信息与知识获取实验二实验报告

2019211283 陈童

2019211287 李隆昊

2019211273 秦志迎

## 实验要求

实现一个信息检索系统，中、英文皆可，数据源可以自选，数据使用爬虫获取，规模不低于100篇文档，进行本地存储。中文处理可以使用开源代码分词，也可以使用字作为基本单元；英文可以直接通过空格分隔。要求构建出数据源的基本倒排索引文件，实现基本的向量空间检索模型的匹配算法；支持用户输入进行查询，查询结果输出按照相关度从大到小排序，列出相关度、题目、主要匹配内容、URL、日期等信息。界面不做要求，可以是命令行，也可以是图形操作界面。

## 使用爬虫进行文档获取

在本次实验中，我们选取了中文马克思主义文库作为资料来源。马克思主义哲学对于当代中国的发展有重要的指导意义，对于社会发展有重要作用。

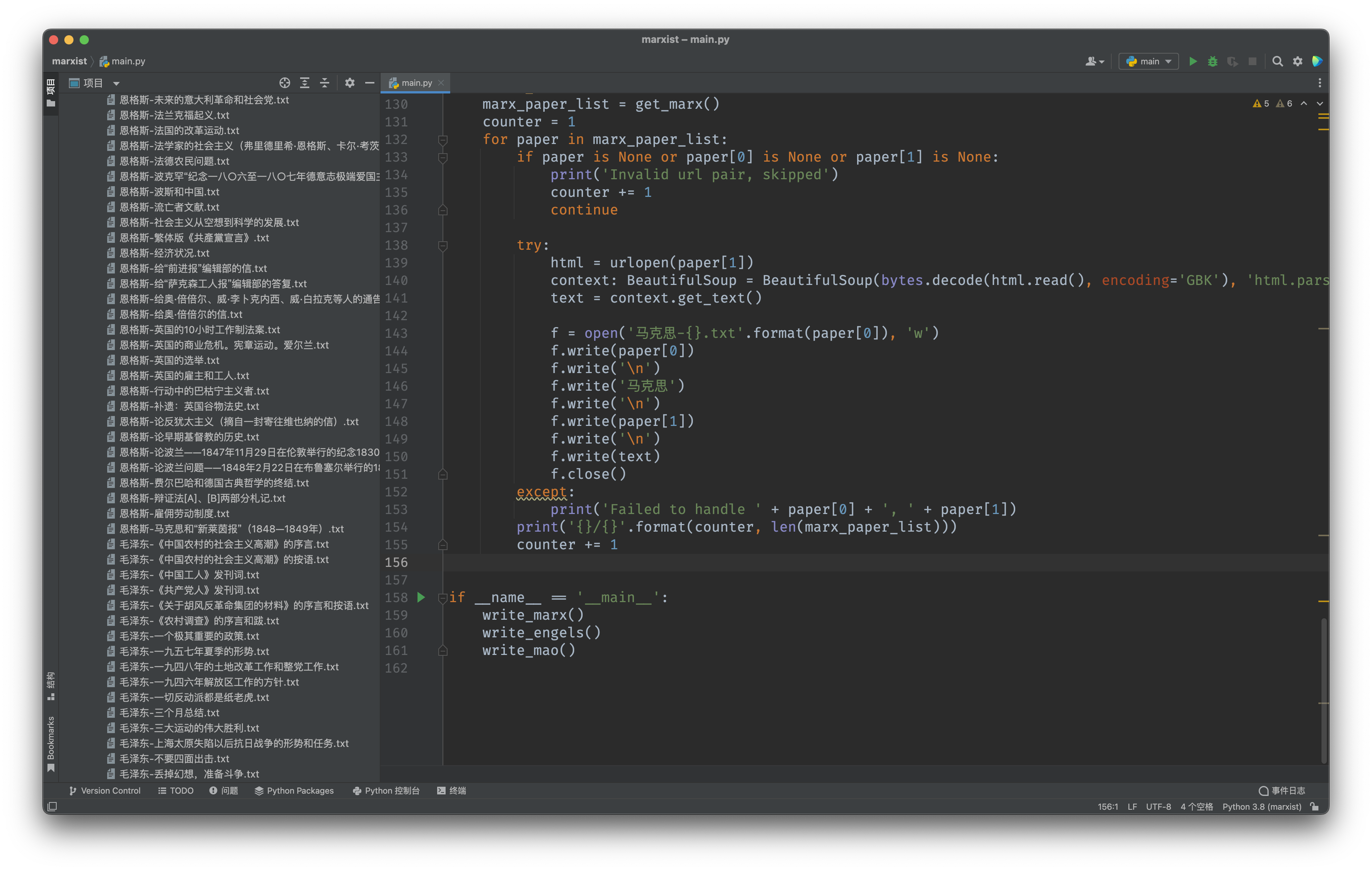
我们使用python作为爬取文档的编程语言。我们使用了urllib库用于请求网页，使用BeautifulSoup库用来解析html。

