[1. 主页上sample 出错 2](#_Toc421809515)

[2. 设置代理 3](#_Toc421809516)

[3. Scrapy简明工作原理 3](#_Toc421809517)

[4. XPath最通俗的教程（ZZ） 3](#_Toc421809518)

[5. 正则表达式 3](#_Toc421809519)

[6. python scrapy parse() function, where is the return value returned to? 4](#_Toc421809520)

[7. Architecture overview 5](#_Toc421809521)

[8. Scrapy Docs 7](#_Toc421809522)

[8-1. Spiders 8](#_Toc421809523)

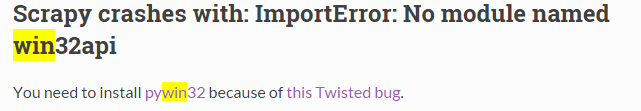
[8-2. 一个spider还是多个spider 9](#_Toc421809524)

[8-3. 对于确实需要多个spider的情况 10](#_Toc421809525)

# 主页上sample 出错



<http://doc.scrapy.org/en/latest/faq.html#scrapy-crashes-with-importerror-no-module-named-win32api>



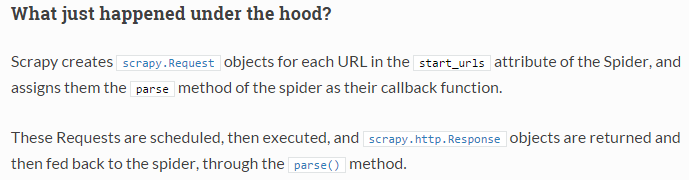
先下载的64位的（因为操作系统是64位的），但安装时提示找不到python的安装路径。READEME提示说可能是版本同Python版本不匹配，换了个32位的，解决

# 设置代理

<http://stackoverflow.com/questions/4710483/scrapy-and-proxies>

在运行之前命令行中先设置set http\_proxy=http://proxy:port

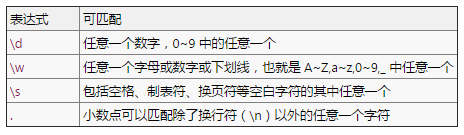
# Scrapy简明工作原理



# XPath最通俗的教程（ZZ）

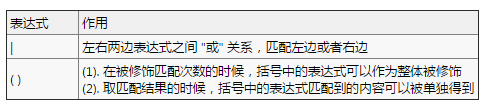
<http://ipointer.cnblogs.com/archive/2005/10/20/258305.html>

# 正则表达式

http://www.regexlab.com/zh/regref.htm  




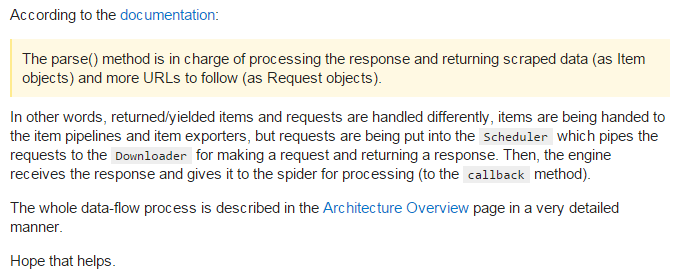




一个比较重要的地方是()里面的子表达式的结果可以单独得到

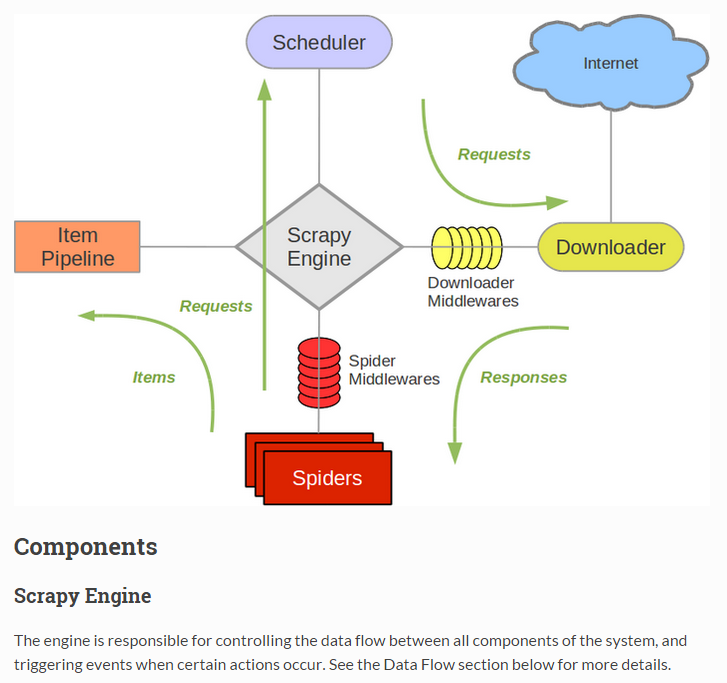
# python scrapy parse() function, where is the return value returned to?

