

# Day09回顾

## settings.py常用变量

```
1  【1】 settings.py中常用变量
2      2.1) 设置日志级别
3          LOG_LEVEL = ''
4      2.2) 保存到日志文件(不在终端输出)
5          LOG_FILE = ''
6      2.3) 设置数据导出编码(主要针对于json文件)
7          FEED_EXPORT_ENCODING = 'utf-8'
8      2.4) 设置User-Agent
9          USER_AGENT = ''
10     2.5) 设置最大并发数(默认为16)
11         CONCURRENT_REQUESTS = 32
12     2.6) 下载延迟时间(每隔多长时间请求一个网页)
13         DOWNLOAD_DELAY = 1
14     2.7) 请求头
15         DEFAULT_REQUEST_HEADERS = {'User-Agent': 'Mozilla/'}
16     2.8) 添加项目管道
17         ITEM_PIPELINES = {'项目目录名.pipelines.类名': 优先级}
18     2.9) cookie(默认禁用,取消注释-True|False都为开启)
19         COOKIES_ENABLED = False
20     2.10) 非结构化数据存储路径
21         IMAGES_STORE = '/home/tarena/images/'
22         FILES_STORE = '/home/tarena/files/'
23     2.11) 添加下载器中间件
24         DOWNLOADER_MIDDLEWARES = {'项目名.middlewares.类名': 200}
```

## 非结构化数据抓取

```
1  【1】 spider
2      yield item['链接']
3
4  【2】 pipelines.py
5      from scrapy.pipelines.images import ImagesPipeline
6      import scrapy
7      class TestPipeline(ImagesPipeline):
8          def get_media_requests(self, item, info):
9              yield scrapy.Request(url=item['url'], meta={'name': item['name']})
10
11          def file_path(self, request, response=None, info=None):
12              name = request.meta['name']
```

```
13         filename = name
14         return filename
15
16 【3】 settings.py
17 IMAGES_STORE = 'D:/Spider/images'
```

## Post请求

### ▪ 方法

```
1 scrapy.FormRequest(url=url, formdata=formdata, callback=self.xxx)
```

### ▪ 使用cookie

```
1 【1】 方法1
2     COOKIES_ENABLED = False
3     DEFAULT_REQUEST_HEADERS = {'Cookie': 'xxx'}
4
5 【2】 方法2
6     COOKIES_ENABLED = True
7     yield scrapy.Request(url=url, cookies={}, callback=self.xxx)
8     yield scrapy.Request(url=url, formdata={}, cookies={}, callback=self.xxx)
9
10 【3】 方法3
11     COOKIES_ENABLED = True
12     class XxxCookieDownloaderMiddleware(object):
13         def process_request(self, request, spider):
14             request.cookies = {}
```

# Day10笔记

## scrapy shell的使用

### ▪ 定义+使用

```
1 【1】 定义
2     1.1) 调试蜘蛛的工具
3     1.2) 交互式shell, 可在不运行spider的前提下, 快速调试 scrapy 代码(主要测试xpath表达式)
4
5 【2】 基本使用
6     scrapy shell URL地址
7
8 【3】 请求对象request属性
9     3.1) request.url : 请求URL地址
10    3.2) request.headers : 请求头 - 字典
11    3.3) request.meta : item数据传递、定义代理
```

```

12
13 【4】 响应对象response属性
14     4.1) response.url      : 返回实际数据的URL地址
15     4.2) response.text    : 响应内容 - 字符串
16     4.3) response.body    : 响应内容 - 字节串
17     4.4) response.encoding : 响应字符编码
18     4.5) response.status   : HTTP响应码

```

#### ▪ scrapy.Request()参数

```

1  【1】 url
2  【2】 callback
3  【3】 headers
4  【4】 meta : 传递数据,定义代理
5  【5】 dont_filter : 是否忽略域组限制,默认False,检查allowed_domains['']
6  【6】 cookies

