操

作

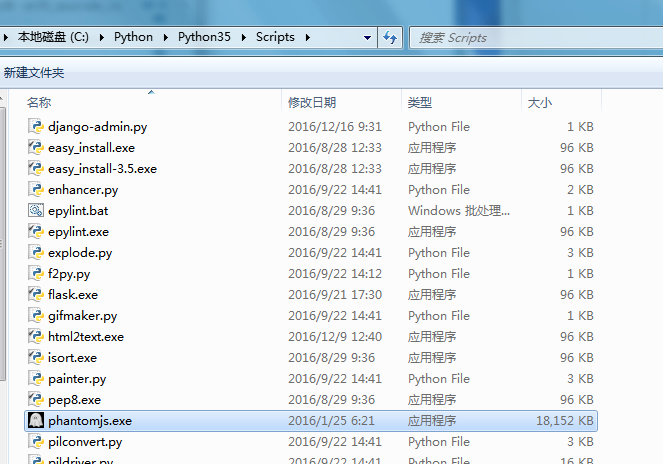
流

程

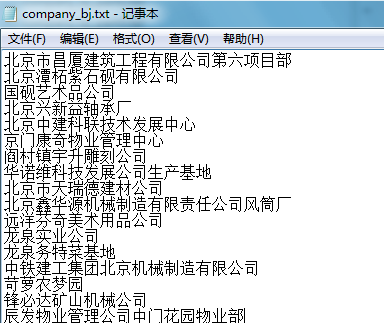
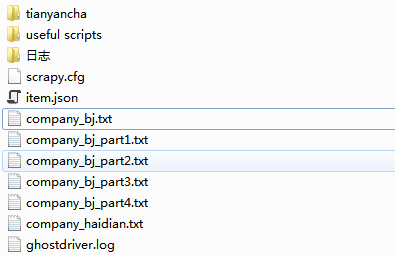
|  |  |  |  |
| --- | --- | --- | --- |
| 版本 | 作者 | 版本描述 | 完成日期 |
| 1.0 | Jerry | 基本实现对天眼查数据抓取及去重 | 2016.12.25 |

使用介绍

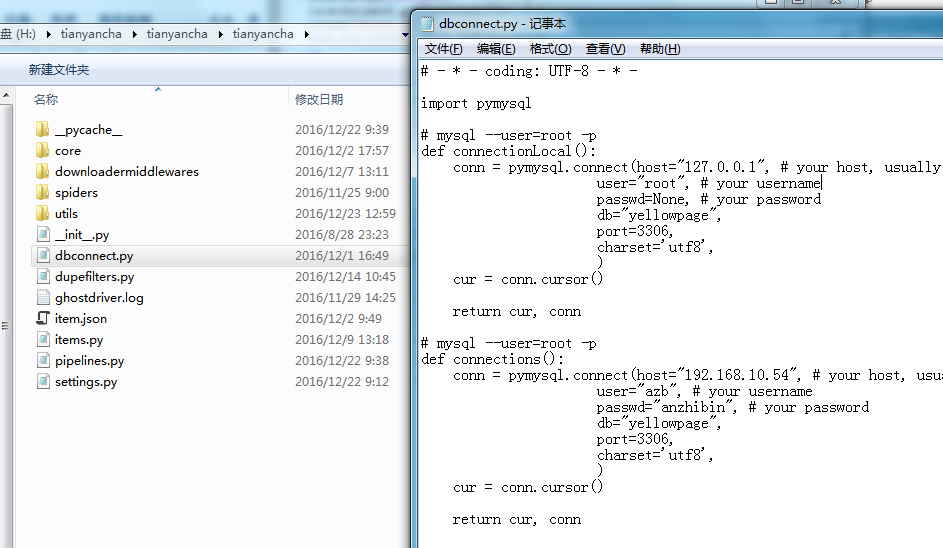
1. 从<http://phantomjs.org/>网站下载对应系统的phantomjs软件，并置于安装目录下



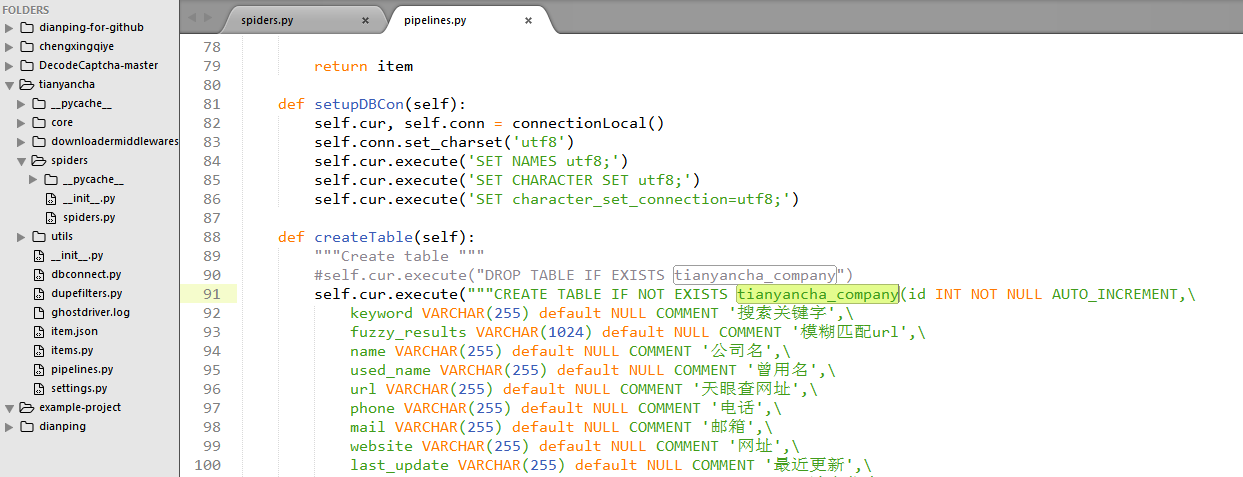
2. 将待查询企业company.txt放置于爬虫项目根目录，并修改spiders.py文件第21行读取该文件。



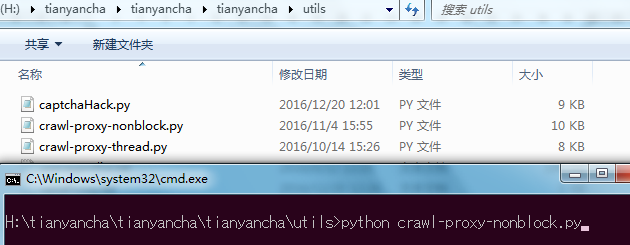
3.修改连接数据库文件tianyancha\tianyancha\dbconnect.py，连接指定数据库



4. 修改项目pipelines.py文件第91行和第130行的数据库tables名字，以防爬虫运行过程中将已有数据表冲掉。



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



同时,如果需要单个代理爬取，则手动修改tianyancha\core\downloader\handlers.py中第62行，其优先级高于downloadmiddleware.py中的IP设置

6.重新开启一个cmd正式启动爬虫进程

