**python爬虫十二：爬取快速ip代理，攻破503**

转：https://zhuanlan.zhihu.com/p/26701898

1.自定爬虫方法

*# -\*- coding: utf-8 -\*-*

import scrapy

import requests

from proxy.items import ProxyItem

class DxdlspiderSpider(scrapy.Spider):

name = 'dxdlspider'

allowed\_domains = ['xicidaili.com']

start\_urls = ['http://www.xicidaili.com/wt/']

def parse(self, response):

items=[]

*#获取网页信息*

add\_list1=response.xpath('//tr[@class=""]');

add\_list2=response.xpath('//tr[@class="odd"]');

add\_list=add\_list1+add\_list2;

print (add\_list)

test=testProxy();

for add in add\_list:

item=ProxyItem();

*#print(add.xpath('./td//text()')[0].extract()+':'+add.xpath('./td//text()')[1].extract())*

address=add.xpath('./td//text()')[0].extract() + ':' + add.xpath('./td//text()')[1].extract()

*#测试是否能用*

*# if test.ipUtils(address):*

*# item['address']=address*

*# items.append(item)*

item['address'] = address

items.append(item)

return items

class testProxy:

def ipUtils(self,proxy):

*# 设置代理头*

proxies = {'http': proxy}

print('正在测试：{}'.format(proxies))

try:

r = requests.get('http://www.baidu.com', proxies=proxies, timeout=3)

if r.status\_code == 200:

print('该代理：{}成功存活'.format(proxy))

return True

except:

print('该代理{}失效！'.format(proxies))

return False

2、items.py

*# -\*- coding: utf-8 -\*-*

*# Define here the models for your scraped items*

*#*

*# See documentation in:*

*# https://doc.scrapy.org/en/latest/topics/items.html*

import scrapy

class ProxyItem(scrapy.Item):

*# define the fields for your item here like:*

*# name = scrapy.Field()*

address=scrapy.Field();

3.pipelins.py存入数据库

*# -\*- coding: utf-8 -\*-*

*# Define your item pipelines here*

*#*

*# Don't forget to add your pipeline to the ITEM\_PIPELINES setting*

*# See: https://doc.scrapy.org/en/latest/topics/item-pipeline.html*

import os

import requests

import json

import codecs

import pymysql

class ProxyPipeline(object):

def process\_item(self, item, spider):

return item

*#将数据存入数据库*

class Rmysql(object):

def process\_item(self, item, spider):

*#抓取数据存入mysql*

*#将数据从item中拿出来*

address=item['address'];

*#和本地的数据库连接起来*

connection=pymysql.connect(

host='localhost', *# 连接的是本地数据库*

user='root', *# 自己的mysql用户名*

passwd='123456', *# 自己的密码*

db='scrapyDB', *# 数据库的名字*

charset='utf8mb4', *# 默认的编码方式：*

cursorclass=pymysql.cursors.DictCursor)

*#插库*

try:

with connection.cursor() as cursor:

*#创建更新数据库的sql*

sql="""INSERT INTO address(address)

VALUES (%s)"""

*#执行sql*

cursor.execute(sql,(address))

*#提交插入数据*

connection.commit()

finally:

*#关闭资源 的第二个参数可以将sql缺省语句补全，一般以元组的格式*

connection.close()

return item

4.settings.py这里要说一下，我是根据知乎一大哥学的python爬虫[点击打开链接](https://zhuanlan.zhihu.com/p/26980300)，一年前这个网站还没有反爬虫他直接获取txt文件，但是我写的时候，发现报了503，很尴尬。所以试着自己写的，最后成功了，这里在settings.py中，模拟user-agent信息

*# -\*- coding: utf-8 -\*-*

import random

*# user agent 列表*

USER\_AGENT\_LIST = [

'MSIE (MSIE 6.0; X11; Linux; i686) Opera 7.23',

'Opera/9.20 (Macintosh; Intel Mac OS X; U; en)',

'Opera/9.0 (Macintosh; PPC Mac OS X; U; en)',

'iTunes/9.0.3 (Macintosh; U; Intel Mac OS X 10\_6\_2; en-ca)',

'Mozilla/4.76 [en\_jp] (X11; U; SunOS 5.8 sun4u)',

'iTunes/4.2 (Macintosh; U; PPC Mac OS X 10.2)',

'Mozilla/5.0 (Macintosh; Intel Mac OS X 10.6; rv:5.0) Gecko/20100101 Firefox/5.0',

'Mozilla/5.0 (Macintosh; Intel Mac OS X 10.6; rv:9.0) Gecko/20100101 Firefox/9.0',

