scrapy框架

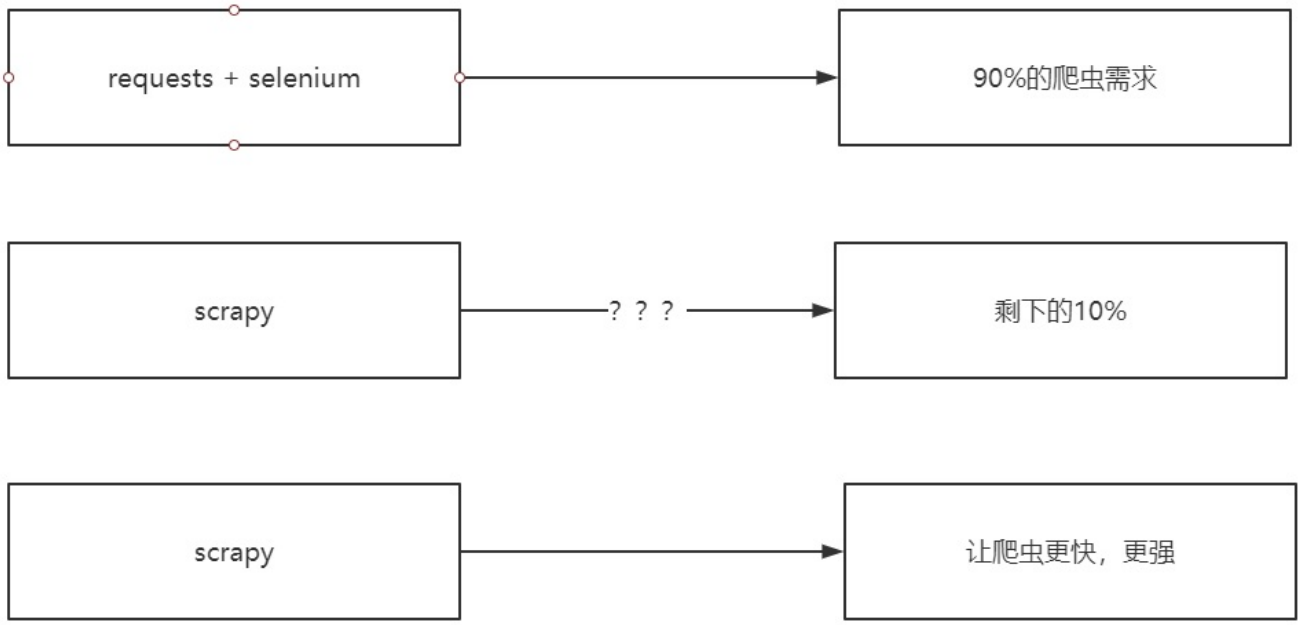
ju7ran

目　录

空白目录

1-scrapy的介绍

为什么要学习scrapy



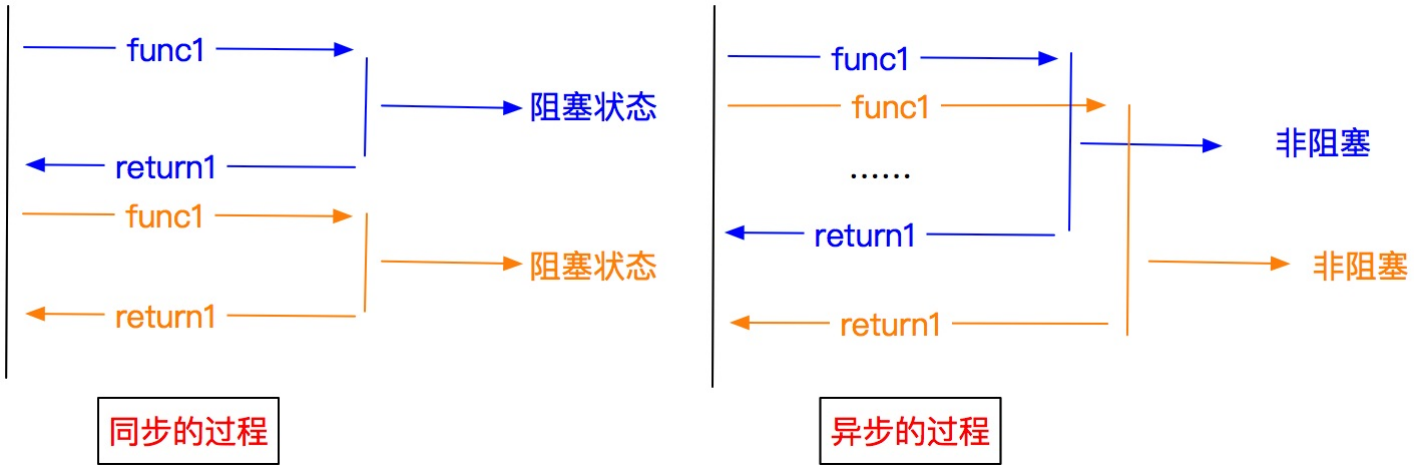
什么是Scrapy