考虑到中文马克思主义文库中资料众多，我们决定选取其中的一部分用于数据处理。我们爬取了马克思、恩格斯与毛泽东三人的部分文章。

在爬取每个人的文章时，我们首先使用urlopen获取其文章索引页面，然后使用BeautifulSoup解析html中的所有a标签。然后，排除掉其中的页面内部跳转以及本次实验中无法处理的pdf、chm文档，剩余的a标签指向的页面即为本次实验中需要的文章链接。将这些文章获取下来，使用BeautifulSoup排除其中的标签信息，抽取出文本信息，即为本次实验中使用的文本数据。

在保存文章时，为避免重名文章，我们在每篇文章的命名前加上了作者的名字。

爬虫的结果展示如下：



左侧为部分爬取的文章，右侧为部分源代码。

## 算法介绍

本次实验使用**空间向量检索模型算法。**实现一个马克思主义文学信息检索系统。

### 3.1 空间向量检索模型介绍

**向量空间模型**是一个把文本文件表示为标识符（比如索引）向量的代数模型。文档和查询都使用向量表示。

descript

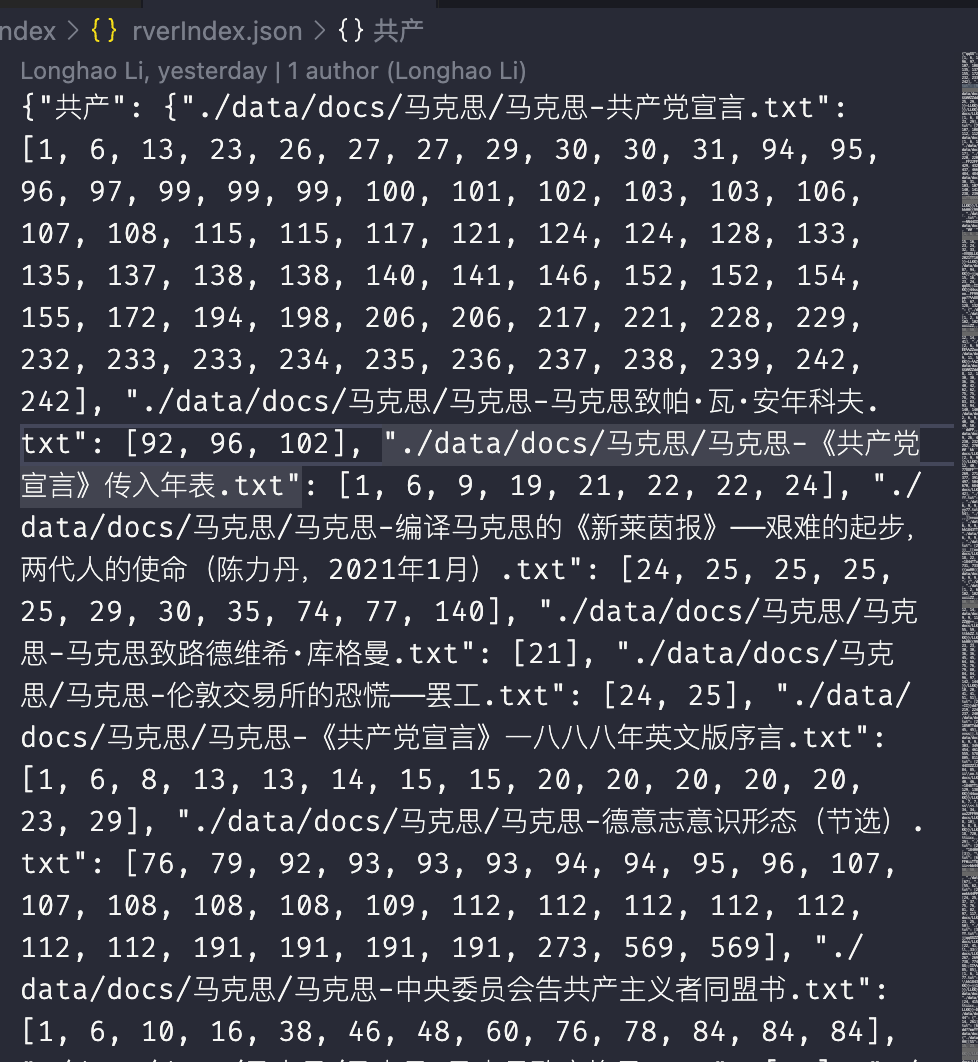
descript

每个位置对应固定的词组。位置的值代表出现过的次数。

输入的词和文档的相关度由各自的向量对应的descript夹角来定义。descript越大，代表双方相关性程度越大。

## 3.2 算法具体实现

1. 使用jieba开源分词库，按行抽取各个文档的内容进行分词，并建立反向索引。

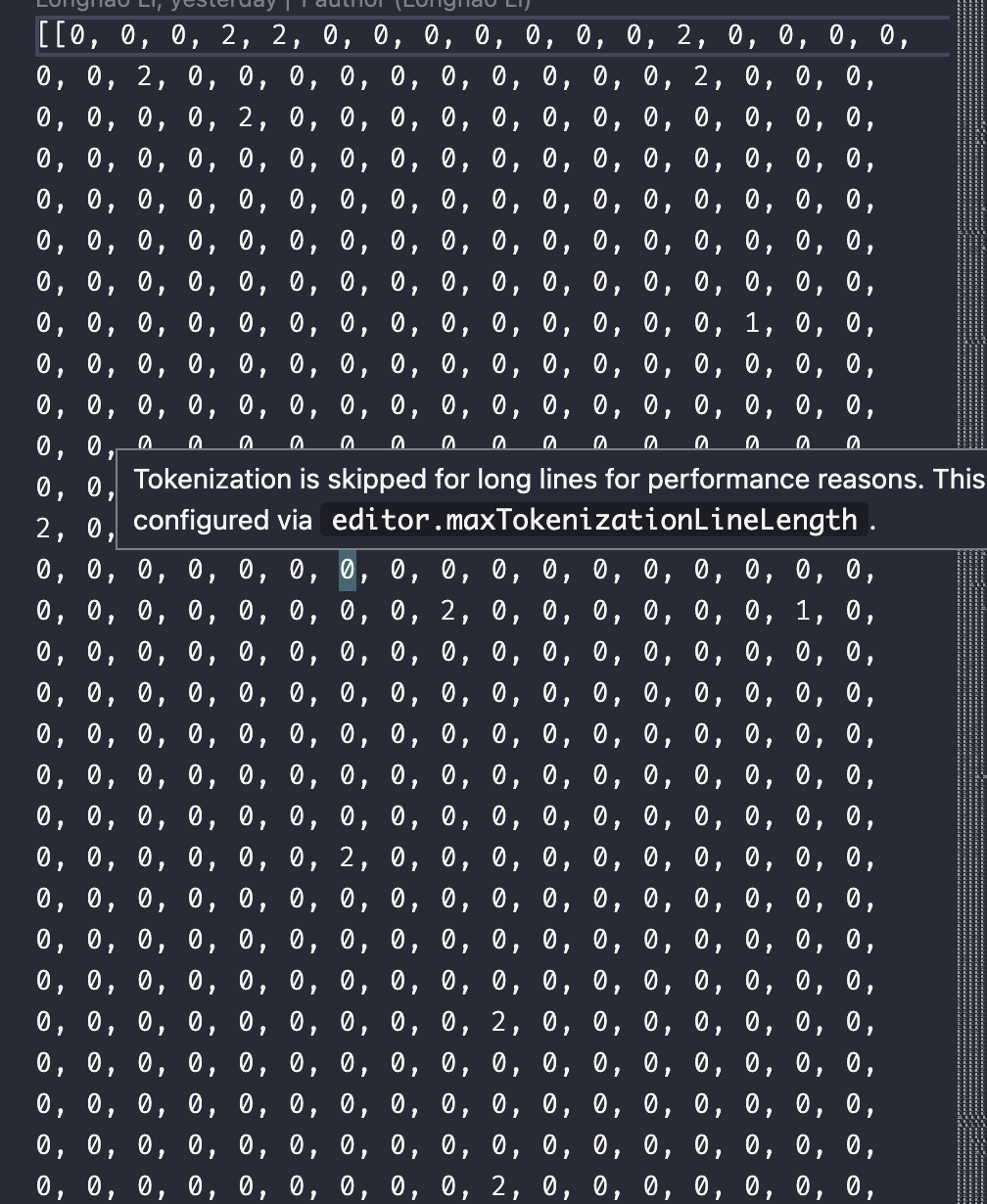