'Mozilla/5.0 (Macintosh; Intel Mac OS X 10.8; rv:16.0) Gecko/20120813 Firefox/16.0',

'Mozilla/4.77 [en] (X11; I; IRIX;64 6.5 IP30)',

'Mozilla/4.8 [en] (X11; U; SunOS; 5.7 sun4u)'

]

*# Scrapy settings for proxy project*

*#*

*# For simplicity, this file contains only settings considered important or*

*# commonly used. You can find more settings consulting the documentation:*

*#*

*# https://doc.scrapy.org/en/latest/topics/settings.html*

*# https://doc.scrapy.org/en/latest/topics/downloader-middleware.html*

*# https://doc.scrapy.org/en/latest/topics/spider-middleware.html*

BOT\_NAME = 'proxy'

SPIDER\_MODULES = ['proxy.spiders']

NEWSPIDER\_MODULE = 'proxy.spiders'

DOWNLOAD\_DELAY = 1

*# 随机生成user agent*

USER\_AGENT = random.choice(USER\_AGENT\_LIST)

ITEM\_PIPELINES ={'proxy.pipelines.Rmysql': 500}

*# Crawl responsibly by identifying yourself (and your website) on the user-agent*

*#USER\_AGENT = 'proxy (+http://www.yourdomain.com)'*

*# Obey robots.txt rules*

ROBOTSTXT\_OBEY = True

*# Configure maximum concurrent requests performed by Scrapy (default: 16)*

*#CONCURRENT\_REQUESTS = 32*

*# Configure a delay for requests for the same website (default: 0)*

*# See https://doc.scrapy.org/en/latest/topics/settings.html#download-delay*

*# See also autothrottle settings and docs*

*#DOWNLOAD\_DELAY = 3*

*# The download delay setting will honor only one of:*

*#CONCURRENT\_REQUESTS\_PER\_DOMAIN = 16*

*#CONCURRENT\_REQUESTS\_PER\_IP = 16*

*# Disable cookies (enabled by default)*

*#COOKIES\_ENABLED = False*

*# Disable Telnet Console (enabled by default)*

*#TELNETCONSOLE\_ENABLED = False*

*# Override the default request headers:*

*#DEFAULT\_REQUEST\_HEADERS = {*

*# 'Accept': 'text/html,application/xhtml+xml,application/xml;q=0.9,\*/\*;q=0.8',*

*# 'Accept-Language': 'en',*

*#}*

*# Enable or disable spider middlewares*

*# See https://doc.scrapy.org/en/latest/topics/spider-middleware.html*

*#SPIDER\_MIDDLEWARES = {*

*# 'proxy.middlewares.ProxySpiderMiddleware': 543,*

*#}*

*# Enable or disable downloader middlewares*

*# See https://doc.scrapy.org/en/latest/topics/downloader-middleware.html*

*#DOWNLOADER\_MIDDLEWARES = {*

*# 'proxy.middlewares.ProxyDownloaderMiddleware': 543,*

*#}*

*# Enable or disable extensions*

*# See https://doc.scrapy.org/en/latest/topics/extensions.html*

*#EXTENSIONS = {*

*# 'scrapy.extensions.telnet.TelnetConsole': None,*

*#}*

*# Configure item pipelines*

*# See https://doc.scrapy.org/en/latest/topics/item-pipeline.html*

*#ITEM\_PIPELINES = {*

*# 'proxy.pipelines.ProxyPipeline': 300,*

*#}*

*# Enable and configure the AutoThrottle extension (disabled by default)*

*# See https://doc.scrapy.org/en/latest/topics/autothrottle.html*

*#AUTOTHROTTLE\_ENABLED = True*

*# The initial download delay*

*#AUTOTHROTTLE\_START\_DELAY = 5*

*# The maximum download delay to be set in case of high latencies*

*#AUTOTHROTTLE\_MAX\_DELAY = 60*

*# The average number of requests Scrapy should be sending in parallel to*

*# each remote server*

*#AUTOTHROTTLE\_TARGET\_CONCURRENCY = 1.0*

*# Enable showing throttling stats for every response received:*

*#AUTOTHROTTLE\_DEBUG = False*

*# Enable and configure HTTP caching (disabled by default)*

*# See https://doc.scrapy.org/en/latest/topics/downloader-middleware.html#httpcache-middleware-settings*

*#HTTPCACHE\_ENABLED = True*

*#HTTPCACHE\_EXPIRATION\_SECS = 0*

*#HTTPCACHE\_DIR = 'httpcache'*

*#HTTPCACHE\_IGNORE\_HTTP\_CODES = []*

*#HTTPCACHE\_STORAGE = 'scrapy.extensions.httpcache.FilesystemCacheStorage'*