苏州大学实验报告

院、系	计算机学院	姓名	赵鹏	学号	2127405037
课程名称	信息检索综合实践				
指导教师	李正华		实验完成日期		2022/5/25

实验名称： 爬虫和某机构主页检索系统

一. 实验目的

1.学习爬虫知识，尝试爬取某机构的完整静态网页，并保存网页源码。

2.利用爬取的网页制作完整的检索系统。

二. 实验内容

1.利用爬虫知识编写代码爬取苏大计算机学院的网站。

2.结合作业 3.4.5.7 制作完整的检索系统，并提供 ui 查询界面。

三. 解决思路（如流程图或伪代码）、遇到的问题和解决方法、运行结果

网页统计数据：

网页数	2988
句子数	1224680
词语数	55048

爬虫部分解决思路:

1.从苏大计科院主页开始，使用 DFS 算法递归爬取未被访问过的网页。

2.对于相对 url 的问题，进行特殊判断，如果不满足 suda.htm 或 html 和 http 在 url 中则将其加上苏大计科院主页的根网址。

3.每次从一个网页中先利用 bs4 解析网页包含的链接，在扩展新网页时，当这些网页全部被添加过则返回，避免无限递归。

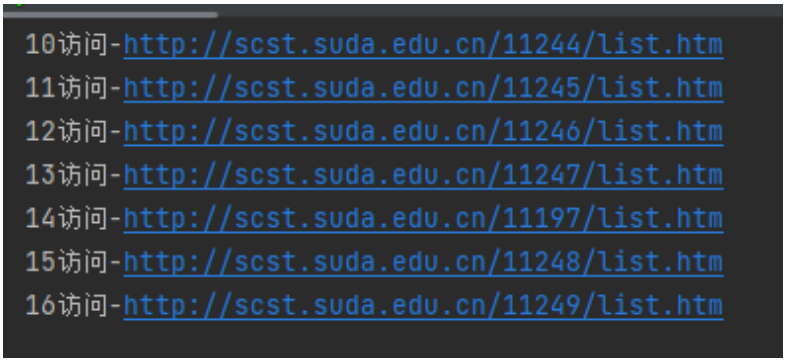
问题 1：部分网页在 get 时会出现时间过长而报错的问题。

解决办法:利用异常处理，将 get 放在 try 中，如果报错则直接 return。

问题 2:因有些网页标题相同，只使用标题作为文件名会导致标题重复的网页被覆盖。

解决办法：爬取网页时进行计数，保存文件时在网页标题最后加上编号作为区分。

运行结果：



综合项目部分解决思路：

1.对于爬取的网页源码，利用作业四的代码进行网页正文提取。

2.编写代码利用 jieba 进行分词，在分句时，以行为单位，根据，。等字符进行分隔。在提取正文时进行这一过程对句子和词语的计数。

3.修改生成倒排索引的代码中最大匹配分词模块，使其能够利用已经分词的语句进行倒排索引中的数据统计。运行程序获得倒排索引。

4.利用 python 的 tkinter 库编写 ui，并利用网页排序模块获得查询结果，将其显示在 ui 中。

因 1.2.3 步大部分代码与作业 3.4.5 相似，故具体细节不再阐述，下面具体说明 4 的思路。

UI 查询模块具体思路：

1.UI 查询界面选择使用 tkinter 进行制作。

2.主界面使用 Label，Entry,Button 等组件进行制作。

3.搜索结果展示界面每页展示 4 个搜索结果，支持翻页操作，使用 Label 显示标题，并给 Label 的文字添加双击鼠标事件，双击可打开提取正文但未分词的文本文件。使用 Text 显示包含查询结果的语句，若无法查询到相关文档则使用弹窗提示。

4.对于每次查询，调用网页排序模块函数进行查询，返回包含的所有网页和每个网页的相关语句。

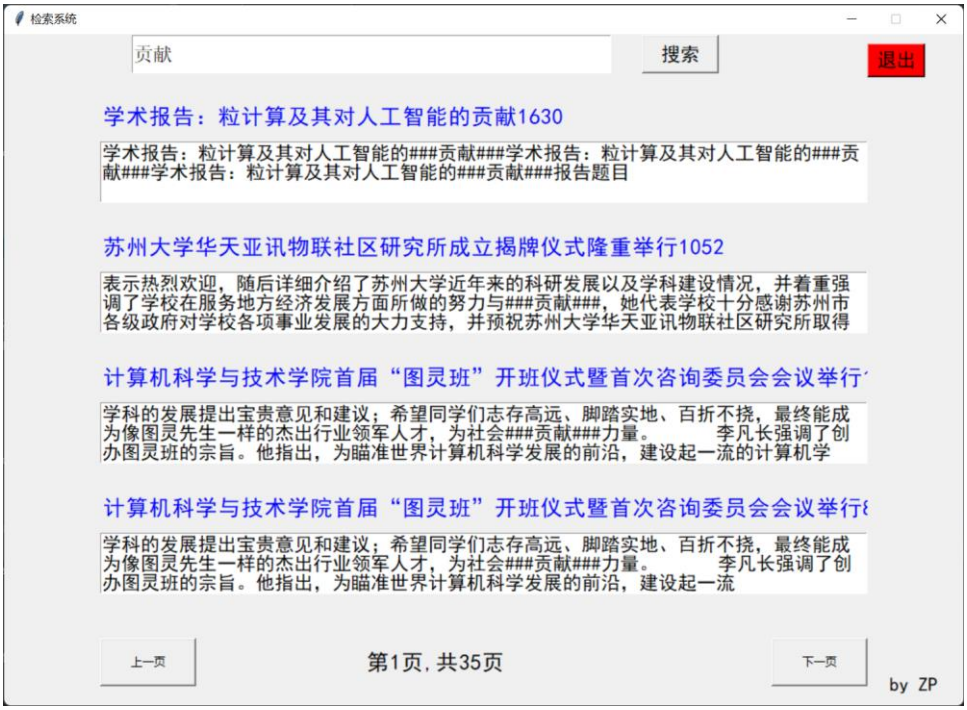
问题 1：当进行多次查询时，显示结果会叠加在一起。

解决办法： 每次显示查询结果结果前销毁前一次的控件。

问题 2：反复翻页时，会出现控件叠加（文字重叠）的问题。

解决办法：只有在控件首次使用的时候进行生成，否则只需显示已被隐藏的即可，无需重新生成新控件。

运行结果：



四. 实验总结

通过本次实验学习到了爬虫的相关知识，能够编写代码使用深度优先搜索算法递归爬取网页。

学习到了 python 的 tkinter 模块制作 ui 界面的方法并能够利用查询接口在界面中显示查询结果。

教务处制