```

## 设置中间件(随机User-Agent)

#### ▪ 少量UA设置 - 不使用中间件

```

1  【1】 方法一 : settings.py
2      1.1) USER_AGENT = ''
3      1.2) DEFAULT_REQUEST_HEADERS = {}
4
5  【2】 方法二 : 爬虫文件
6      yield scrapy.Request(url,callback=函数名,headers={})

```

#### ▪ 大量UA设置 - 使用middlewares.py中间件

```

1  【1】 获取User-Agent方式
2      1.1) 方法1 : 新建useragents.py,存放大量User-Agent, random模块随机切换
3      1.2) 方法2 : 使用fake_useragent模块
4          from fake_useragent import UserAgent
5          agent = UserAgent().random
6
7  【2】 middlewares.py新建中间件类
8      class RandomUseragentMiddleware(object):
9          def process_request(self, request, spider):
10             agent = UserAgent().random
11             request.headers['User-Agent'] = agent
12
13  【3】 settings.py添加此下载器中间件
14      DOWNLOADER_MIDDLEWARES = {'' : 优先级}

```

## 设置中间件(随机代理)

#### ▪ 代理IP中间件

```

1 class RandomProxyDownloaderMiddleware(object):
2     def process_request(self,request,spider):
3         request.meta['proxy'] = xxx
4
5         # 捕获异常的方法,一旦代理不能用,会被此方法捕获,并重新包装请求再次发送
6     def process_exception(self,request,exception,spider):
7         return request

```

## 设置中间件(Cookie)

### Cookie中间件

```

1 class BaiduCookieDownloaderMiddleware(object):
2     def process_request(self,request,spider):
3         cookies = self.get_cookies()
4         print('middleware3', cookies)
5         # 利用请求对象request的cookies属性
6         request.cookies = cookies
7
8
9     def get_cookies(self):
10        costr = ''
11        cookies = {}
12        for kv in costr.split('; '):
13            cookies[kv.split('=')[0]] =kv.split('=')[1]
14
15        return cookies

```

### 练习

1 将有道翻译案例的cookie使用中间件的方式来实现

## 分布式爬虫

### 分布式爬虫介绍

```

1 【1】原理
2     多台主机共享1个爬取队列
3
4 【2】实现
5     2.1) 重写scrapy调度器(scrapy_redis模块)
6     2.2) sudo pip3 install scrapy_redis

```

### 为什么使用redis

```

1 【1】 Redis基于内存,速度快
2 【2】 Redis非关系型数据库,Redis中集合,存储每个request的指纹

```

# scrapy\_redis详解

## ■ GitHub地址

```
1 | https://github.com/rmax/scrapy-redis
```

## ■ settings.py说明

```
1 | # 重新指定调度器：启用Redis调度存储请求队列
2 | SCHEDULER = "scrapy_redis.scheduler.Scheduler"
3 |
4 | # 重新指定去重机制：确保所有的爬虫通过Redis去重
5 | DUPEFILTER_CLASS = "scrapy_redis.dupefilter.RFPDupeFilter"
6 |
7 | # 不清除Redis队列：暂停/恢复/断点续爬(默认清除为False,设置为True不清除)
8 | SCHEDULER_PERSIST = True
9 |
10 | # 优先级队列 (默认)
11 | SCHEDULER_QUEUE_CLASS = 'scrapy_redis.queue.PriorityQueue'
12 | #可选用的其它队列
13 | # 先进先出
14 | SCHEDULER_QUEUE_CLASS = 'scrapy_redis.queue.FifoQueue'
15 | # 后进先出
16 | SCHEDULER_QUEUE_CLASS = 'scrapy_redis.queue.LifoQueue'
17 |
18 | # redis管道
19 | ITEM_PIPELINES = {
20 |     'scrapy_redis.pipelines.RedisPipeline': 300
21 | }
22 |
23 | #指定连接到redis时使用的端口和地址
24 | REDIS_HOST = 'localhost'
25 | REDIS_PORT = 6379
```

