Scrapy

1、新建项目：scrapy startproject ProjectName

2、编写爬虫

**import** **scrapy**

**class** **QuotesSpider**(scrapy.Spider):

name = "quotes" # 爬虫名，运行时指定此名

**def** start\_requests(self): # 定义入口url (可以用start\_urls=[]替代此函数)

urls = [

'http://quotes.toscrape.com/page/1/',

'http://quotes.toscrape.com/page/2/',

]

**for** url **in** urls:

**yield** scrapy.Request(url=url, callback=self.parse) # 可以有多个parse回调函数

**def** parse(self, response): # 每一个url读取后的回调函数

# 此函数调用yield 返回一个dict，作为Log值记录在控制台或者文件中(-o log.json)

page = response.url.split("/")[-2]

filename = 'quotes-*%s*.html' % page

yield {“log” : “hello”} # 生成日志

yield scrapy.Request(“next\_page.html”, callback=self.parse) # 产生下级页面

**with** open(filename, 'wb') **as** f:

f.write(response.body) # 保存文件

self.log('Saved file *%s*' % filename)

# scrapy.follow 可以使用带url 的dom元素。

3、开始爬数据

scrapy crawl spider\_name

4、测试scrapy的selector 功能

scrapy shell "http://quotes.toscrape.com/page/1/"

（linux下用单引号）

此时可以直接操作response 变量。

5、CSS selector

获取属性值： response.css(“a::attr(href)”)