动态HTML处理

第三部分课程概要

1. 动态html介绍
2. selenium和phantomjs
3. 机器视觉和tesseract介绍

后续爬虫代码的建议

尽量减少请求次数	1、能抓列表页就不抓详情页 2、保存获取到的html页面，供查错和重复请求使用
关注网站的所有类型的页面	1、wap页面，触屏版页面 2、H5页面 3、APP
多伪装	1、动态的UA 2、代理ip 3、不使用cookie
利用多线程分布式	在不被ban的请求下尽可能的提高速度

动态HTML技术了解

JS	是网络上最常用的脚本语言。它可以收集用户的跟踪数据,不需要重载页面直接提交表单，在页面嵌入多媒体文件，甚至运行网页游戏
jQuery	Query是一个快速、简洁的JavaScript框架，封装了JavaScript常用的功能代码
Ajax	AJAX 可以使网页实现异步更新，可以在不重新加载整个网页的情况下，对网页的某部分进行更新

对搜索引擎不友好
对爬虫也不友好

Selenium和PhantomJS

- Selenium

Selenium是一个Web的自动化测试工具，最初是为网站自动化测试而开发的，Selenium 可以直接运行在浏览器上，它支持所有主流的浏览器（包括PhantomJS这些无界面的浏览器），可以接收指令，让浏览器自动加载页面，获取需要的数据，甚至页面截屏

- PhantomJS

PhantomJS 是一个基于Webkit的“无界面”(headless)浏览器，它会把网站加载到内存并执行页面上的 JavaScript

http://selenium-python-zh.readthedocs.io/en/latest/waits.html

Selenium和PhantomJS入门

1. 加载网页：
   - from selenium import webdriver
   - driver = webdriver.PhantomJS(“c:…/pantomjs.exe”)
   - driver.get("http://www.baidu.com/")
   - driver.save_screenshot("长城.png")
2. 定位和操作：
   - driver.find_element_by_id(“kw”).send_keys(“长城”)
   - driver.find_element_by_id("su").click()
3. 查看请求信息：
   - driver.page_source
   - driver.get_cookies()
   - driver.current_url
4. 退出
   - driver.close() #退出当前页面
   - driver.quit() #退出浏览器

页面元素定位

- 用法：
  - find_element_by_id (返回一个)
  - find_elements_by_xpath （返回一个列表）
  - find_elements_by_link_text
  - find_elements_by_partial_link_text
  - find_elements_by_tag_name
  - find_elements_by_class_name
  - find_elements_by_css_selector
- 注意点:
  1. find_element 和find_elements的区别：返回一个和返回一个列表
  2. by_link_text和by_partial_link_text的区别：全部文本和包含某个文本
  3. by_css_selector的用法： #food span.dairy.aged
  4. by_xpath中获取属性和文本需要使用get_attribute() 和.text

动手练习

1. 模拟登陆豆瓣网

cookie

- Cookie相关用法：
  - {cookie[‘name’]: cookie[‘value’] for cookie in driver.get_cookies()}
  - driver.delete_cookie("CookieName")
  - driver.delete_all_cookies()

页面等待

- 为什么需要等待
  - 如果网站采用了动态html技术，那么页面上的部分元素出现时间便不能确定，这个时候就可以设置一个等待时间，强制要求在时间内出现，否则报错
- 强制等待
  - time.sleep(10)
- 显式等待(了解)
  - 显式等待指定某个条件，然后设置最长等待时间。如果在这个时间还没有找到元素，那么便会抛出异常了。
  - WebDriverWait(driver, 10).until(
    EC.presence_of_element_located((By.ID, "myDynamicElement"))
- 隐式等待(了解)
  - 就是简单地设置一个最大等待时间，单位为秒。
  - driver.implicitly_wait(10)

phantomjs安装指南

#官网下载，apt-get可能会报错

--解压文件

tar -xvf phantomjs-1.9.7-linux-x86_64.tar.bz2

-将程序移到一个合适的位置

sudo mv phantomjs-1.9.7-linux-x86_64 /usr/local/src/phantomjs

--创建软链接到环境变量中。这样可以直接在shell中使用phantomjs命令

sudo ln -sf /usr/local/src/phantojs/bin/phantomjs /usr/local/bin/phantomjs

--检查是否正常工作

phantomjs --version

chromedirver下载地址:https://npm.taobao.org/mirrors/chromedriver

phantomjs下载地址:http://phantomjs.org/download.html

动手练习

1. 爬取斗鱼直播平台的所有房间信息
   https://www.douyu.com/directory/all

Selenium总结

一．应用场景：
1. 动态html页面请求
2. 登录获取cookies

二．如何使用
1. 导包并且实例化driver
2. 发送请求
3. 定位获取数据
4. 保存
5. 退出driver

三．Cookies相关方法：
- get_cookies()

四．页面等待
- 强制等待

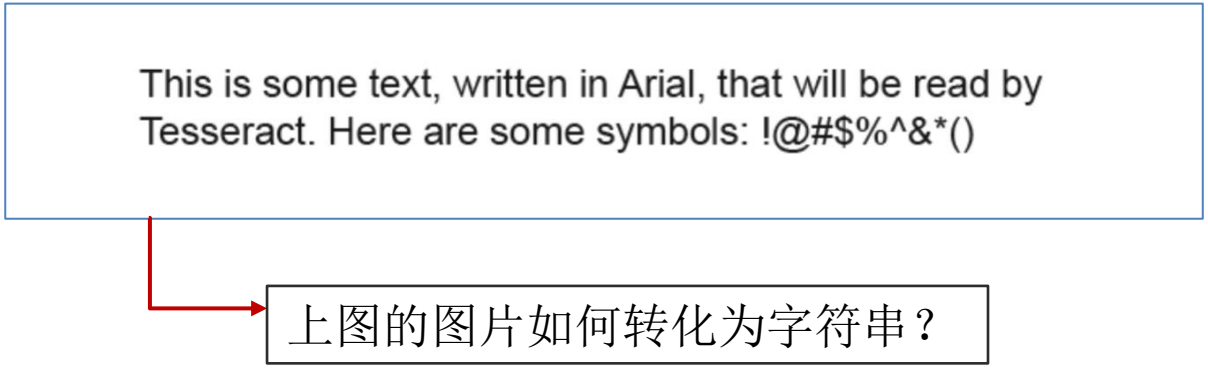
Tesseract

1. 定义：
   - Tesseract是一个将图像翻译成文字的OCR库(光学文字识别，Optical Character Recognition)

2. 安装：
   - sudo apt-get install tesseract-ocr

3. 在python中调用Tesseract
   - pip install pytesseract

Tesseract处理规范的文字



- 在终端中：
  tesseract test.jpg text

- 在python代码中
  import pytesseract
  from PIL import Image
  image = Image.open(jpg)
  pytesseract.image_to_string(image)

Thank You!

改变中国 IT 教育，我们正在行动