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

DZ1833026 王子成

outline

- 目标
- 实现细节
- 实验结果
- 总结

目标

- 通过 baidu.com 精度搜索 site: nju.edu.cn获取南京大学的全部网站
- 利用 selenium 包和 chrome 浏览器和Firefox浏览器进行链接准确性测试和兼容性测试
- 编写 python 脚本，自动化打开、分析、并关闭每个网页，查看网页是否能在规定时间内打开， title数据与搜索数据是否相关，链接地址是否正确
- 统计出错链接的数量和原因，输出分析结果

实现细节

- 1.网页加载时间控制
- 基于 selenium 包的隐式 等待 driver.implicitly_wait(30)
- time.sleep(5)以防万一，进行整体等待其他脚本完成

```
driver.implicitly_wait(30)
time.sleep(5)
```

```
ions.StaleElementReferenceException: Message: stale element reference: element is
e=71.0.3578.98)
driver=2.45.615291 (ec3682e3c9061c10f26ea9e5cdcf3c53f3f74387),platform=Windows NT
```

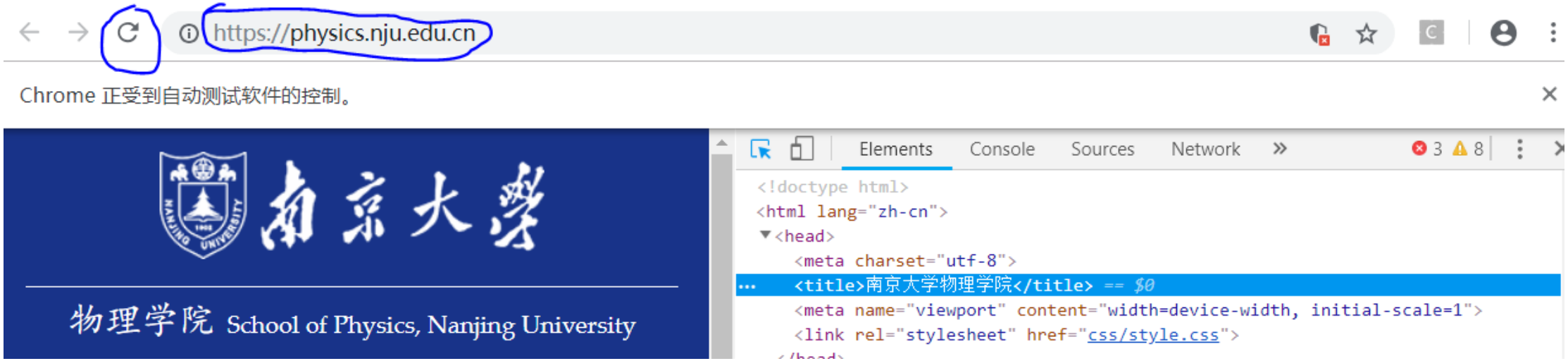
实现细节

- 2.选项卡控制
- 共2个 handles,分别是： 百度精确搜索 handle, 南京大学 handle
- 当分析网页时， driver 选择南京大学网页 handle
- 当选择网页时， driver选择百度搜索handle

