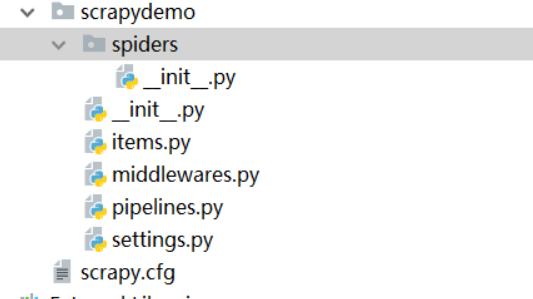
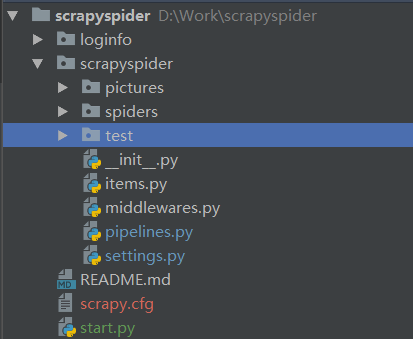
1. scrapy框架介绍



这就是一个scrapy项目的初始结构，我们写的代码都包含在上图的文件中。

然后实际项目中的代码目录是这样



比上图多的几个文件夹的含义

1. loginfo 这是由我们自己定义的用来保存爬虫当前运行状态的文件夹，位置由我们自己决定，这里我们用的相对路径，所以项目保存在运行起点的文件夹里。
2. pictures 是我们用来保存爬取的图片的文件夹，位置可以由我们自己修改，如果想要改变图片的存储路径，修改settings.py文件里的IMAGES\_STORE。如下图



(3) spiders文件夹存储我们写的spider文件

1. 如何启动爬虫

启动爬虫需要在命令行中启动，启动的路径取决于scrapy.cfg 文件所在的位置

比如在我电脑中scrapy.cfg 文件位于,

那么我启动的位置就该在这个地方

启动的命令

scrapy crawl name –s JOBDIR =路径

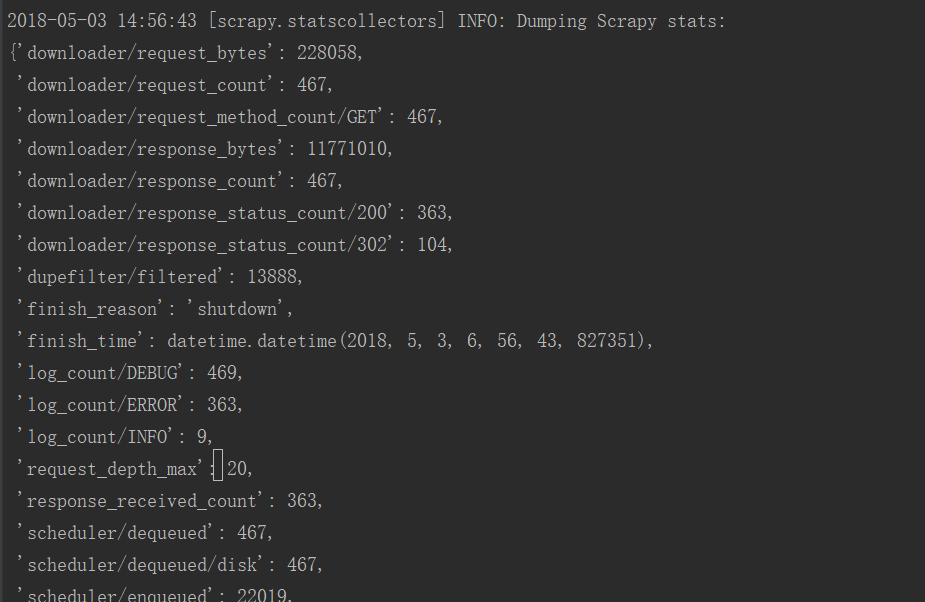
实例：scrapy crawl baike -s JOBDIR=loginfo/001

Name 是一个爬虫的唯一标识，百科的爬虫的name就是 baike， 图片爬虫的name就是pic,JOBDIR 后面就是我们存储爬虫当前运行状态的路径（注意这里每个爬虫当前信息的保存路径不能是一样的）所以如果要在自己的地方运行，那么启动命令就应该这么写

百科爬虫：scrapy crawl baike -s JOBDIR=loginfo/baike（后面的loginfo/baike可以自由修改）

图片爬虫: scrapy crawl pic -s JOBDIR=loginfo/pic

上面是启动命令，如果我们需要暂时结束今天的爬虫任务，那么我们只需要在命令行中使用按下 ctrl+c 就行（注意按下之后需要等大概10秒,出现类似下图的图片就可以终止了）



如果我们需要重新启动爬虫，只需要输入启动命令我们就可以重新启动爬虫了。

1. 需要在本地修改的地方
2. 当前运行状态的保存

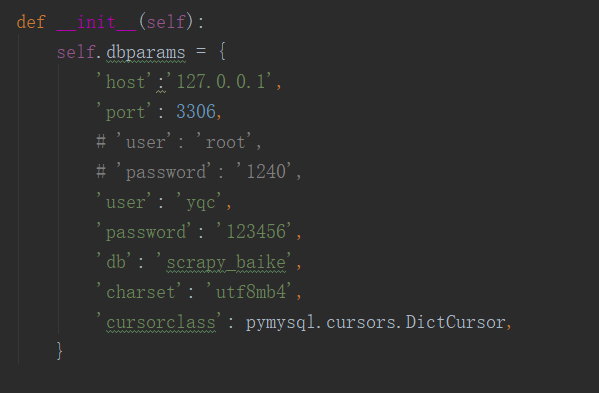
在启动命令的JOBDIR后修改即可，这里使用相对路径即可，请不要使用绝对路径

1. 图片的存储目录

前面已经说了，请查看第一点

1. 数据库的连接

数据库的地址，端口，用户名，密码等请在pipeline.py 文件中BaiKeSpiderPipeline类里面的\_init\_函数修改，如下图,一般只需要修改user和password



1. 附录 SQL语句

数据库名称是：scrapy\_baike(GUI方式创建即可)

表的创建语句:

CREATE TABLE entity\_table (

id BIGINT UNSIGNED AUTO\_INCREMENT PRIMARY KEY,

oid TEXT NOT NULL,

name TEXT NOT NULL,

descrip TEXT NOT NULL,

infobox TEXT,

infolink TEXT,

tag TEXT,

image\_list TEXT

)

CREATE TABLE synonym\_table (

id BIGINT UNSIGNED AUTO\_INCREMENT PRIMARY KEY,

name TEXT NOT NULL,

descrip TEXT NOT NULL,

oid TEXT NOT NULL

)

需要安装的packages

Scrapy,urllib,pymysql

1. 可能会遇到的情况

有时会出现pic爬虫突然正常中止，这是因为baike爬虫收集的数据里的OID都已经被读取，这个时候只需要等待一会，用同样的启动命令启动pic爬虫即可。（这种情况很少出现）