百度精确搜索的南京大学网页链接正确性和多浏览器兼容性测试脚本文档

DZ1833025 王子成

目录

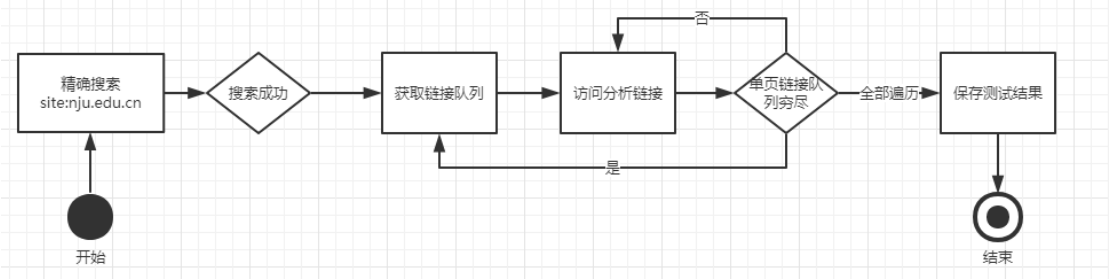
一 前言

通过百度搜索引擎的精确搜索功能（在搜索框中输入 site:nju.edu.cn 即可精确搜索南京大学的相关链接）获取南京大学的全部链接，使用 Chrome 和 Firefox 两种浏览器对这些链接进行访问测试，通过检测链接指向网页的 title 信息，链接地址和加载时间，分析链接信息是否准确，链接地址是否属于南京大学顶级域名，链接指向网站是否能够顺利加载，以及链接指向网页是否使用 https 加密增强安全性。收集以上全部信息，对比两个浏览器的相应结果，检测南京大学链接指向网页的兼容性。
由于链接数目较大，因此使用 python 语言和 selenium 包编写测试脚本，自动化访问搜索页面收集的全部链接，并分析收集到的相关数据。

二 项目介绍

2.1 脚本执行流程

使用 webdriver 打开 https://baidu.com，搜索 “site:nju.edu.cn” 获取有关南京大学的全部热门链接（经测试，百度搜索引擎的爬虫共获取超过一百万个链接，精确搜索仅提供热门链接，两浏览器均能获取 76 页共 757 个南京大学相关热门链接。获取单独页面中的全部链接，分别使用 webdriver 模拟自动化操作点击每个链接，若单页链接被全部遍历则自动化点击下一页重复此过程，点击链接并访问指向网页后，对该网页的相关数据进行分析，将分析结果保存到本地环境中。流程图如图所示。



2.2 脚本文件介绍

Test.py 脚本功能：自动化收集，分析测试获取的全部链接并生成结果文件。
nextpage(driver)：当搜索结果页面的链接被全部分析之后，调用 webdriver 进行翻页操作。
initdriver(browser) ：根据浏览器的类型创建 selenium webdriver，打开百度搜索页面，等待用户登录，登陆成功后输入 “site:nju.edu.cn” 进行精确搜索，返回 webdriver。
trim() 链接 title 可能存在空格等干扰因素，去除这些干扰因素，返回被清洗过的数据。
operate(driver,browser)：根据 driver 和 browser 自动化测试分析获取的全部链接数据，返

回结果列表。
writeTocsv(errorlist, browser)：将结果列表按照 csv 格式写入文件。

Analysis.py 脚本功能 :读取 test.py 脚本生成的数据，对数据进行分析并绘制图像表征数据。

2.3 数据格式及含义

Title	链接指向网页的 title 信息
url	链接地址
WARNING0	链接指向的网页未使用 HTTPS 进行加密
ERROR0	链接指向网页的信息与搜索结果页面链接信息不符
ERROR1	链接指向地址与搜索结果页面链接地址不符
ERROR2	链接指向网页加载发生异常