# 腾讯招聘分布式改写

## ■ 分布式爬虫完成步骤

```
1 | 【1】首先完成非分布式scrapy爬虫：正常scrapy爬虫项目抓取
2 | 【2】设置,部署成为分布式爬虫
```

## ■ 分布式环境说明

```
1 | 【1】分布式爬虫服务器数量：2 (其中1台Windows,1台Ubuntu虚拟机)
2 | 【2】服务器分工：
3 |     2.1) Windows：负责数据抓取
4 |     2.2) Ubuntu：负责URL地址统一管理,同时负责数据抓取
```

## ■ 腾讯招聘分布式爬虫 - 数据同时存入1个Redis数据库

```

1  【1】完成正常scrapy项目数据抓取（非分布式 - 拷贝之前的Tencnet）
2
3  【2】设置settings.py, 完成分布式设置
4      2.1-必须) 使用scrapy_redis的调度器
5          SCHEDULER = "scrapy_redis.scheduler.Scheduler"
6
7      2.2-必须) 使用scrapy_redis的去重机制
8          DUPEFILTER_CLASS = "scrapy_redis.dupefilter.RFPDupeFilter"
9
10     2.3-必须) 定义redis主机地址和端口号
11         REDIS_HOST = '192.168.1.107'
12         REDIS_PORT = 6379
13
14     2.4-非必须) 是否清除请求指纹, True:不清除 False:清除（默认）
15         SCHEDULER_PERSIST = True
16
17     2.5-非必须) 在ITEM_PIPELINES中添加redis管道, 数据将会存入redis数据库
18         'scrapy_redis.pipelines.RedisPipeline': 200
19
20 【3】把代码原封不动的拷贝到分布式中的其他爬虫服务器, 同时开始运行爬虫
21
22 【结果】：多台机器同时抓取, 数据会统一存到Ubuntu的redis中, 而且所抓数据不重复

```

#### ■ 腾讯招聘分布式爬虫 - 数据存入MySQL数据库

```

1  """和数据存入redis步骤基本一样, 只是变更一下管道和MySQL数据库服务器的IP地址"""
2  【1】 settings.py
3      1.1) SCHEDULER = 'scrapy_redis.scheduler.Scheduler'
4      1.2) DUPEFILTER_CLASS = 'scrapy_redis.dupefilter.RFPDupeFilter'
5      1.3) SCHEDULER_PERSIST = True
6      1.4) REDIS_HOST = '192.168.1.107'
7      1.5) REDIS_POST = 6379
8      1.6) ITEM_PIPELINES = {'Tencent.pipelines.TencentMysqlPipeline' : 300}
9      1.7) MYSQL_HOST = '192.168.1.101'
10
11 【2】将代码拷贝到分布式中所有爬虫服务器
12
13 【3】多台爬虫服务器同时运行scrapy爬虫
14
15 # 赠送腾讯MySQL数据库建库建表语句
16 """
17 create database tencentdb charset utf8;
18 use tencentdb;
19 create table tencenttab(
20 job_name varchar(1000),
21 job_type varchar(200),
22 job_duty varchar(5000),
23 job_require varchar(5000),
24 job_address varchar(200),
25 job_time varchar(200)
26 )charset=utf8;
27 """

```

## 新的篇章

# 腾讯招聘分布式改写之方法二

### ■ 使用redis\_key改写 (同时存入MySQL数据库)

```
1  【1】 settings.py和方法一中写法一致
2      1.1) SCHEDULER = 'scrapy_redis.scheduler.Scheduler'
3      1.2) DUPEFILTER_CLASS = 'scrapy_redis.dupefilter.RFPDupeFilter'
4      1.3) SCHEDULER_PERSIST = True
5      1.4) REDIS_HOST = '192.168.1.107'
6      1.5) REDIS_POST = 6379
7      1.6) ITEM_PIPELINES = {'Tencent.pipelines.TencentMysqlPipeline' : 300}
8      1.7) MYSQL_HOST = '192.168.1.101'
9
10 【2】 爬虫文件:tencent.py (必须基于start_urls)
11     from scrapy_redis.spiders import RedisSpider
12     class TencentSpider(RedisSpider):
13         # 1. 去掉start_urls
14         # 2. 定义redis_key
15         redis_key = 'tencent:spider'
16         def parse(self, response):
17             pass
18
19 【3】 把代码复制到所有爬虫服务器, 并启动项目
20
21 【4】 到redis命令行, 执行LPUSH命令压入第一个要爬取的URL地址
22     >LPUSH tencent:spider 第1页的URL地址
23
24 【注意】: 项目爬取结束后无法退出, 如何退出?
25 setting.py
26 CLOSESPIDER_TIMEOUT = 3600
27 # 到指定时间(3600秒)时, 会自动结束并退出
```