Scrapy是一个为了爬取网站数据，提取结构性数据而编写的应用框架，我们只需要实现少量的代码，就能够快速的抓取

Scrapy使用了Twisted异步网络框架，可以加快我们的下载速度

http://scrapy-chs.readthedocs.io/zh_CN/1.0/intro/overview.html

异步和非阻塞的区别

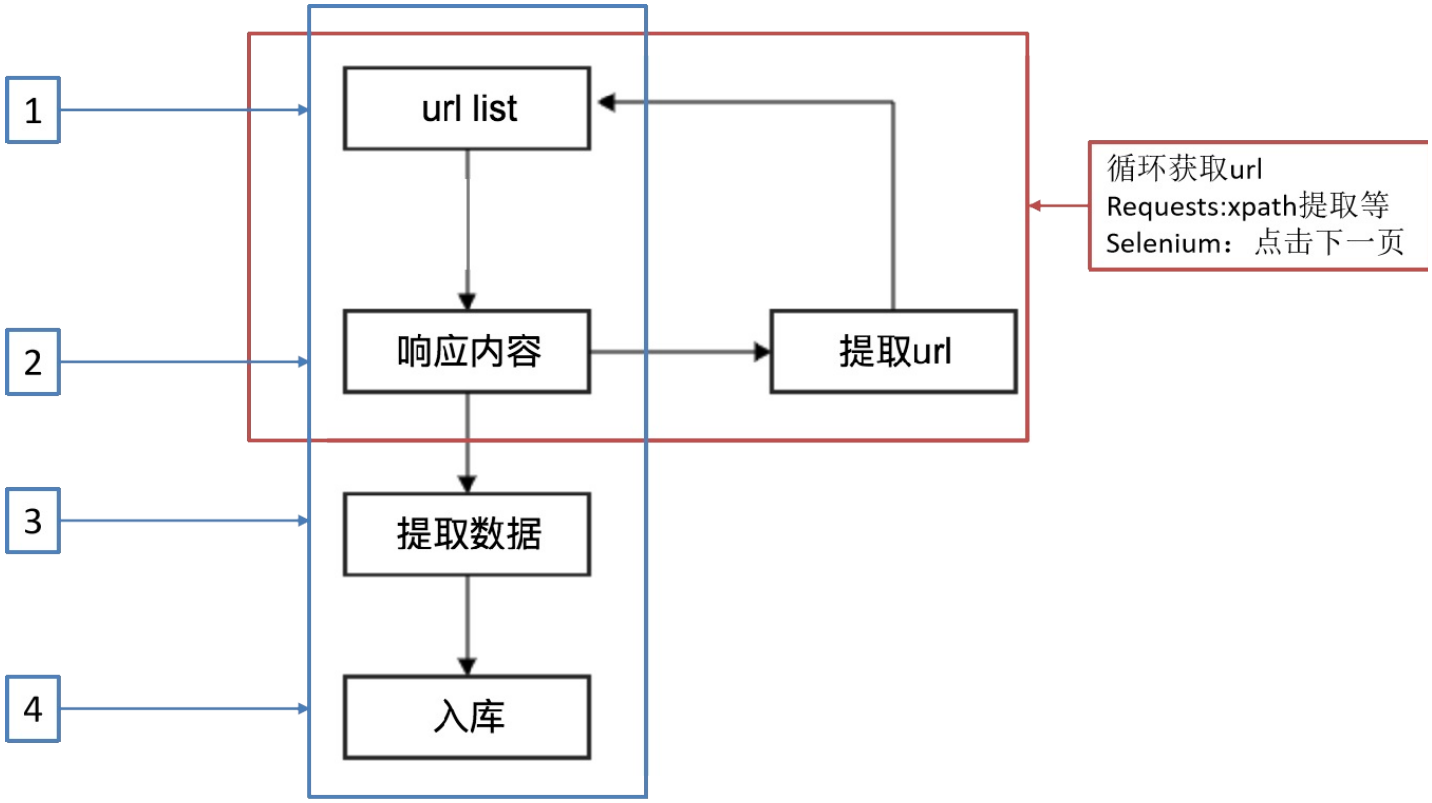


异步：调用在发出之后，这个调用就直接返回，不管有无结果

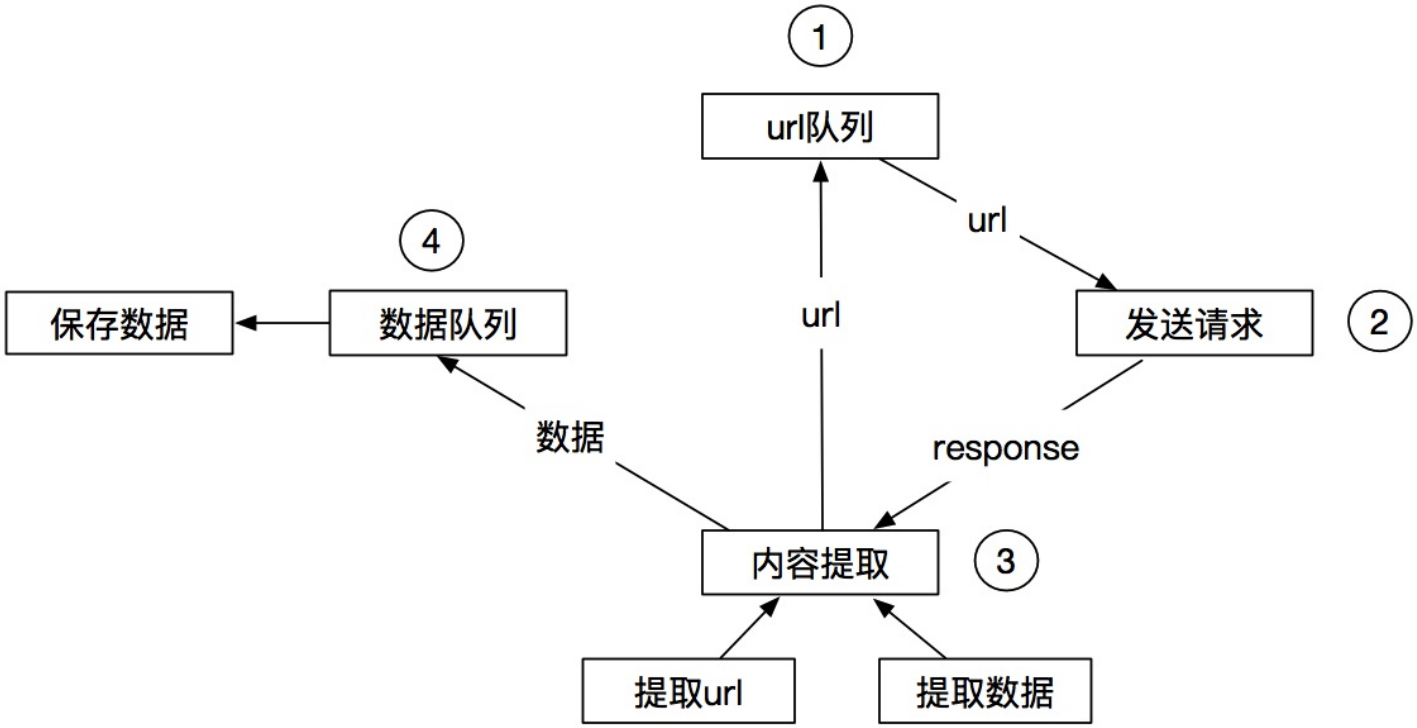
非阻塞：关注的是程序在等待调用结果时的状态，指在不能立刻得到结果之前，该调用不会阻塞当前线程。