实现细节

- 3.链接正确性分析
- Title
- 链接地址
- 网页加载

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



实现细节

- 4.多次实验进行结果对比
- 5.当百度搜索页面链接穷尽自动点击下一页

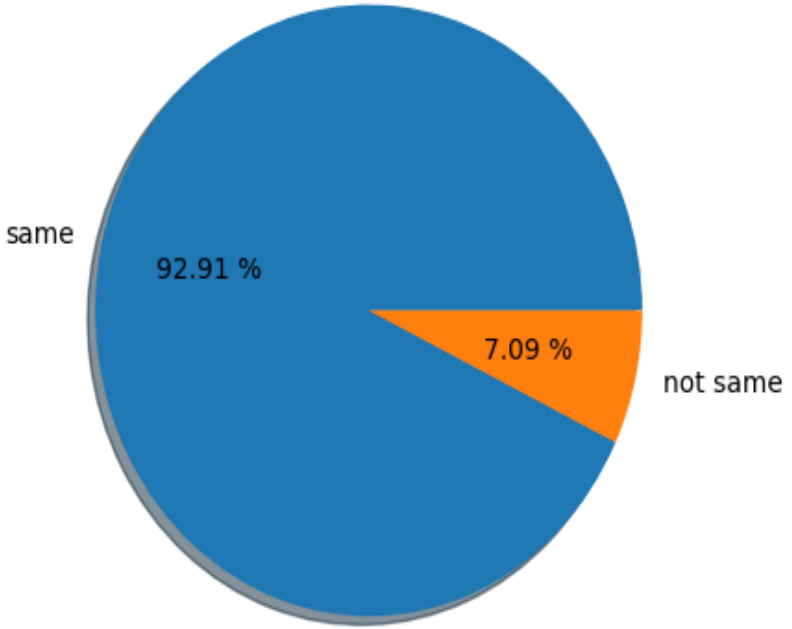

```
def nextpage():
    driver.implicitly_wait(30)
    ele = driver.find_elements_by_class_name('n')
    # ele = driver.find_elements_by_link_text('下一页&gt;')
    if len(ele) == 2:
        driver.implicitly_wait(30)
        time.sleep(2)
        ele[1].click()
    else:
        driver.implicitly_wait(30)
        time.sleep(2)
        ele[0].click()
```

实验结果

- 1.每10 s 分析一个网页
- 2.共分析10页搜索页面， title 不一致，加载问题，其中“南海研究中心”被错误指向百度搜索“南京大学物理系很强吗？”
- 百度搜索页面搜索量较大时需要登录
- Selenium driver加载的过程中会被误识别为病毒威胁。

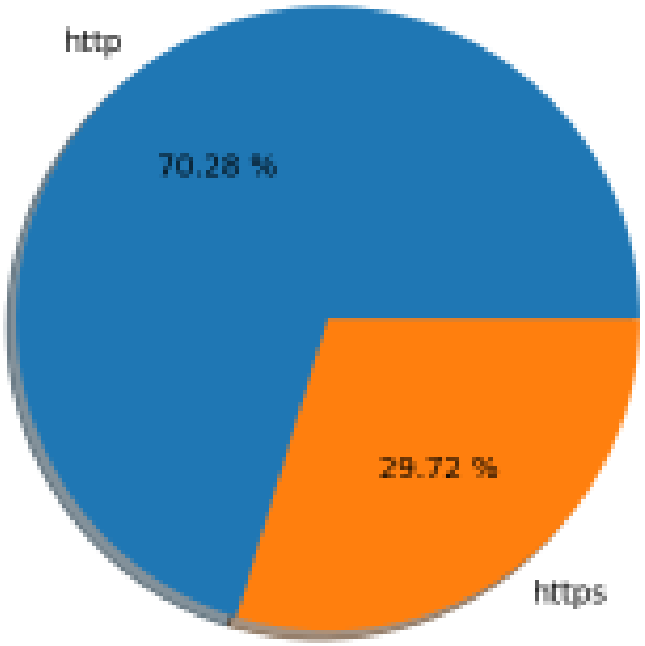
实验结果

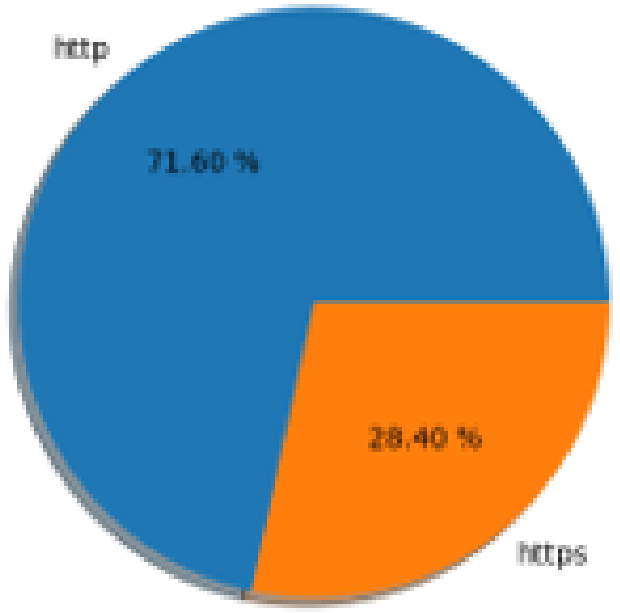
- Chrome和Firefox两浏览器分别分析757个最热门链接（百度搜索引擎爬虫仅提供757个热门链接），然而两个浏览器结果存在差异，经过统计一共收集783个网页，其中搜索结果一致的共有635个链接。相同链接中有45个存在差异
- 深究原因可能有1）浏览器对相同代码信息渲染结果存在差异，2）网络状态随时间变化，难以控制



