Python安装

<https://www.runoob.com/python/python-install.html>

配置环境变量 python pip3

参考git文档

<https://github.com/scrapy/scrapy>

安装 scrapy，注意报错有可能是网络问题

pip install scrapy

报错解决

[https://www.lfd.uci.edu/~gohlke/pythonlibs/#twisted](https://www.lfd.uci.edu/~gohlke/pythonlibs/" \l "twisted)

下载whl文件安装

pip install wheel-0.35.1-py3-none-any.whl

下载对应chromedriver的驱动

<http://npm.taobao.org/mirrors/chromedriver/>

http://chromedriver.storage.googleapis.com/index.html

开始code

scrapy startproject myscrapy

cd myscrapy

scrapy genspider example example.com

scrapy crawl taobao

Xpath

<https://www.w3school.com.cn/xpath/xpath_syntax.asp>

Python：yield关键字以及next、send函数的作用

<https://blog.csdn.net/mingC0758/article/details/53783001?utm_medium=distribute.pc_relevant.none-task-blog-BlogCommendFromMachineLearnPai2-1.channel_param&depth_1-utm_source=distribute.pc_relevant.none-task-blog-BlogCommendFromMachineLearnPai2-1.channel_param>

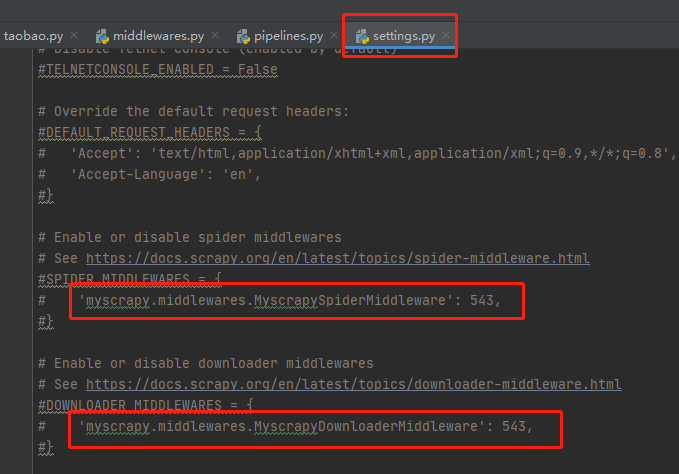
我们得出以下结论：

1、带有yield关键字的函数自动变成生成器

2、生成器被调用时不会立即执行

next函数与send函数很相似，都能获得生成器的下一个yield后面表达式的值，不同的是****send函数可以向生成器传参****：

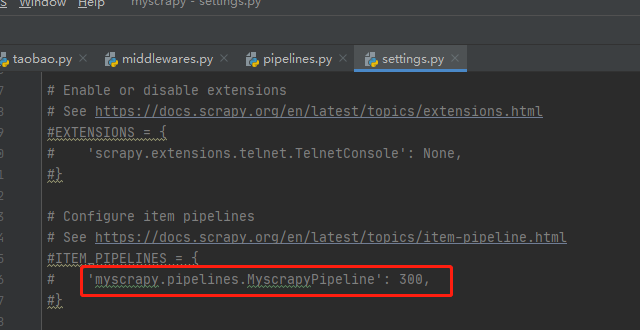
=== Middleware ===



Selenium 在 SPIDER\_MIDDLEWARES 中

=== pipelines ===

管道 对应 取 yield 的内容



ITEM\_PIPELINES

<https://blog.csdn.net/harry5508/article/details/86486777/>

### **selenium规避检测**

现在不少大网站有对selenium采取了监测机制。比如正常情况下我们用浏览器访问淘宝等网站的 window.navigator.webdriver的值为 undefined。  
而使用selenium访问则该值为true。

### **解决办法**

from selenium.webdriver import Chromefrom selenium.webdriver import ChromeOptions

option = ChromeOptions()

option.add\_experimental\_option('excludeSwitches', ['enable-automation'])

driver = Chrome(r'浏览器驱动程序路径',options=option)

# 其他操作不用变,再去console一下window.navigator.webdriver就显示undefined了