Scrapy工作流程

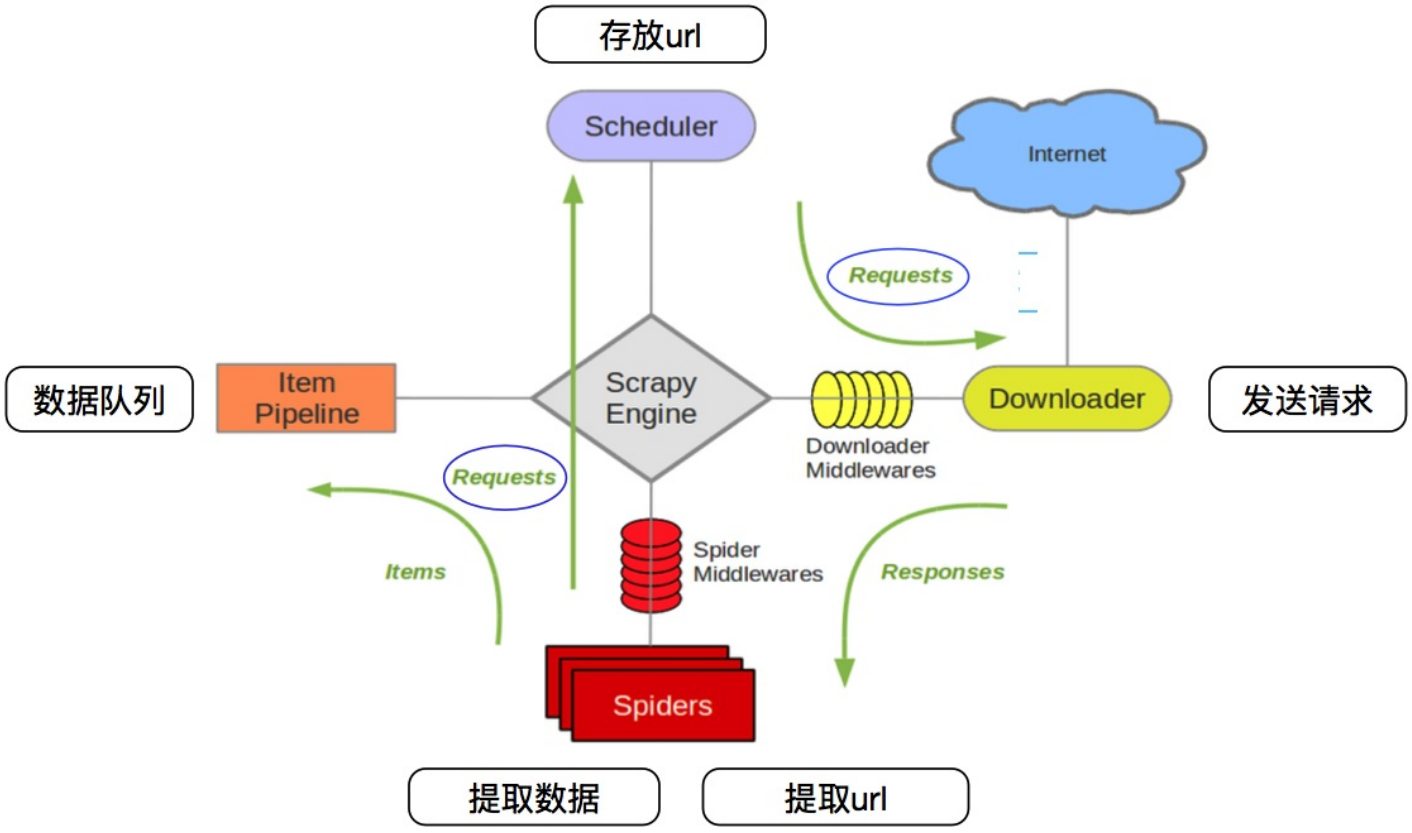
回顾之前的爬虫流程



另外一种爬虫方式



Scrapy的爬虫流程



Scrapy engine(引擎)	总指挥:负责数据和信号的在不同模块间的传递	scrapy已经实现
Scheduler(调度器)	一个队列,存放引擎发过来的request请求	scrapy已经实现
Downloader(下载器)	下载把引擎发过来的requests请求,并返回给	scrapy已经实现

Downloader(下载器)	引擎	scrapy已经实现
Spider(爬虫)	处理引擎发来的response,提取数据,提取url,并交给引擎	需要手写
Item Pipline(管道)	处理引擎传过来的数据,比如存储	需要手写
Downloader Middlewares(下载中间件)	可以自定义的下载扩展，比如设置代理	一般不用手写
Spider Middlewares(中间件)	可以自定义requests请求和进行response过滤	一般不用手写