## 机器视觉与tesseract

### ■ 概述

```
1  【1】作用
2      处理图形验证码
3
4  【2】三个重要概念 - OCR、tesseract-ocr、pytesseract
5      2.1) OCR
6          光学字符识别(Optical Character Recognition),通过扫描等光学输入方式将各种票据、报刊、书
7          籍、文稿及其它印刷品的文字转化为图像信息,再利用文字识别技术将图像信息转化为电子文本
8
9      2.2) tesseract-ocr
10         OCR的一个底层识别库(不是模块,不能导入),由Google维护的开源OCR识别库
11
12     2.3) pytesseract
13         Python模块,可调用底层识别库,是对tesseract-ocr做的一层Python API封装
```

#### ■ 安装tesseract-ocr

```
1  【1】Ubuntu安装
2      sudo apt-get install tesseract-ocr
3
4  【2】Windows安装
5      2.1) 下载安装包
6      2.2) 添加到环境变量(Path)
7
8  【3】测试(终端 | cmd命令行)
9      tesseract xxx.jpg 文件名
```

#### ■ 安装pytesseract

```
1  【1】安装
2      sudo pip3 install pytesseract
3
4  【2】使用示例
5      import pytesseract
6      # Python图片处理库
7      from PIL import Image
8
9      # 创建图片对象
10     img = Image.open('test1.jpg')
11     # 图片转字符串
12     result = pytesseract.image_to_string(img)
13     print(result)
```

## 在线打码平台

#### ■ 为什么使用在线打码

```
1  tesseract-ocr识别率很低,文字变形、干扰,导致无法识别验证码
```

#### ■ 云打码平台使用步骤



- 1 【1】 下载并查看接口文档
- 2 【2】 调整接口文档，调整代码并接入程序测试
- 3 【3】 真正接入程序，在线识别后获取结果并使用

## 破解云打码网站验证码

### ■ 1 - 下载并调整接口文档，封装成函数，打码获取结果

```
1 import http.client, mimetypes, urllib, json, time, requests
2
3 #####
4
5 class YDMHttp:
6
7     apiurl = 'http://api.yundama.com/api.php'
8     username = ''
9     password = ''
10    appid = ''
11    appkey = ''
12
13    def __init__(self, username, password, appid, appkey):
14        self.username = username
15        self.password = password
16        self.appid = str(appid)
17        self.appkey = appkey
18
19    def request(self, fields, files=[]):
20        response = self.post_url(self.apiurl, fields, files)
21        response = json.loads(response)
22        return response
23
24    def balance(self):
25        data = {'method': 'balance', 'username': self.username, 'password': self.password,
26'appid': self.appid, 'appkey': self.appkey}
27        response = self.request(data)
28        if (response):
29            if (response['ret'] and response['ret'] < 0):
30                return response['ret']
31            else:
32                return response['balance']
33        else:
34            return -9001
35
36    def login(self):
37        data = {'method': 'login', 'username': self.username, 'password': self.password,
38'appid': self.appid, 'appkey': self.appkey}
39        response = self.request(data)
40        if (response):
41            if (response['ret'] and response['ret'] < 0):
42                return response['ret']
43            else:
44                return response['uid']
```

```

43         else:
44             return -9001
45
46     def upload(self, filename, codetype, timeout):
47         data = {'method': 'upload', 'username': self.username, 'password': self.password,
48 'appid': self.appid, 'appkey': self.appkey, 'codetype': str(codetype), 'timeout':
49 str(timeout)}
50         file = {'file': filename}
51         response = self.request(data, file)
52         if (response):
53             if (response['ret'] and response['ret'] < 0):
54                 return response['ret']
55             else:
56                 return response['cid']
57         else:
58             return -9001
59
60     def result(self, cid):
61         data = {'method': 'result', 'username': self.username, 'password': self.password,
62 'appid': self.appid, 'appkey': self.appkey, 'cid': str(cid)}
63         response = self.request(data)
64         return response and response['text'] or ''
65
66     def decode(self, filename, codetype, timeout):
67         cid = self.upload(filename, codetype, timeout)
68         if (cid > 0):
69             for i in range(0, timeout):
70                 result = self.result(cid)
71                 if (result != ''):
72                     return cid, result
73                 else:
74                     time.sleep(1)
75             return -3003, ''
76         else:
77             return cid, ''
78
79     def report(self, cid):
80         data = {'method': 'report', 'username': self.username, 'password': self.password,
81 'appid': self.appid, 'appkey': self.appkey, 'cid': str(cid), 'flag': '0'}
82         response = self.request(data)
83         if (response):
84             return response['ret']
85         else:
86             return -9001
87
88     def post_url(self, url, fields, files=[]):
89         for key in files:
90             files[key] = open(files[key], 'rb');
91         res = requests.post(url, files=files, data=fields)
92         return res.text
93
94 #####
95 def get_result(filename):
96     # 用户名
97     username = 'yibeizi001'
98     # 密码
99     password = 'zhishouzhethian001'