Chrome浏览器和Firefox浏览器https使用情况

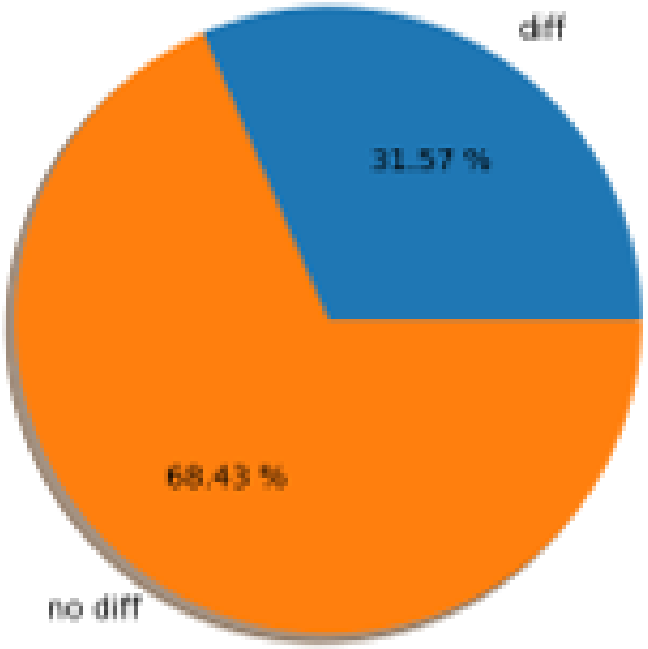
- 二者类似，仅有约30%的网站使用https进行保护。

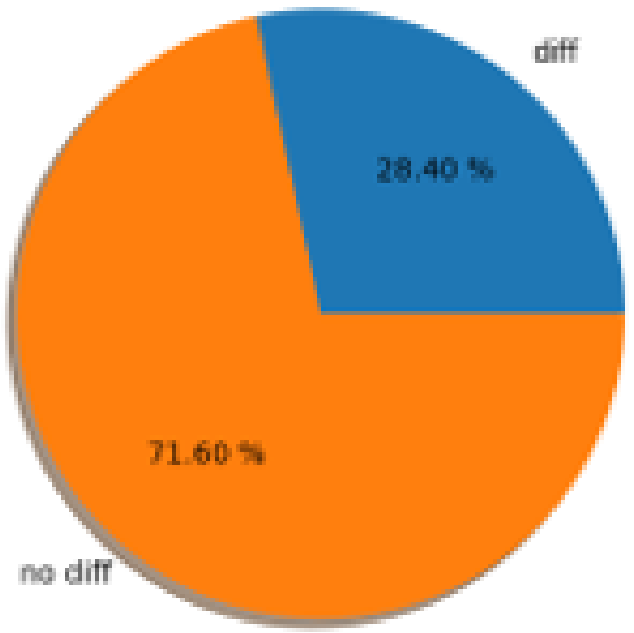




Chrome浏览器和Firefox浏览器title信息与搜索结果页面提供信息不符情况

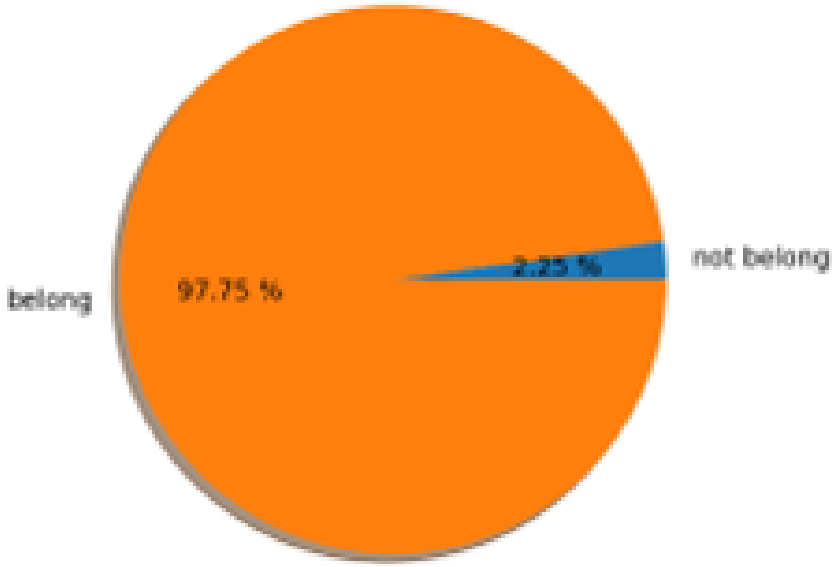
- 均有30%左右的数据与搜索结果页面不符，可能是百度搜索引擎的爬虫更新信息不及时。