Scrapy入门

```
1 创建一个scrapy项目
scrapy startproject mySpider

2 生成一个爬虫
scrapy genspider demo "demo.cn"

3 提取数据
完善spider 使用xpath等

4 保存数据
pipeline中保存数据
```

创建一个scrapy项目

2-pipeline-item-shell

1 使用pipeline

从pipeline的字典形可以看出来，pipeline可以有多个，而且确实pipeline能够定义多个

为什么需要多个pipeline：

1 可能会有多个spider，不同的pipeline处理不同的item的内容

2 一个spider的内容可以要做不同的操作，比如存入不同的数据库中

注意：

1 pipeline的权重越小优先级越高

2 pipeline中process_item方法名不能修改为其他的名称

2 logging模块的使用

爬虫文件

```
import scrapy
import logging

logger = logging.getLogger(__name__)

class QbSpider(scrapy.Spider):
    name = 'qb'
    allowed_domains = ['qiushibaike.com']
    start_urls = ['http://qiushibaike.com/']

    def parse(self, response):
        for i in range(10):
            item = {}
            item['content'] = "haha"
            # logging.warning(item)
            logger.warning(item)
            yield item
```

pipeline文件

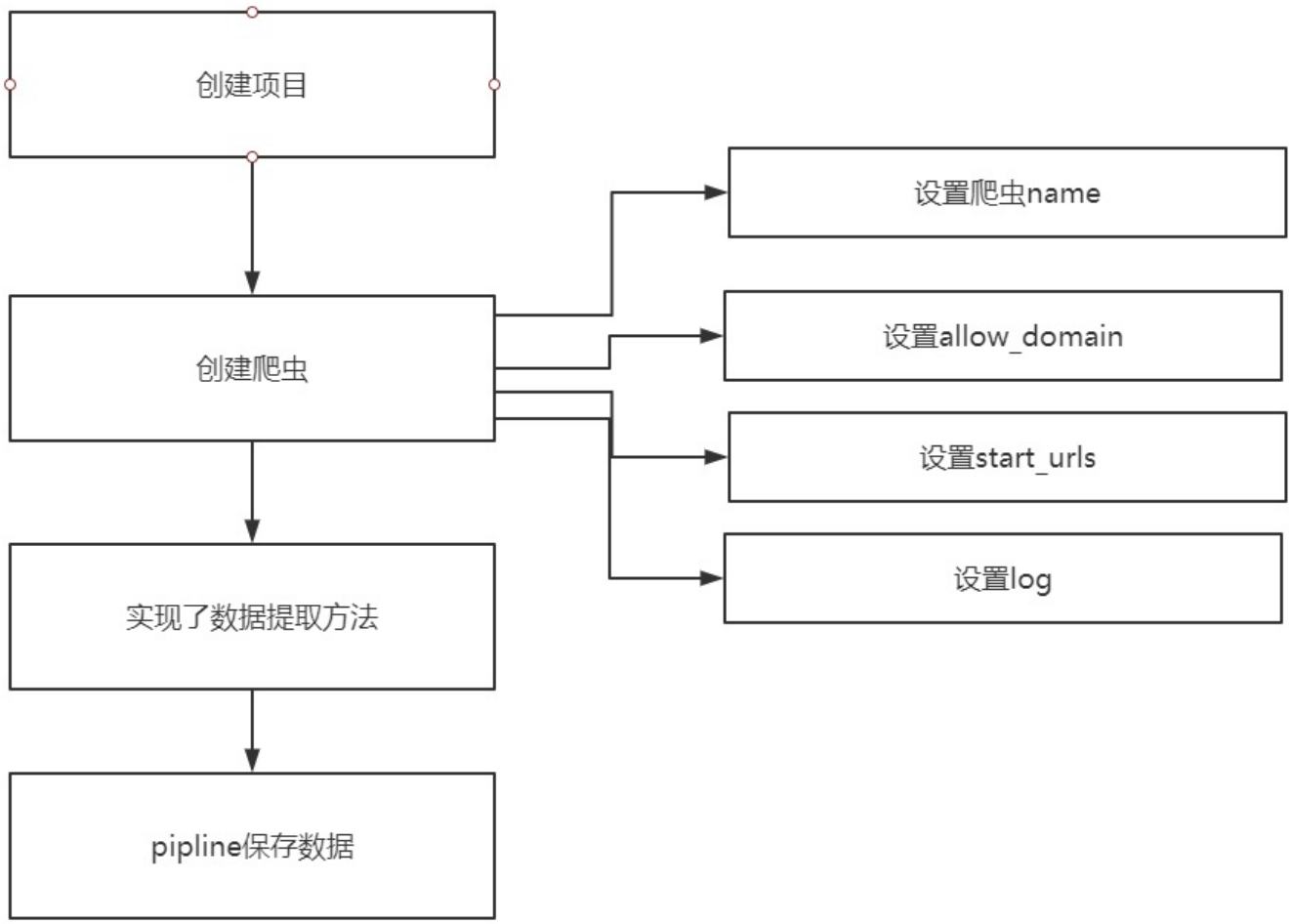
```
import logging

logger = logging.getLogger(__name__)

class MyspiderPipeline(object):
    def process_item(self, item, spider):
        # print(item)
        logger.warning(item)
        item['hello'] = 'world'
        return item
```

