

班级

学号

姓名

时 间

智能

201900161140

张文浩

2021年9月16日

实验报告-实验一

信息检索与数据挖掘

实验内容

Inverted index and Boolean Retrieval Model

– 使用我们介绍的方法，在tweets数据集上构建inverted index;

– 实现Boolean Retrieval Model，使用TREC 2014 test topics 进行测试；

– Boolean Retrieval Model：

• Input：a query (like Ron and Weasley)

• Output: print the qualified tweets.

• 支持and, or ,not；查询优化可以选做；

# 实验环境：

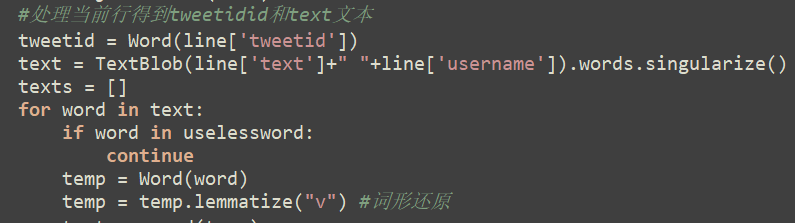
win10+spyder（anaconda）

实验步骤

# 一、数据预处理

读入数据，将“tweetid”作为docid，并提取出来。将username作为text的一部分加入待处理的文本。

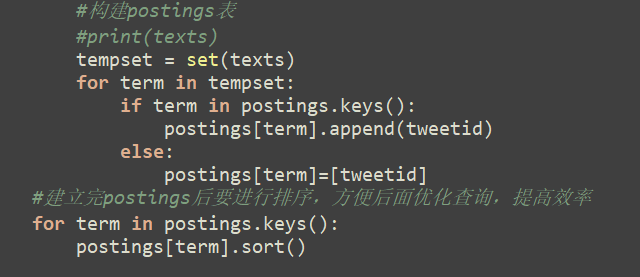
进行数据预处理，先将所有大写转小写，利用TextBlob库对text文本进行词干提取（名词变单复数等操作），利用lemmatize函数进行词形还原



# 二、构建postings

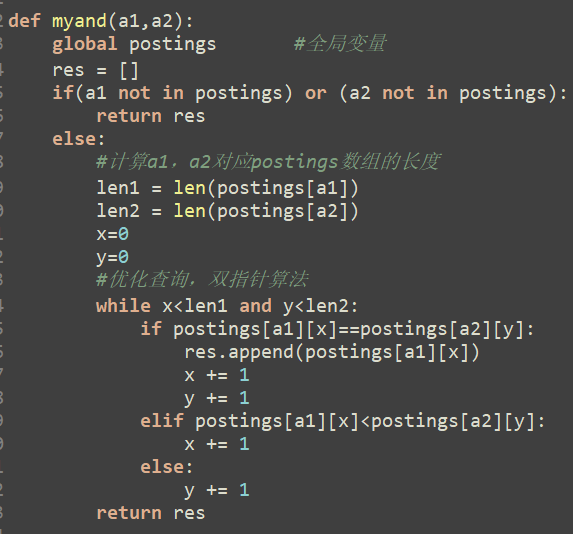
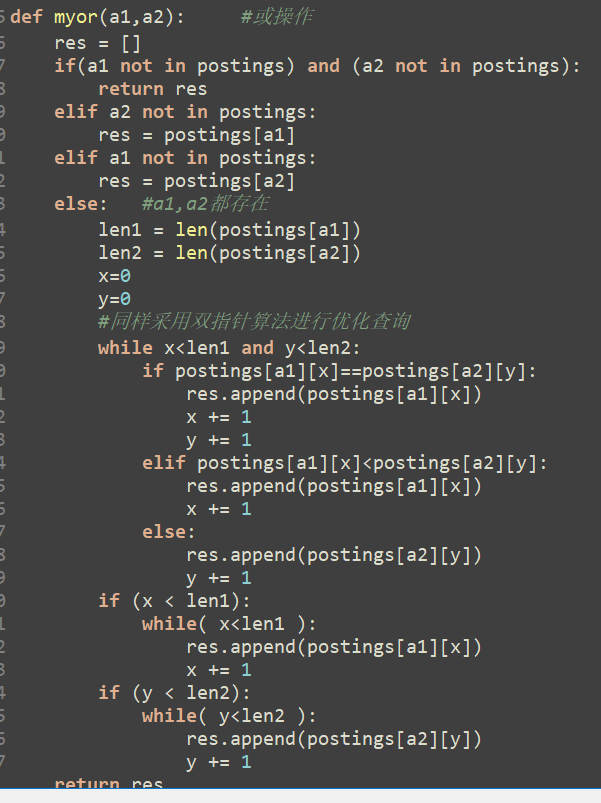
建立一个处理整个文本的函数，来构建postings表。

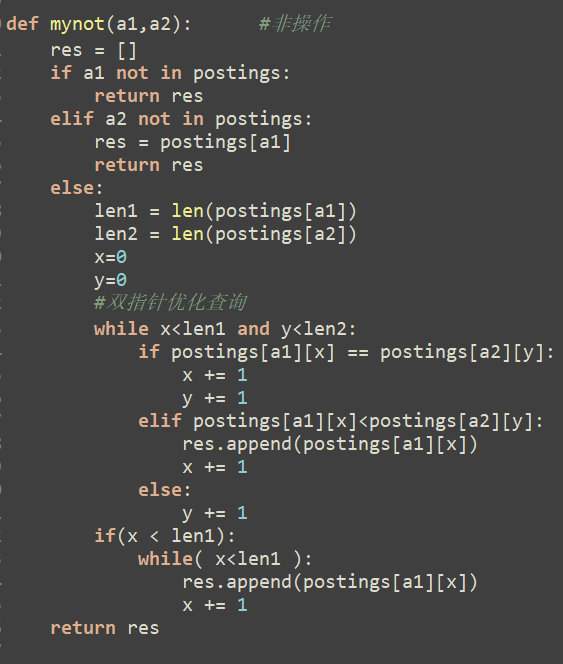
一行一行读入，先将所有字母变为小写，调用上面的函数，得到每一个document的docid和terms，构建postings表。在处理完所有document之后对postings表进行排序操作，按照docid进行排序，便于优化查询处理