```

```

96     # 软件ID, 开发者分成必要参数。登录开发者后台【我的软件】获得!
97     appid      = 1
98     # 软件密钥, 开发者分成必要参数。登录开发者后台【我的软件】获得!
99     appkey     = '22cc5376925e9387a23cf797cb9ba745'
100    # 验证码类型, # 例: 1004表示4位字母数字, 不同类型收费不同。请准确填写, 否则影响识别率。在此查
    询所有类型 http://www.yundama.com/price.html
101    codetype    = 5000
102    # 超时时间, 秒
103    timeout     = 60
104    # 初始化
105    yundama = YDMHttp(username, password, appid, appkey)
106    # 登陆云打码
107    uid = yundama.login()
108    # 查询余额
109    balance = yundama.balance()
110    # 开始识别, 图片路径, 验证码类型ID, 超时时间(秒), 识别结果
111    cid, result = yundama.decode(filename, codetype, timeout)
112
113    return result
114    #####

```

## ■ 2 - 访问云打码网站, 获取验证码并在线识别

```

1  '''识别云打码官网的验证码'''
2  from selenium import webdriver
3  from ydmapi import *
4  from PIL import Image
5
6  class YdmSpider(object):
7      def __init__(self):
8          self.options = webdriver.ChromeOptions()
9          # 浏览器窗口最大化
10         self.options.add_argument('--start-maximized')
11         self.browser = webdriver.Chrome(options=self.options)
12
13         self.browser.get('http://www.yundama.com/')
14
15     # 获取验证码图片截取出来
16     def get_image(self):
17         # 1. 获取页面截图
18         self.browser.save_screenshot('index.png')
19         # 2. 获取验证码节点坐标, 把图片截取出来
20         # location: 获取节点左上角的坐标(x y)
21         location = self.browser.find_element_by_xpath('//*[@id="verifyImg"]').location
22         # size: 获取节点的大小(宽度和高度)
23         size = self.browser.find_element_by_xpath('//*[@id="verifyImg"]').size
24         # 四个坐标
25         left_x = location['x']
26         left_y = location['y']
27         right_x = left_x + size['width']
28         right_y = left_y + size['height']
29         # 从index.png中截图图片, 注意crop()方法参数为元组
30         img = Image.open('index.png').crop((left_x, left_y, right_x, right_y))
31         img.save('verify.png')
32
33     # 获取识别结果

```

```

34     def get_result(self):
35         result = get_result('verify.png')
36         print('识别结果:', result)
37
38     # 入口函数
39     def run(self):
40         self.get_image()
41         self.get_result()
42         self.browser.close()
43
44 if __name__ == '__main__':
45     spider = YdmSpider()
46     spider.run()