保存到本地，在setting文件中 `LOG_FILE = './log.log'`

回顾



问题来了：如何实现翻页请求

回忆：

requests模块是如何发送翻页的请求的?

1、找到下一页地址

2、之后调用requests.get(url)

思路：

1、找到下一页的地址

2、构造一个关于下一页url地址的request请求传递给调度器

3 腾讯爬虫

通过爬取腾讯招聘的页面的招聘信息,学习如何实现翻页请求

http://hr.tencent.com/position.php

```
创建项目
scrapy startproject tencent

创建爬虫
scrapy genspider hr tencent.com
```

scrapy.Request知识点

```
scrapy.Request(url, callback=None, method='GET', headers=None, body=None,cookies
=None, meta=None, encoding='utf-8', priority=0,
dont_filter=False, errback=None, flags=None)

常用参数为：
callback：指定传入的URL交给那个解析函数去处理
meta：实现不同的解析函数中传递数据，meta默认会携带部分信息,比如下载延迟，请求深度
dont_filter:让scrapy的去重不会过滤当前URL，scrapy默认有URL去重功能，对需要重复请求的URL
有重要用途
```

4 item的介绍和使用

```
items.py

import scrapy

class TencentItem(scrapy.Item):
    # define the fields for your item here like:
```

```
    title = scrapy.Field()
    position = scrapy.Field()
    date = scrapy.Field()
```

在hr.py中直接导入items,实例化TencentItem，就可以使用了

为什么要定义一个字典？

在获取到数据的时候，使用不同的item来存放不同的数据

在把数据交给pipeline的时候，可以通过isinstance(item,MyspiderItem)来判断数据是属于哪一个item，进行不同的数据处理

5 阳光政务平台

http://wz.sun0769.com/index.php/question/questionType?type=4&page=0

6 debug信息的认识

```
[scrapy.utils.log] INFO: Overridden settings:自己设置的setting的信息
[scrapy.middleware] INFO: Enabled extensions:启动的扩展，默认有一堆
[scrapy.middleware] INFO: Enabled downloader middlewares:启动的下载中间价，默认一堆
[scrapy.middleware] INFO: Enabled spider middlewares:启动的爬虫中间件，默认一堆
[scrapy.middleware] INFO: Enabled item pipelines: 启动的管道
[scrapy.extensions.telnet] DEBUG: 爬虫运行的时候能够使用talent命令对爬虫做一些控制，比如暂停等
[scrapy.statscollectors] INFO: Dumping Scrapy stats:爬虫结束时候的一些统计信息，比如请求响应数量等
[scrapy.core.scraper] DEBUG: Scraped from <200 http://wz.sun0769.com/html/question/201707/3
40346.shtml>{'content':.....}     :每次yield item的时候回提示item的内容以及这个item来自的url地址
```

Scrapy深入之scrapy shell

Scrapy shell是一个交互终端，我们可以在未启动spider的情况下尝试及调试代码，也可以用来测试XPath表达式

使用方法：

scrapy shell http://www.itcast.cn/channel/teacher.shtml

```
response.url:当前相应的URL地址
response.request.url:当前相应的请求的URL地址
response.headers:响应头
response.body:响应体，也就是HTML代码，默认是byte类型
response.requests.headers:当前响应的请求头
```