上海理工大学光电信息与计算机工程学院

**《知识图谱》作业**

2023-2024**学年第一学期**

|  |  |
| --- | --- |
| **专　　业** | 数据科学与大数据技术 |
| **姓 名** | 罗震宇 |
| **学　　号** | 2135060321 |
| **年 级** | 2021级 |
| **指导教师** | 彭敦陆 |
| **课程代码** | 12004310 |
| **成 绩** |  |
| **教师签字** |  |

****

**目录**

[作业 构建一个搜索引擎 4](#_Toc154060967)

[1. 任务描述 4](#_Toc154060968)

[2. 作业内容 4](#_Toc154060969)

[3. 实验环境 4](#_Toc154060970)

[4. 实现过程 5](#_Toc154060971)

[4.1 爬取并解析网页 5](#_Toc154060972)

[4.2 建立并存储倒排索引 7](#_Toc154060973)

[4.3 网页链接图和实现PageRank算法 7](#_Toc154060974)

[4.3 实现基于关键字的搜索 9](#_Toc154060975)

[5 实验小结 10](#_Toc154060976)

作业 构建一个搜索引擎

## 1. 任务描述

利用爬虫技术、倒排索引、Hits/PageRank等算法，实现一个搜索引擎原型系统。

## 2. 作业内容

1. 实现网页爬虫，爬取不小于10000万网页

2. 对爬取的网页建立倒排索引（inverted list)

3. 构建一个网页连接图（建议连接阵200\*200，或其它能说明算法特点大小的），并求各节点的PageRank值

4. 基于所建索引实现基于关键字的搜索，并将结果按PageRank排序输出

## 3. 实验环境

处理器 AMD Ryzen 7 5800H with Radeon Graphics

内存 16 GB

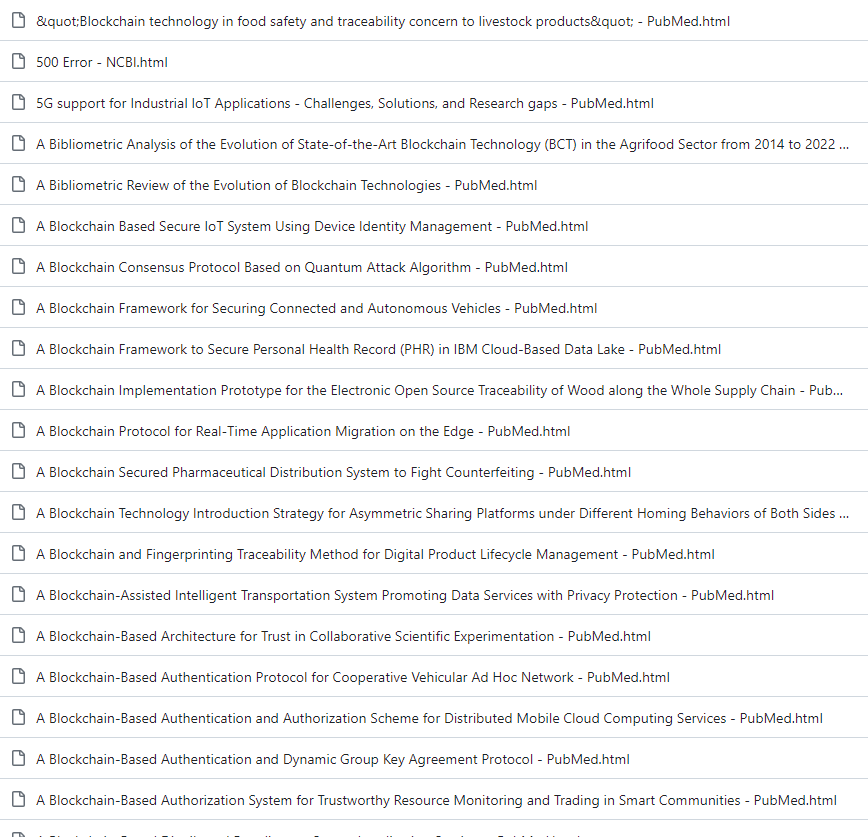
系统 Windows 11  
IDE PyCharm Professional 2023.2.4

## 4. 实现过程

### 4.1 爬取并解析网页

本次作业选取[PubMed (nih.gov)](https://pubmed.ncbi.nlm.nih.gov/)数据库作为数据源，爬取了以“中国”、“计算机”、“区块链”、“数据库”为关键词的相关论文网页。本系统属于美国国立卫生研究院下属的国家医学图书馆维护的Entrez信息检索系统的一部分。PubMed取名是“Public Medicine”（公共医疗）的意思。

