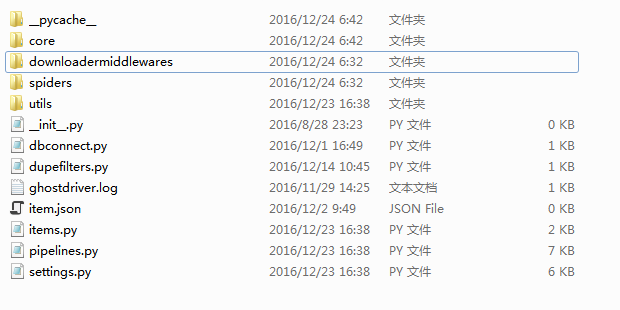
天眼查爬虫目录如下: tianyancha\tianyancha



1.

 定义抓取的网页信息条目，例如：楼盘名，网址等… 用类字典表示

2.

数据库连接文件，连接本地数据库或远程数据库

3.

 爬虫scrapy中个性化设置

4.

 数据库定义及写入的管道

5.

 用来记录重复url， scrapy默认Filtered duplicate request:该语句只记录一次，此处修改为处处记录

6.  定义爬虫的抓取具体逻辑实现

7.  定义爬虫中间件：

其中，

Retry.py为重试中间件

rotate\_proxy为代理ip中间件

rotate\_useragent为代理user-agent中间件

8.  存放user-agent池和ip池

其中，crawl-proxy.py或crawl-proxy-nonblock.py或crawl-proxy-thread.py为获取ip代理的脚本文件;

useragentlist.txt为UA池

validProxy为代理ip池

9.  core\downloader\handlers.py

定义与PhantomJS的连接文件，并将phantomjs作为下载器进行网页下载。