魂黄页公司数据。

5. 除了用sqlite保存到本地bjYellowpage.db数据库外，同时也保存了一份到Mysql数据库，见<http://127.0.0.1/phpmyadmin/>数据库yellowpage下的company table



该数据库的连接文件见dbconnect.py：