如上图所示，反向索引显示“共产”一词在马克思-共产党宣言的文章中分别在1，6，13..等行数出现。注意，因为有可能分出没有意义的词，所以在建立反向索引时会过滤掉所有的含有非中文，英文和数字的词语，提高模型的准确度，降低本地空间占用。

该功能由 buildReverseIndex()完成

1. 根据反向索引建立每个文档的词频向量，并存储在fileFreqAddr.json中。

且为了加快计算速度，所有的向量用python的科学计算库ndarray类型存储。并且所有的文章按照顺序排列，组成矩阵。矩阵的计算会更快。

输出结果如图所示

为了保证每个词对应的位置在所有的数据结构中一致，构建反向索引后的所有数据结构的词典都来自reverseIndex.json,且按照字典序排列。

该部分功能由get\_file\_word\_freq2(path: str) -> Dict[str, Dict[str, int]]完成。

1. 获取输入向量

向用户问询一行输入，转换成向量。该部分内容由inputVector（）完成。

1. 计算相关度，返回文章

通过ndarray内置的矩阵乘法，计算相关度。

计算之前，使用sklearn中的向量归一化函数将向量转化为l2范数的标准化向量。如果不进行归一化处理，那么长的文章会更有优势。归一化后模型可以通过词的含量百分比进行判断，而非单纯的词频。

1. 用户反馈

返回后，用户将有用的文章编号返回给系统，系统根据用户的输入，将对应的文章的词频增加，添加下次搜索该关键词时有用文章的相关度。

## 系统展示

本次实验过程中，我们的搜索引擎系统最终可以在命令行运行，当程序开始运行之后，系统会提示我们输入我们想要搜索的语句，输入回车之后，系统将对我们输入的语句在本地爬虫所得到的文档仓库进行检索，并且通过空间向量模型生成文档和查询的向量，