一、实验内容及功能说明

本程序实现了一个基于 PageRank 和分词算法的在 news.sina.com.cn 下进行新闻关键字搜索的搜索引擎。

只需在框内输入想要搜寻的关键字,点击搜索按钮即可获得搜索结果。

二、设计方案与设计思路

本程序由三个主要部分构成: 爬虫、排序、GUI。

爬虫部分爬去从 news.sina.com.cn 为根节点的,深度为 2 以内的,具有相同的前四级域名的新闻网站。这其中涉及到几次过滤。第一次过滤是根据四级域名进行过滤,这个可以直接使用正则表达式解决。如果出链网站具有不同的四级域名,则不将其放入队列。第二次过滤是根据 url 进行网页性质的判断,即判断一个网页是属于门户(即拥有诸多出链但是本身没有内容)还是属于文本内容(新闻主页面)、文件(pdf, mp3 等)、视频网站(主内容为一个视频播放器)。判断的方法如下: 如果 url 的结尾是一个/,或者它的名字中包含 index,则认为它是一个门户网站; 如果结尾为文件格式,则认为是文件; 如果名字中包含 video,则认为它是一个视频网站; 否则它是一个内容网站。最后一层过滤是根据内容网站的 HTML 提取正文。处理办法是使用正则表达式匹配和,这样会得到页眉页脚的一些不重要信息,但是不影响搜索结果。

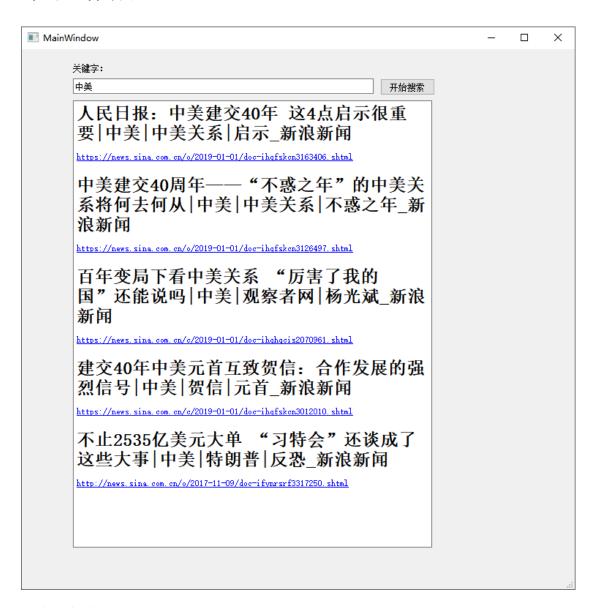
排序则是使用了两个标准加权之后得到一个最终的权重进行排序。这个权重的组成方式如下: pagerank (10%), 关键字在正文中的出现频率(90%)。可以预料的是, 如果关键字在任何一个网页中都找不到匹配时, 返回的结果总是一样的(即

学号:	17231164	姓名:	丁元杰	2019-01-04
		_		=

pagerank 排名最高的 5 个网页)

GUI 则使用了 pyQt5 库, 并没有请设计师设计, 虽然丑但是可以用。

三、程序运行效果



一个运行良好的结果。

学号: __17231164 ____ 姓名: __丁元杰 ____ 2019-01-04

Window	-	
关 键字:		
中美		
人民日报:中美建交40年 这4点启示很重要 中美 中美关系 启示_新浪新闻		
https://news.sina.com.cn/o/2019-01-01/doc-ihqfskcn3163406.shtml		
中美建交40周年——"不惑之年"的中美关系将何去何从 中美 中美关系 不惑之年_新浪新闻		
https://news.sina.com.cn/o/2019-01-01/doc-ihqfskon3126497.shtml		
百年变局下看中美关系 "厉害了我的国"还能说吗 中美 观察者网 杨光斌_新浪新闻		
https://news.sina.com.cn/c/2019-01-01/doc-ihqhqcis2070961.shtml		
建交40年中美元首互致贺信:合作发展的强 烈信号 中美 贺信 元首_新浪新闻		
https://news.sina.com.cn/c/2019-01-01/doc-ihqfskcn3012010.shtml		
不止2535亿美元大单 "习特会"还谈成了 这些大事 中美 特朗普 反恐_新浪新闻		
http://news.sina.com.cn/o/2017-11-09/doc-ifynrsrf3317250.shtml		

不具代表性的搜索结果。



找不到关键字的搜索结果。

四、设计亮点(自选应用必填)

采用非常简单的手法实现了一个基本可用的程序。

五、实验总结

(1) 在实验中遇到了哪些问题? 是如何解决的?

首先的头疼问题是找库和研究每个类的功能,网上的教程大多数都是使用三方库的,但 是强行搜一搜还是能够找到官方库的函数用法。

学号:	17231164	姓名:	丁元杰	2019-01-04

其次的问题是 GUI 并不好做,选择什么模块编写 GUI、如何显示结果、如何合理划分模块都是只学过 OPP 的我无法一次安排清楚的。不过跟着往上的无数教程,还是成功地将 GUI 的设计与功能的实现基本分离。

(2) 收 (jiao) 获 (xun) 和体 (tu) 会 (cao)

还是控制台程序写起来令人神清气爽。