第一周课程学习心得及笔记;

用pip install Scrapy安装

然后在控制台执行如下代码：

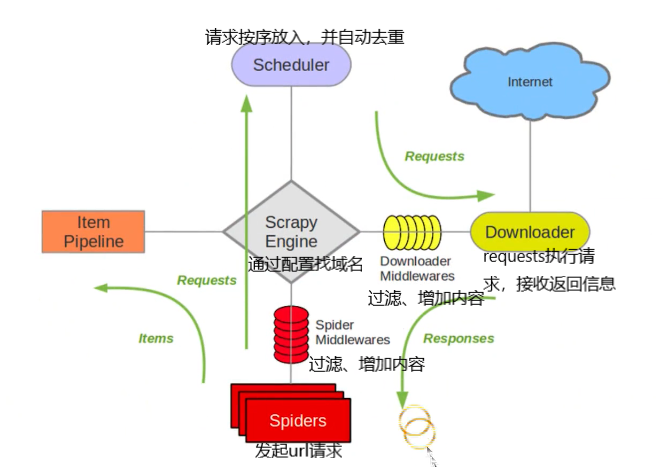
**scrapy startproject spiders**

**cd spiders\spiders**

**scrapy genspider movies douban.com(movies.py 相当于下面的pachong.py)**

****

**各组件调用结构如下：**

****

**一般情况下settings.py保持不动，只有当爬取文件返回异常如4\*\*时，解决返爬虫，改user\_\_agent等，对爬虫进行优化。**

**user\_\_agent、**DOWNLOAD\_DELAY = 3取消注释

**下载延迟一般要打开。**

**movies.py内容如下：**

import scrapy  
from bs4 import BeautifulSoup # bs4 库 BeautifulSoup为包  
from spiders.items import SpidersItem  
  
class MoviesSpider(scrapy.Spider):  
 name = 'douban'#定义爬虫名称，运行的时候用  
 allowed\_domains = ['movie.douban.com'] **#允许访问的域名，网站有向外的链接**

#定义URL起始列表  
 start\_urls = ['https://movie.douban.com/top250'] **#第一次向哪个域名发起请求，twisted异步框架需先发起请求才能启动，另外scrapy 底层事先从start\_urls获取了http头部信息。**  
#注释掉，默认的parse函数从start\_urls里取内容，实际我们是从里面链接取  
 # def parse(self, response):  
 # pass  
**# parse真正要爬取的内容，response请求返回信息，函数首字母小写**

#爬虫启动时，引擎自动调用该方法，并且只会被调用一次，用于生成初始的请求对象（Request）。  
 #start\_requests（）方法默认情况下读取start\_urls列表中的URL并生成Request对象，发送给引擎，引擎再指挥其他组件向网站服务器发送请求，下载网页  
 #这里对start\_requests进行改写  
 def start\_requests(self):  
 for i in range(0,10):  
 url=f'https://movie.douban.com/top250?start={i\*25}'  
 yield scrapy.Request(url=url,callback=self.parse)#yield 暂时当做return  
#Requests调用downloaders，真正发起请求  
#callback回调函数，引擎回将下载好的网页（Response对象）发给该方法，执行数据解析  
#这里可以使用callback指定新的函数，不是用parse作为默认的回调参数  
 def parse(self, response):  
 items=[]  
 soup=BeautifulSoup(response.text,'html.parser')  
 title\_list=soup.find\_all('div',attrs={'class':'hd'})  
 for i in range(len(title\_list)):  
 item=SpidersItem()  
 title=title\_list[i].find('a').find('span',).text  
 link=title\_list[i].find('a').get('href')  
 item['title']=title  
 item['link']=link  
 items.append(item)  
 return items

**items.py里定义两个新的item：**

import scrapy  
class SpidersItem(scrapy.Item):  
 # define the fields for your item here like:  
 # name = scrapy.Field()  
 #pass  
 title=scrapy.Field()  
 link=scrapy.Field()

**运行爬虫：在cmd终端用命令scrapy crawl douban**