```

## Fiddler抓包工具

### ■ 配置Fiddler

```

1  【1】 Tools -> Options -> HTTPS
2      1.1) 添加证书信任: 勾选 Decrypt Https Traffic 后弹出窗口, 一路确认
3      1.2) 设置之抓浏览器的包: ...from browsers only
4
5  【2】 Tools -> Options -> Connections
6      2.1) 设置监听端口 (默认为8888)
7
8  【3】 配置完成后重启Fiddler (重要)
9      3.1) 关闭Fiddler,再打开Fiddler

```

### ■ 配置浏览器代理

```

1  【1】 安装Proxy SwitchyOmega谷歌浏览器插件
2
3  【2】 配置代理
4      2.1) 点击浏览器右上角插件SwitchyOmega -> 选项 -> 新建情景模式 -> myproxy(名字) -> 创建
5      2.2) 输入 HTTP:// 127.0.0.1 8888
6      2.3) 点击 : 应用选项
7
8  【3】 点击右上角SwitchyOmega可切换代理
9
10 【注意】: 一旦切换了自己创建的代理,则必须要打开Fiddler才可以上网

```

### ■ Fiddler常用菜单

```

1  【1】 Inspector : 查看数据包详细内容
2      1.1) 整体分为请求和响应两部分
3
4  【2】 Inspector常用菜单
5      2.1) Headers : 请求头信息
6      2.2) WebForms: POST请求Form表单数据 : <body>
7              GET请求查询参数: <QueryString>
8      2.3) Raw : 将整个请求显示为纯文本

```

# 移动端app数据抓取

## ■ 方法1 - 手机 + Fiddler

1 | 设置方法见文件夹 - 移动端抓包配置

## ■ 方法2 - F12浏览器工具

## 有道翻译手机版破解案例

```
1 import requests
2 from lxml import etree
3
4 word = input('请输入要翻译的单词:')
5
6 post_url = 'http://m.youdao.com/translate'
7 post_data = {
8     'inputtext':word,
9     'type':'AUTO'
10 }
11
12 html = requests.post(url=post_url,data=post_data).text
13 parse_html = etree.HTML(html)
14 xpath_bds = '//ul[@id="translateResult"]/li/text()'
15 result = parse_html.xpath(xpath_bds)[0]
16
17 print(result)
```

## 爬虫总结

```
1 # 1、什么是爬虫
2 爬虫是请求网站并提取数据的自动化程序
3
4 # 2、robots协议是什么
5 爬虫协议或机器人协议,网站通过robots协议告诉搜索引擎哪些页面可以抓取, 哪些页面不能抓取
6
7 # 3、爬虫的基本流程
8 1、请求得到响应
9 2、解析
10 3、保存数据
11
12 # 4、请求
13 1、urllib
14 2、requests
15 3、scrapy
16
17 # 5、解析
18 1、re正则表达式
```

```
19 2、lxml+xpath解析
20 3、json解析模块
21
22 # 6、selenium+browser
23
24 # 7、常见反爬策略
25 1、Headers：最基本的反爬手段，一般被关注的变量是UserAgent和Referer，可以考虑使用浏览器中
26 2、UA：建立User-Agent池，每次访问页面随机切换
27 3、拉黑高频访问IP
28 数据量大用代理IP池伪装成多个访问者，也可控制爬取速度
29 4、Cookies
30 建立有效的cookie池，每次访问随机切换
31 5、验证码
32 验证码数量较少可人工填写
33 图形验证码可使用tesseract识别
34 其他情况只能在线打码、人工打码和训练机器学习模型
35 6、动态生成
36 一般由js动态生成的数据都是向特定的地址发get请求得到的，返回的一般是json
37 7、签名及js加密
38 一般为本地JS加密，查找本地JS文件，分析，或者使用execjs模块执行JS
39 8、js调整页面结构
40 9、js在响应中指向新的地址
41
42 # 8、scrapy框架的运行机制
43
44 # 9、分布式爬虫的原理
45 多台主机共享一个爬取队列
```