针对每个链接进行上述问题分析，当链接及链接指向网页不存在问题时则只记录链接 title 和链接地址，一旦出现任何问题则以 csv 格式进行数据记录。

三 项目的部署

1. 从互联网获取最新稳定开发版本的 Chrome 和 Firefox 浏览器，并安装在测试环境中。如测试环境已有浏览器，请检测更新。
2. 从 selenium 项目的官方网站下载对应浏览器的 webdriver 驱动，下载后解压在项目目录下，并记录在环境中的绝对地址。
3. 下载并安装 python3 及 python3-pip，试用 pip 工具下载 selenium 包并安装在测试环境中。
4. 将 test.py 和 analysis.py 脚本部署在相同项目目录下，打开 test.py，对 initdriver 函数中的两处 webdriver 地址进行更改，更改为上一步记录的项目目录地址。

```
def initdriver(browser):
    if browser == 'firefox':
        driver = webdriver.Firefox(executable_path='C:/Program Files/Mozilla Firefox/geckodriver.exe')
    else:
        driver = webdriver.Chrome('C:/Program Files (x86)/Google/Chrome/Application/chromedriver.exe')
```

5. 修改过后可以开始执行。

四 项目执行

1. 执行脚本 test.py，当听到测试环境的蜂鸣器发出蜂鸣声时，可以看到百度网页已经被打

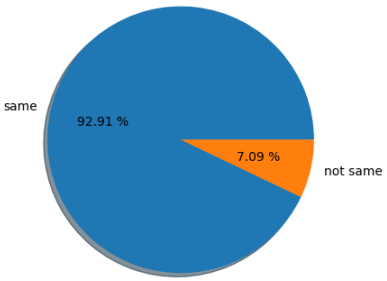
开，请在两分钟内点击登录按钮，并使用手机百度 APP 进行登录，两分钟计时结束开始自动化测试脚本。

2. 当第二次听到蜂鸣器发出蜂鸣声时，第一个测试浏览器的全部链接已经执行完成，开始第二个浏览器的测试流程，请在听到蜂鸣声后在两分钟内使用手机百度 APP 登录第二个浏览器，两分钟结束后，开始自动化测试收集到的全部链接。
3. 当全部测试完成后可以在项目目录中收集到 error-chrome.csv 和 error-firefox.csv 两个文件。
4. 执行 analysis.py 可以对两个文件进行分析，分析结果将在控制台中被显示，同时会弹出已经保存过的图片窗口。所有图片均被保存到项目目录下。

项目执行的相关视频地址 https://youtu.be/qyykhh-YcWE ，直接点击即可访问查看脚本执行情况。

五 结果分析

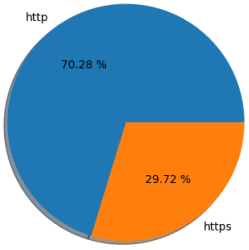
Chrome 和 Firefox 两浏览器分别分析 757 个最热门链接（百度搜索引擎爬虫仅提供 757 个热门链接），然而两个浏览器结果存在差异，经过统计一共收集 783 个网页，其中搜索结果一致的共有 635 个链接，其余均存在一定差异。两个浏览器收集到的 635 个相同链接中有 45 个存在差异，差一点均为 ERROR0 结果不一致，深究原因可能有 1）浏览器对相同代码信息渲染结果存在差异，使得 selenium 包获得的结果产生不同。2）网络状态随时间变化，难以控制，网络状态不同导致加载时间存在差异。如图所示：

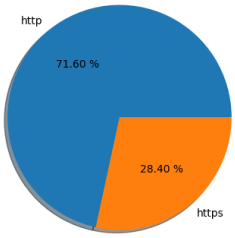


下面分别进行讨论：

5.1 Chrome 浏览器和 Firefox 浏览器 https 使用情况

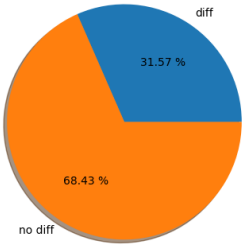
二者类似，仅有约 30%的网站使用 https 进行保护。

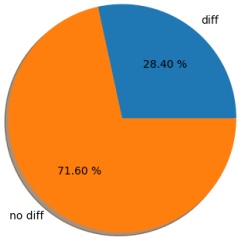




5.2 Chrome 浏览器和 Firefox 浏览器 title 信息与搜索结果页面提供信息不符情况

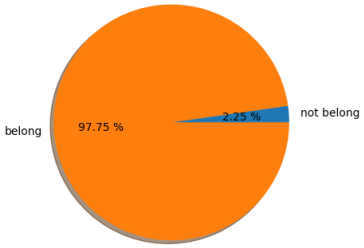
均有 30%左右的数据与搜索结果页面不符，可能是百度搜索引擎的爬虫更新信息不及时。