自1971年至1997年，MEDLINE数据库主要通过大学图书馆等学术机构访问。PubMed于1996年1月发行后允许用户免费在线检索MEDLINE。1997年6月PubMed正式对公众开放。



图表 1 爬取的网页

爬虫程序的工作原理如下：

1.程序使用requests库向目标网站发送一个GET请求，获取网页的HTML内容。

2.程序使用正则表达式从HTML内容中提取出总页数。

3.程序会让用户输入想要获取的网页数目。

4.程序会根据用户输入的数目，构造新的URL，然后再次发送GET请求获取新的HTML内容。

5.在获取的HTML内容中，程序会使用正则表达式提取出每个页面的URL。

6.对于每个页面的URL，程序会发送GET请求，获取页面的HTML内容。

7.程序会从每个页面的HTML内容中提取出标题，并将HTML内容保存到本地文件。

8.程序会打印出每个页面的标题。

程序在发送GET请求时，会使用headers参数来模拟浏览器的User-Agent，以防止被网站屏蔽。同时，程序也包含了异常处理机制，当访问某个URL时出现错误，程序会忽略这个错误并继续运行。

在爬取的网页当中，通过正则表达式解析出论文的标题、日期、作者、摘要等内容，并且在存储后进行分词。由于爬取的都是英文网页，所以使用Keras的Tokenizer类进行分词。这是一个文本标记化工具，用于将文本转换为数字序列，即将每个词用数字编码，它也可以对文本进行向量化，将每个词转换为一个向量。最后将分词结果保存至本目录下。



