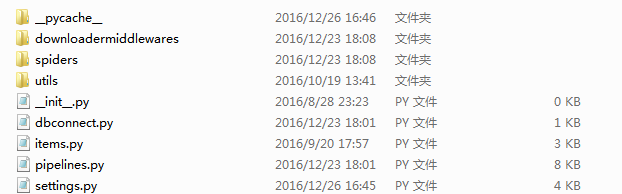
安居客爬虫目录如下:



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.  定义爬虫的抓取具体逻辑实现

6.  定义爬虫中间件：

其中， redirect\_with\_capthca.py 为重定向中间件

rotate\_proxy\_every\_request为代理ip中间件

rotate\_useragent\_every\_request为代理user-agent中间件

7.  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