哈尔滨工业大学

<<信息检索>>

实验报告

(2023 年度春季学期)

姓名：	李浩桢
学号：	7203610321
学院：	计算机学院
教师：	张宇

实验一 网页文本的预处理

一、实验目的

本次实验目的是对信息检索中网页文本预处理的流程和涉及的技术有一个全面的了解，包括抓去网页、网页正文提取、分词处理、停用词处理等环节。本次实验所要用到的知识如下：

1.基本编程能力（文件处理、网页爬取等）

2.分词、停用词处理

二、实验内容

1. 网页的抓取和正文提取

通过爬虫工具爬取网页（至少 1000 个，其中包含附件的网页不少于 100 个，多线程实现爬虫可加分），然后提取网页标题和网页正文，以及网页中的附件并保存附件到本地，然后将附件名称记录在 file_name 字段中，附件必须是文本文档（txt、doc、docx、xlsx 等）而不能是图片。网页正文和网页标题可以自行定义，但一般应该是网页中你最关注的内容。最后将爬取下来的数据保存为 json 格式。

2. 分词处理、去停用词处理

将提取的网页文本进行分词和去停用词处理，并将结果保存。分词工具使用由我校社会计算与信息检索研究中心开发的语言技术平台-LTP。停用词表采用由我校社会计算与信息检索研究中心发布的停用词表(stop_words.txt)。如果该页面存在文档，则将文档下载并将文档名称保存在 file_name 字段中。

三、实验过程及结果

1. 网页的选择与抓取：

选择“jwc.hit.edu.cn”教务处网站作为我们爬虫的起点网站，并限制爬虫程序只爬取此网站域名下的网页。首先我们设置 Http 请求报文的部分字段，将 Cookie

字段，User-Agent 字段，Connection 字段加入。接着从起始网址出发，初步筛选（将非 jwc.hit 域下的 url 筛去，将不允许爬取的 url 筛去）并爬取到 1100 个教务处网站下的网页 url，并将其存入线程安全的 Queue 中，方便后续爬取网页内容和下载网页附件。接着我们需要根据网页内容判断该网页是否合法，即：该网页是否有标题<title>，是否有正文 ，并检查是否有附件，附件是否能够正常下载。

2. 附件的下载：

首先我们需要识别附件，可以通过 F12 查看教务处网站附件的基本格式，均为 jwc.hit.edu.cn/_upload/…/xxx.doc(xlxs, doc, txt, etc.)。所以我们只需要找到结尾是 xlxs, doc, docx, txt 等且包含‘_upload’的 url 即可。利用 BeautifulSoup 抓取网页上所有的<a>标签的 href 属性值（即 url），接着用正则表达式筛选附件 url 即可。

3. 多线程抓取页面内容和下载：

利用 threading 库中的 Thread, Lock 类实现了多线程抓取网页和下载附件。继承 Thread 类并重写 run 函数实现 Mythread 类。run()函数中利用死循环（while True）实现了不同线程不断抓取不同 url 的过程，需要注意的是，在访问全局变量或记录全局变量时需要使用 Lock.acquire()函数获取锁并阻塞进程，防止全局变量访问值错误。在访问完全局变量后要及时释放锁 Lock.release()。合理利用锁我们能够为每个下载下来的附件创建对应 json 文件行数的目录，方便我们检查爬取的文件是否正确，json 文件写入是否正确。页面内容的抓取根据 F12 查看的网页 HTML 源码找到对应标签即可。需要注意的是对于某些列表网页或网址导航主页，其正文部分的属性为，利用 bs4 抓取对应属性下的 text 即可。多线程使程序抓取网页速度明显变快。

4. 礼貌规则：

利用 urllib 的 robotparser 模块访问每个网站的../robot.txt 目录并根据爬虫协议解读（本人爬虫名为 HIT_IR_CRAW_ROBOT）本爬虫能够爬取哪些网页。同时设置爬取时间间隔，craw_tlimit，在重写的 run 函数中，每爬取一次网页则 sleep 一段时间后再次爬取。

5. 分词和去停用词处理：

利用本校的 LTP 模块实现了对标题和正文的分词处理，并根据停用词表删去了部分词。同时完成了一些文本错误的清除，分词结果中的非法空格的删除等工

作。

6. 错误的处理：

无论是访问网页超时，下载附件失败，还是网页无标题无正文的情况，程序都做了对应的错误处理和错误提示输出，方便后续调试。

四、实验心得

多线程实现中 Lock 的使用尤为关键，若想要记录当前网页是第几个网页则先需要使用锁将 cnt += 1 再用局部变量拷贝一份，防止前后 cnt 值不一。爬取下的正文往往会包含一些不可见空格，分词后，简单地 strip()无法完全去除，还需匹配除去\u00a0 等不可见符（或扫一遍 word 查看是否有不可见符号，有就删去）。

本次实验让我对爬虫程序产生了浓厚的兴趣，同时锻炼了自己编写代码，debug 的能力。加深了对各模块的理解，同时也学习到了一些 HTML 的编码格式，受益匪浅。