图表 2 词汇表目录



图表 3 某一个网页的词汇表

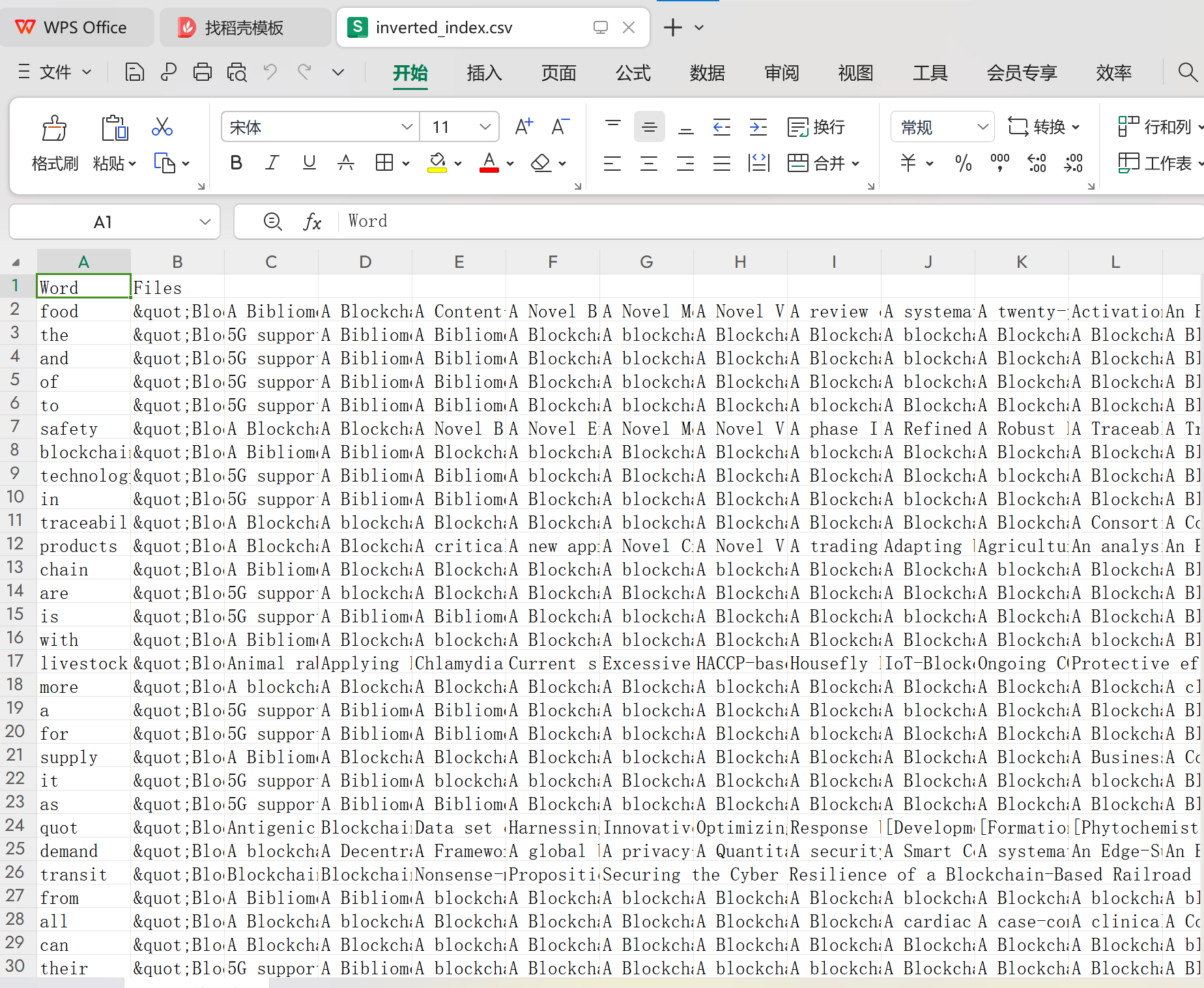
### 4.2 建立并存储倒排索引

倒排索引是一种数据结构，用于存储每个单词出现在哪些文档中，以下是代码的工作流程：

1.定义一个名为inverted\_index的字典，用于存储倒排索引。字典的键是单词，值是一个列表，存储包含该单词的文件路径。

2.使用os.walk函数遍历词汇表目录中的所有文件。对于每个文件，读取其中的所有单词，并将文件路径添加到对应单词的文件列表中。

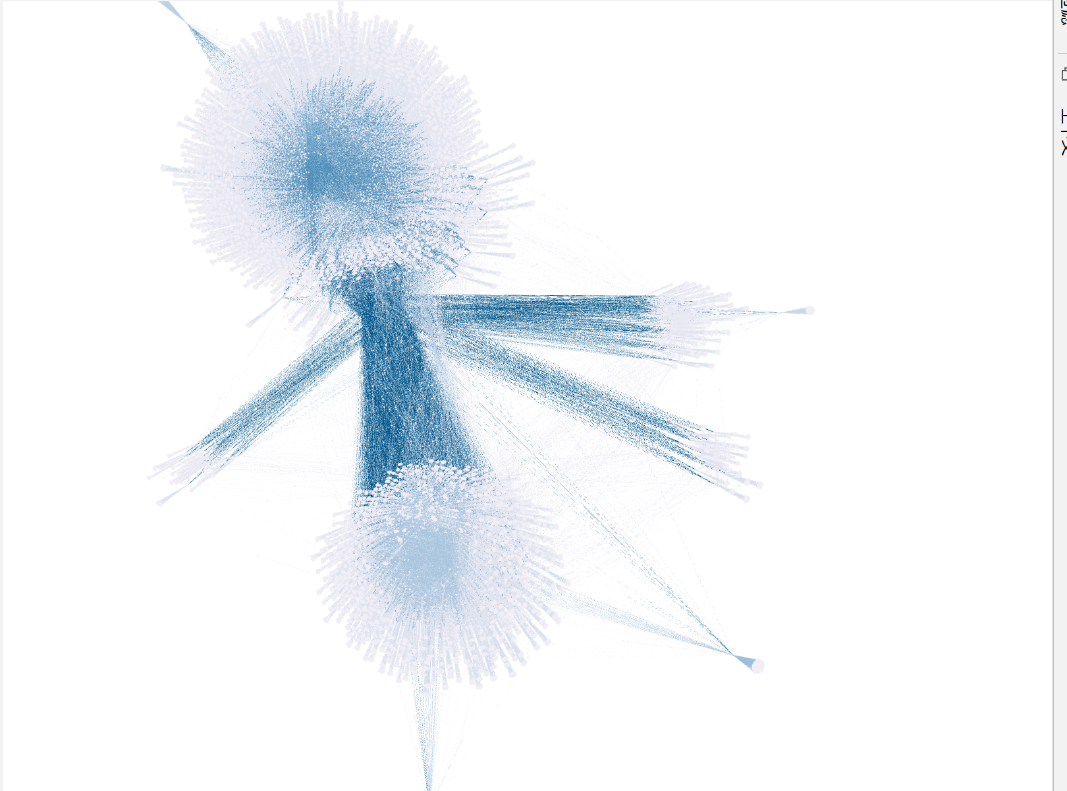
3.所有文件处理完毕后，将倒排索引写入到CSV文件中。首先写入表头，然后对于inverted\_index字典中的每个条目，写入一个CSV行，行的第一个元素是单词，后面的元素是包含该单词的文件名。