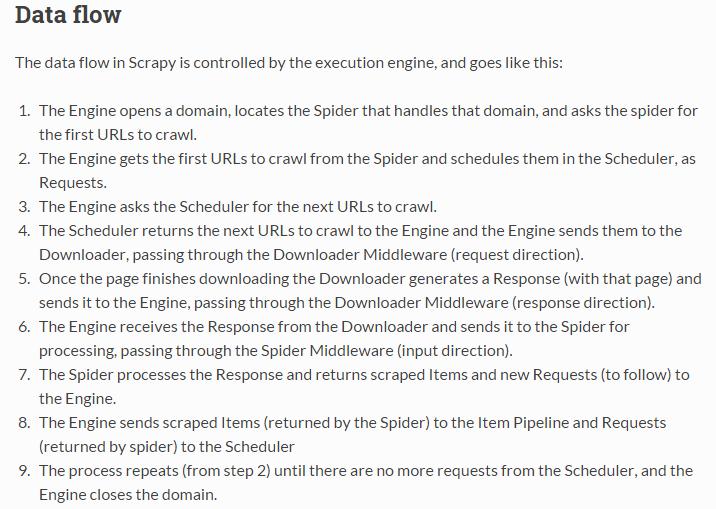
<http://stackoverflow.com/questions/26195982/python-scrapy-parse-function-where-is-the-return-value-returned-to>