# 三、构建查询函数

构建是函数分别处理and，or，not操作，三个函数均采用双指针算法进行优化查询，在O（n）的时间复杂度下解决问题

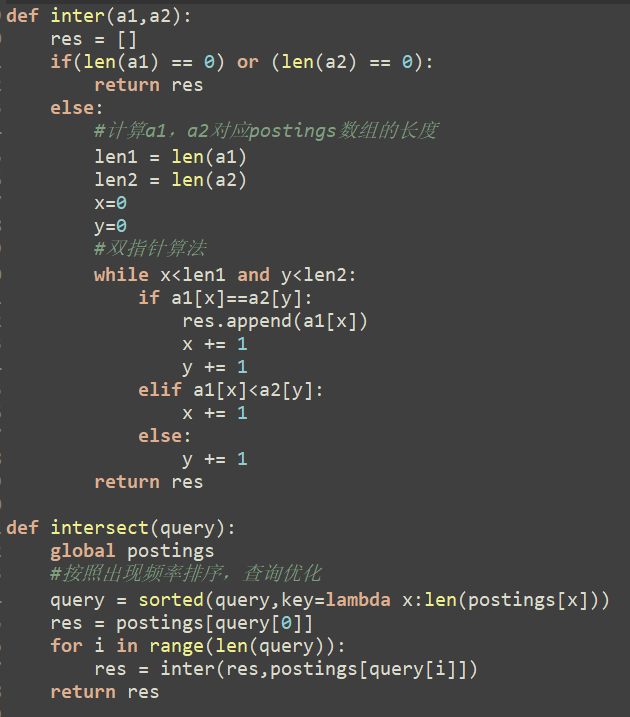
 



优化查询：

仅做了and的优化查询操作，输入时可输入多个单词，单词之间用空格分隔开，查询这几个单词同时出现的docid。

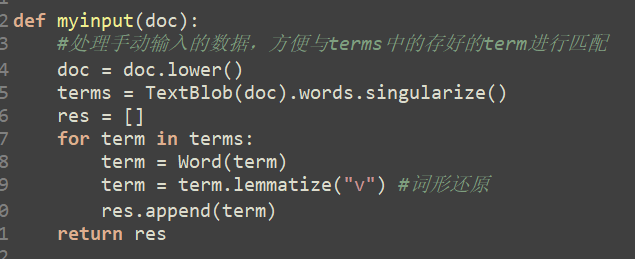
函数实现如下：



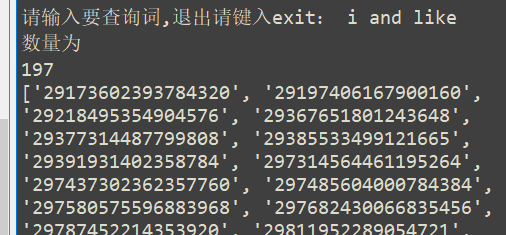
inter函数是将两个序列的取交集操作，待intersect调用

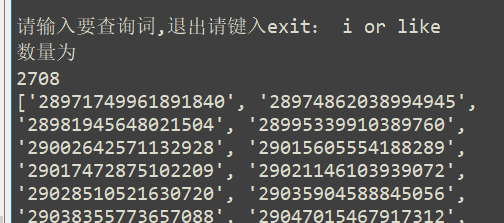
在intersect函数中先对输入的单词按照在document中出现的次数从小到大排序，进行优化查询，再进行两两合并。

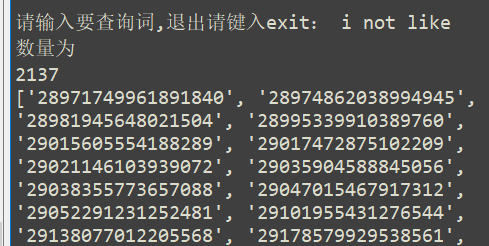
构建处理input的search查询函数，在接收输入的单词之后也要进行singularize和lemmatize数据归一化词形还原的预处理操作。

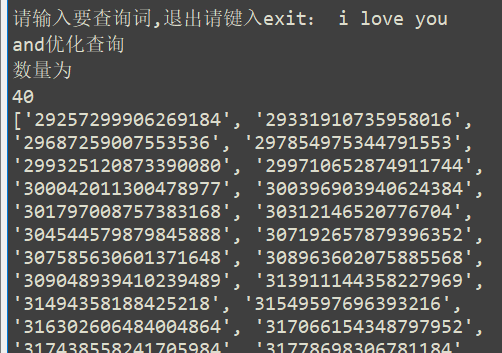


实验结果

and：

or：

not 

and 优化查询

实验总结

通过本次实验，我实现在tweets数据集上构建inverted index ，学习了如何利用TextBlob库进行文本预处理操作，实践了and，or，not的优化查询，对课本上的内容有了更加清晰的认识