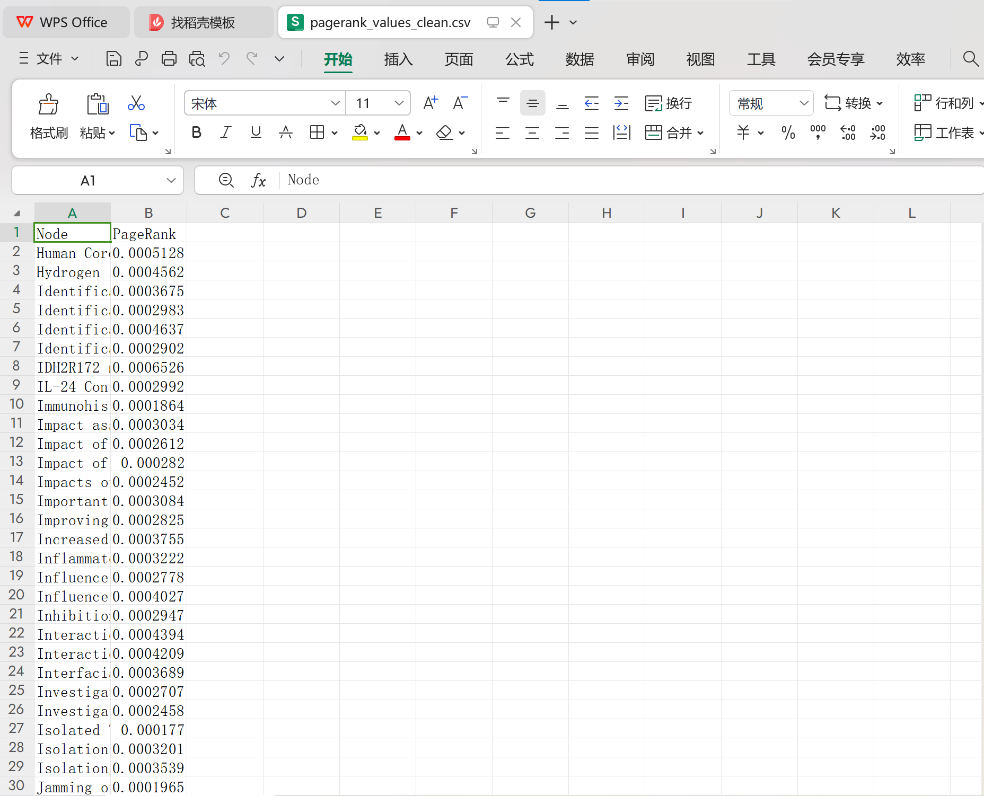
图表 4 倒排索引内容

### 4.3 网页链接图和实现PageRank算法

按照实验要求，需要形成网页链接图，并且需要使用PageRank算法对关键字检索的网页链接进行排序。经过计算得到PageRank值，部分如下所示：



图表 5 网页链接图可视化



图表 6 部分pagerank计算

### 4.3 实现基于关键字的搜索

使用Python的flask框架搭建web服务器，自己设计前端页面，与服务器后端进行连接，前端用户通过搜索框输入自己的想要找寻的关键词，系统则将请求发送到服务器，服务器再将搜索结果以json文件格式返回前端。

## 5 实验小结

本次实验加深了我对搜索引擎原理的理解，提高了数据处理与分析以及网络编程等技能，加强了对编程项目的兴趣和动手能力。

首先，我对搜索引擎的工作流程有了更深入的理解。我了解到，搜索引擎是通过爬虫获取网页信息，然后对网页内容进行分词和建立倒排索引，最后根据用户的查询返回相关的网页。

在实现搜索引擎的过程中，我学会了如何使用Python的requests库进行网络爬取，如何使用正则表达式提取网页的关键信息，如何使用Keras的Tokenizer进行文本分词，如何构建倒排索引，以及如何使用PageRank算法对网页进行排序。

在项目中，我认为我做得好的部分是：在爬虫部分和网页存储及解析部分做得比较好，这些地方都充分考虑到了实际情况，学习到了很多工程应用的知识。自己也在Python编程方面更加熟练了。

然而，我认为我做得不好的部分是：在处理网页链接图部分时，我没有很好地处理无效链接，这导致在进行PageRank计算时产生了很多无效的结果，需要进行额外的数据清洗；在搜索引擎部分，我没有对用户的查询进行有效的处理，例如，我没有实现查询扩展或查询修正的功能，这可能会影响搜索结果的质量。

总的来说，这次实验让我对搜索引擎的工作原理有了更深入的理解，也让我学会了如何使用Python实现搜索引擎的各个部分。虽然还有许多需要改进的地方，但我相信通过不断的学习和实践，我能够做得更好。