# Architecture overview

<http://doc.scrapy.org/en/latest/topics/architecture.html>

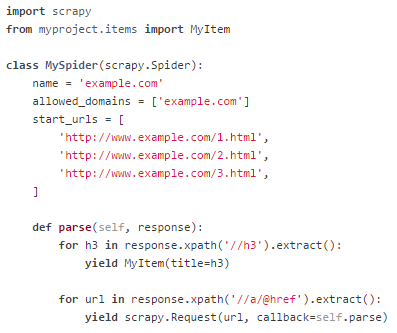


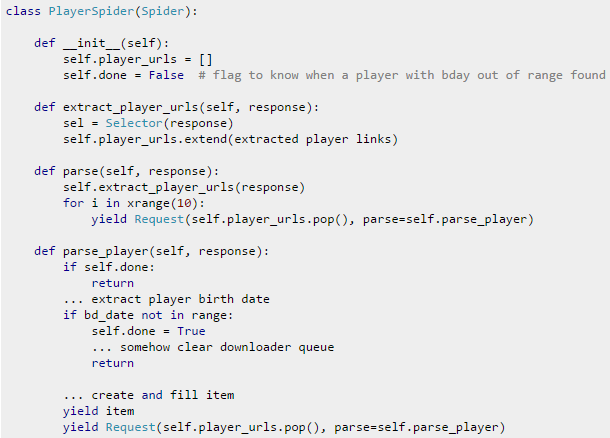


# Scrapy Docs

<http://doc.scrapy.org/en/latest/topics/spiders.html>

## Spiders

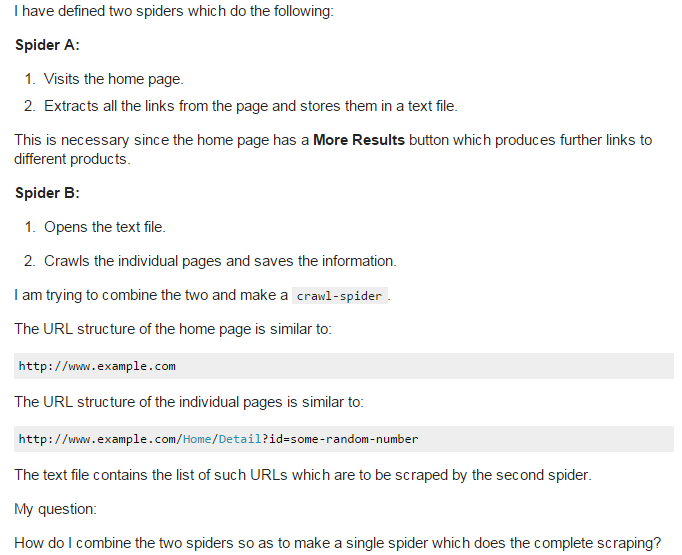


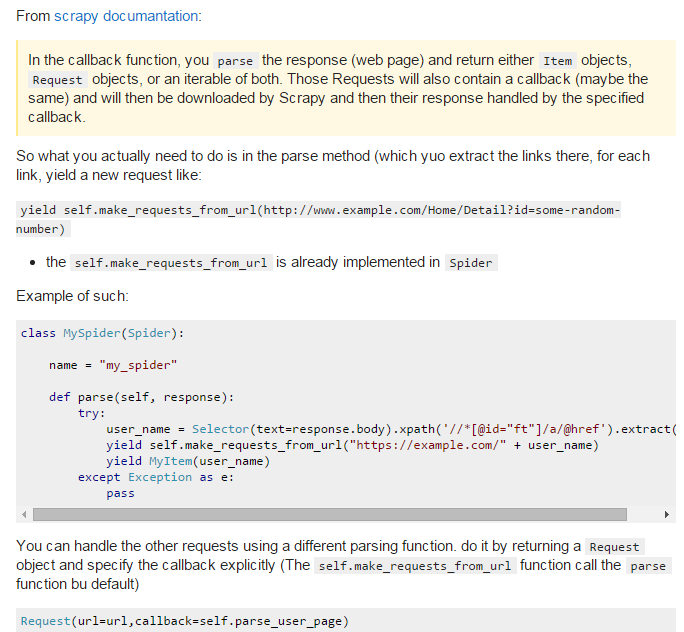


## 一个spider还是多个spider

对于先分析一个主页，然后对主页里的各个连接分别处理的场景，不用多个spider，而是用一个spider的多个parse函数而已。

[Scrapy multiple spiders](http://stackoverflow.com/questions/28023996/scrapy-multiple-spiders)

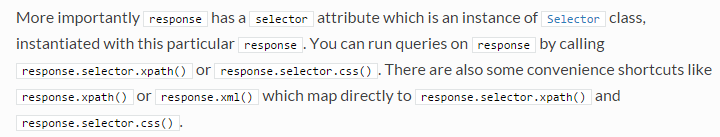




## 对于确实需要多个spider的情况

[Running multiple scrapy spiders programmatically](http://kirankoduru.github.io/python/multiple-scrapy-spiders.html)