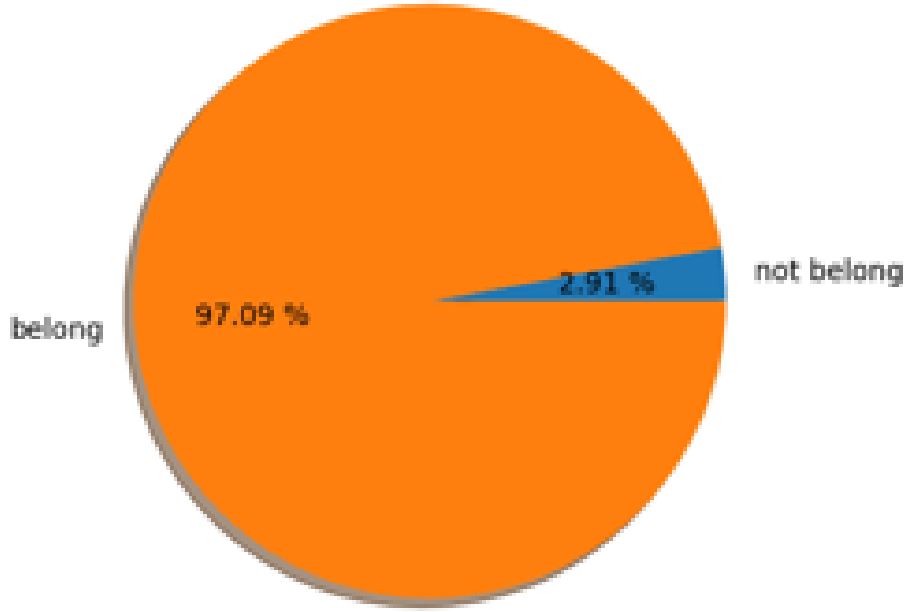




Chrome浏览器和Firefox浏览器搜索引擎获取的链接不属于nju.edu.cn顶级域名的情况

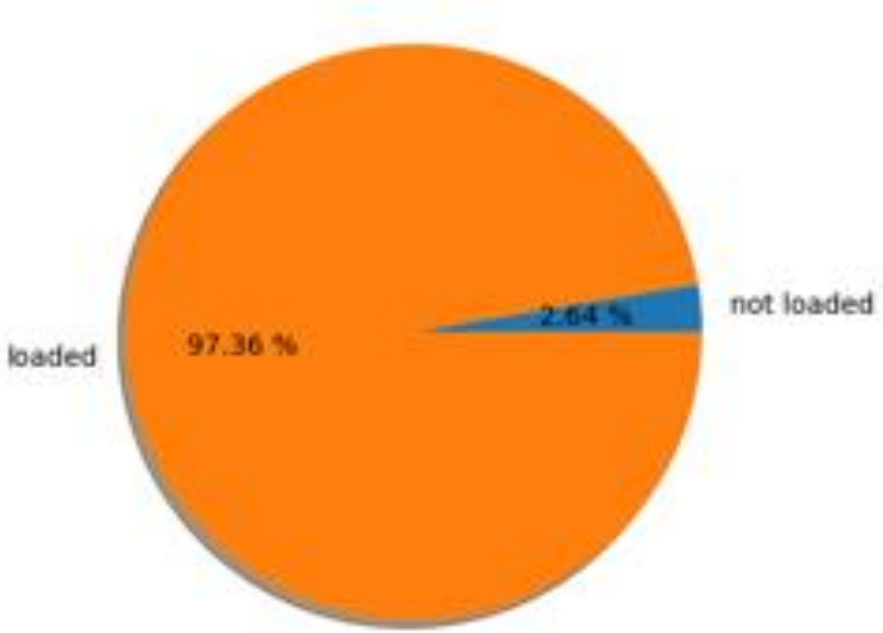
- 约有2%-3%的网页链接不属于nju.edu.cn，原因可能是加载速度太慢，脚本测试分析到的是百度的链接重定向页面。

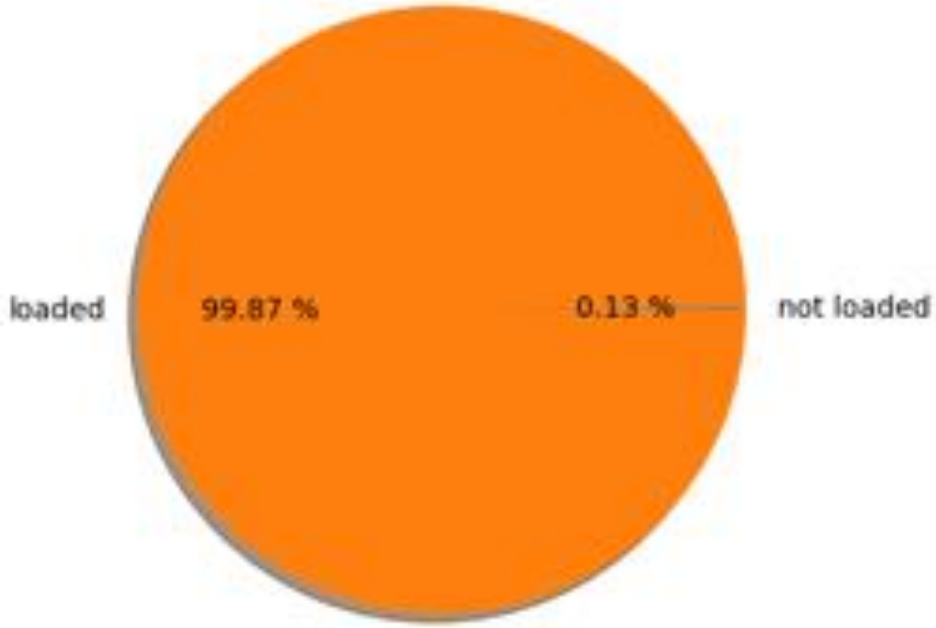




Chrome浏览器和Firefox浏览器链接指向网页无法加载的情况

- 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。





总结

- 1.南京大学部分机构和部门的网站负责人员安全意识较为淡薄，约70%的网页都没有使用https加密，有很高的安全风险，建议统一设置相同证书，一方面可以增强安全性，另一方面也可以防止三级域名被劫持伪造。甚至缴费系统的网站也没有使用https，例如：南大缴费系统,http://net.nju.edu.cn/,WARNING0,,
- 2.网页优化情况存在差异，需要设置等待时间
- 3.网页情况复杂，需要设置异常捕获防止脚本被异常终止
- 4.部分网页，例如vpn2.nju.edu.cn等网站不对外网用户开放（测试环境为外网服务器），但被百度爬虫获取，因此检测被判定为无法加载。

Thank you