并且计算相关度，在计算相关度完成之后，系统将对查询结果中各篇文章的相关度由高到低进行排序，并且选择相关度前五的文章进行输出其相关信息，包括文章序号、相关度、题目、主要匹配内容、URL、日期等信息，在输出相关信息之后，系统会提示用户进行用户评价，选择出对于用户有用的文章序号，系统会根据用户的选择在文档库中修改对应文章的词频，为之后用户进行搜索的结果进行优化，以便用户可以搜索得出更加可靠的结果。在用户完成一次搜索之后系统会提示用户进行下一次搜索，如果用户想要推出搜索系统输入quit完成退出。

系统运行示例如下：

我们首先输入我们想要搜索的词汇：共产马克思恩格斯社会财富革命资本基本，在完成搜索后系统将展示搜索结果，并提示用户进行用户评价，我们输入我们感兴趣的文章的序号，系统将对本地文档库进行优化，在完成这次搜索之后，我们下一次搜索同样的信息，可以发现搜索出来的结果中，我们之前感兴趣的文章的相关度发生了变化，这证明我们每一次用户评价都有助于提升本系统的准确性。



## 实验心得

秦志迎：通过本次实验我对信息检索系统的工作原理有了基本的了解，同时也对课上学习过的倒序索引以及空间向量模型相关的知识掌握的更加透彻，同时我的编码能力也得到了一定的提升，收益颇多。

李隆昊：通过本次实验，我学会了使用python爬虫从网络上获取信息，对于信息检索的过程有了更加深刻的理解。除此之外，我对空间向量模型算法也有了更好的理解，这使我对于搜索引擎检索数据的原理也有了更加深入的了解。

陈童： 通过本次试验，我更加理解了信息检索中向量空间模型的原理。通过实践我了解了自然语言处理的输入输出规范化以及使用中间文件减少计算量的方法。更加熟悉了python语言中sklearn等机器学习框架的使用。