# 为什么能对response调用xpath函数



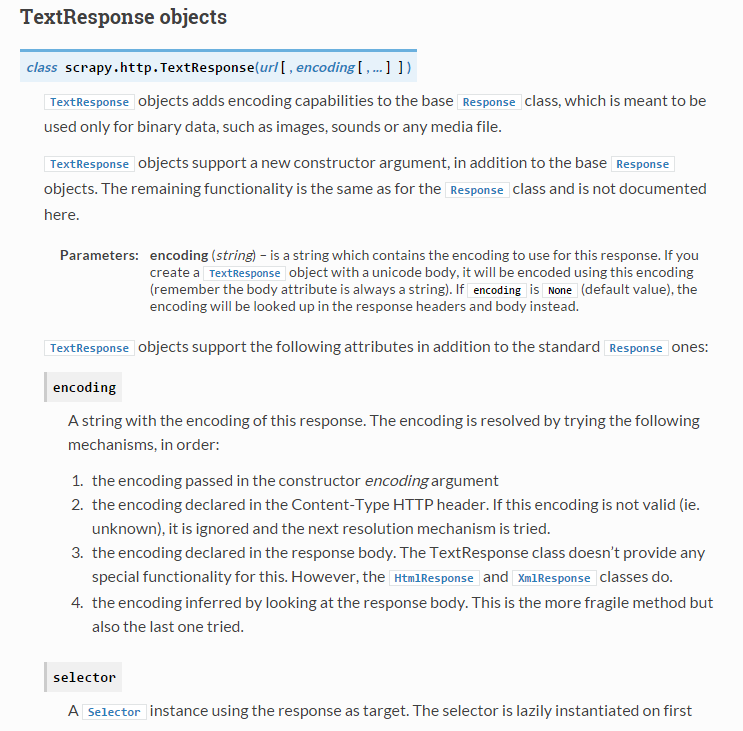
# Request and Response

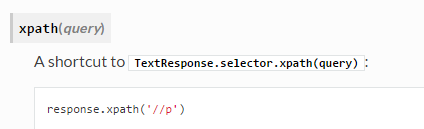
都是引自 scrapy.http package



在我们的parse函数中返回的reponse， 应该不是Response的基类，而是HtmlReponse类，而HtmlResponse又继承自TextResponse

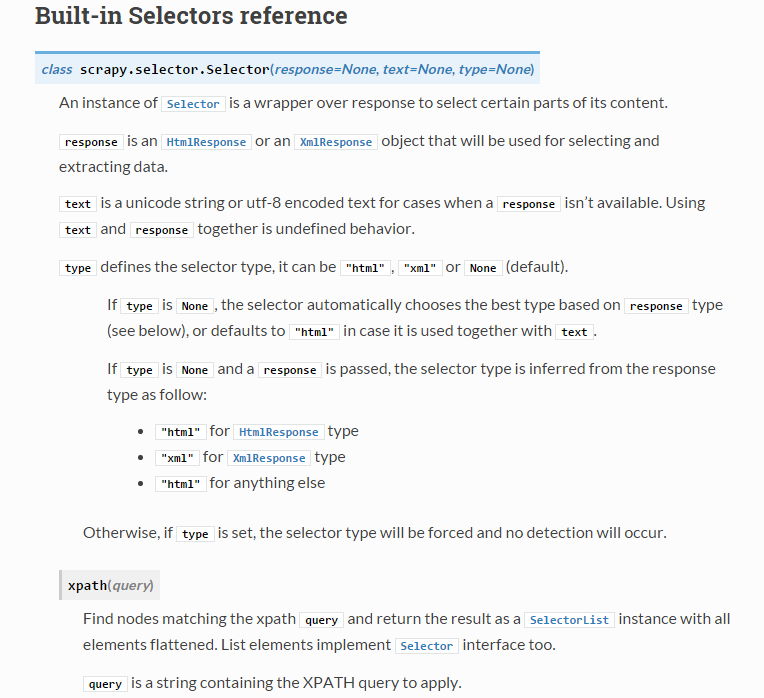
TextResponse包含有selector属性和xpath方法。





# Selector 和SelectorList

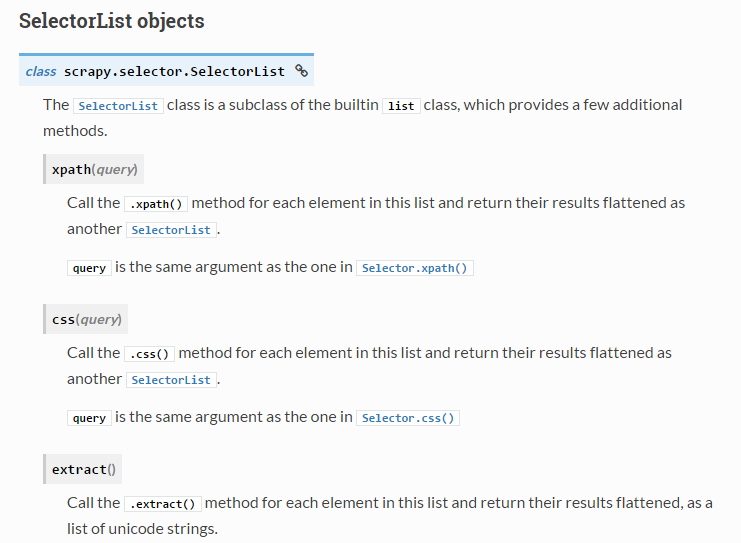
<http://doc.scrapy.org/en/latest/topics/selectors.html>



注意xpath返回的是一个SelectionList 集合

SelectionList也实现了Selector 接口，同样有xpath属性和extract方法

Extracat方法返回的是字符串



## Relative XPaths

相对路径，前面要么不加/，要么加上”.”, 任何以’/’ 或者’//’ 开头的xpath都是相对于根，而不是相对于前面你的selector