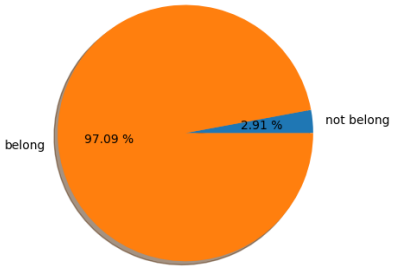




5.3 Chrome 浏览器和 Firefox 浏览器搜索引擎获取的链接不属于 nju.edu.cn 顶级域名的情况

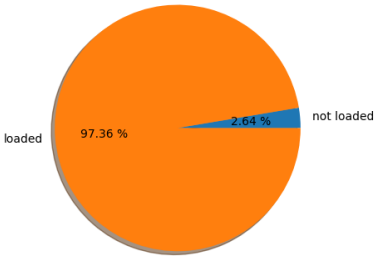
约有 2%-3%的网页链接不属于 nju.edu.cn，原因可能是加载速度太慢，脚本测试分析到的是百度的链接重定向页面。

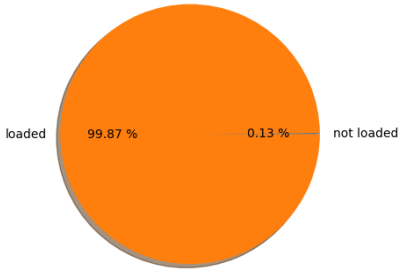




5.4 Chrome 浏览器和 Firefox 浏览器链接指向网页无法加载的情况

Chrome 浏览器的无法加载情况大于 Firefox 浏览器无法加载情况，存在该版本的 Firefox 浏览器比 Chrome 浏览器兼容性更好的可能。





六 未来工作

1. 本次项目收集到的全部链接来自百度搜索引擎提供的精确搜索结果，然而该结果仅包含 76 页 757 个可测试链接，根据以往经验者 757 个可测试链接通常表示在 nju.edu.cn 域名下最热门的 757 个链接。根据百度搜索引擎精确搜索的统计，南京大学的机构和部门中包含的链接超过一百万个，因此仅对 757 个链接进行测试远远不够，未来我们可以编写脚本通过递归迭代的方式获取全部南京大学域名下的全部链接，同时，可以用更多时间使用更多类别的浏览器进行更加全面的链接准确性测试。
2. 由于相同计算机下的 selenium 不能并行运行，因此我们可以借鉴分布式思维，同时使用多台机器对链接进行搜集测试，增强脚本测试的效率。

七 总结

本次实验的脚本的提交时间较为仓促，然而我还是尽最大的可能完成了全部任务，并且编写的脚本具有较强的健壮性，能够应对复杂的网络状态变化，成功自动化测试分析，并保存结果。同时我还利用了电脑的蜂鸣器提示用户需要登录，具有较好的人机交互性。最后收集到的数据使用 csv 格式进行保存，csv 格式的文件不仅能够被编程语言快速处理成二维数组，同时能够被 Excel 一类表格编辑软件读取，便于被更广泛的利用。

南京大学机构和部门的网页作为向公众开放的窗口基本能做到全面覆盖，然而安全意识较为淡薄，不仅绝大多数链接没有使用 https 进行保护，而且部分财务和缴费网页可以被外网访问的同时也没有使用 https 保护，一旦攻击者利用安全漏洞进行攻击，后果将非常严重。此外，很多网页的 title 与百度搜索引擎结果页面提供的信息严重不符，其中很有可能是百度搜索引擎爬虫更新数据不及时或 CDN 数据更新不及时，不过网站的相关工作人员在对网页进行较大改动的时候也应该主动向搜索引擎提交变更信息，便于信息及时更新。

最后，部分南京大学的机构或部门网页的 title 仅仅写了部门的名称，甚至部门的简称，没有包括南京大学这一关键信息，很容易造成误导，希望学校的相关部门能够提升对学校信息安全建设的重视程度。