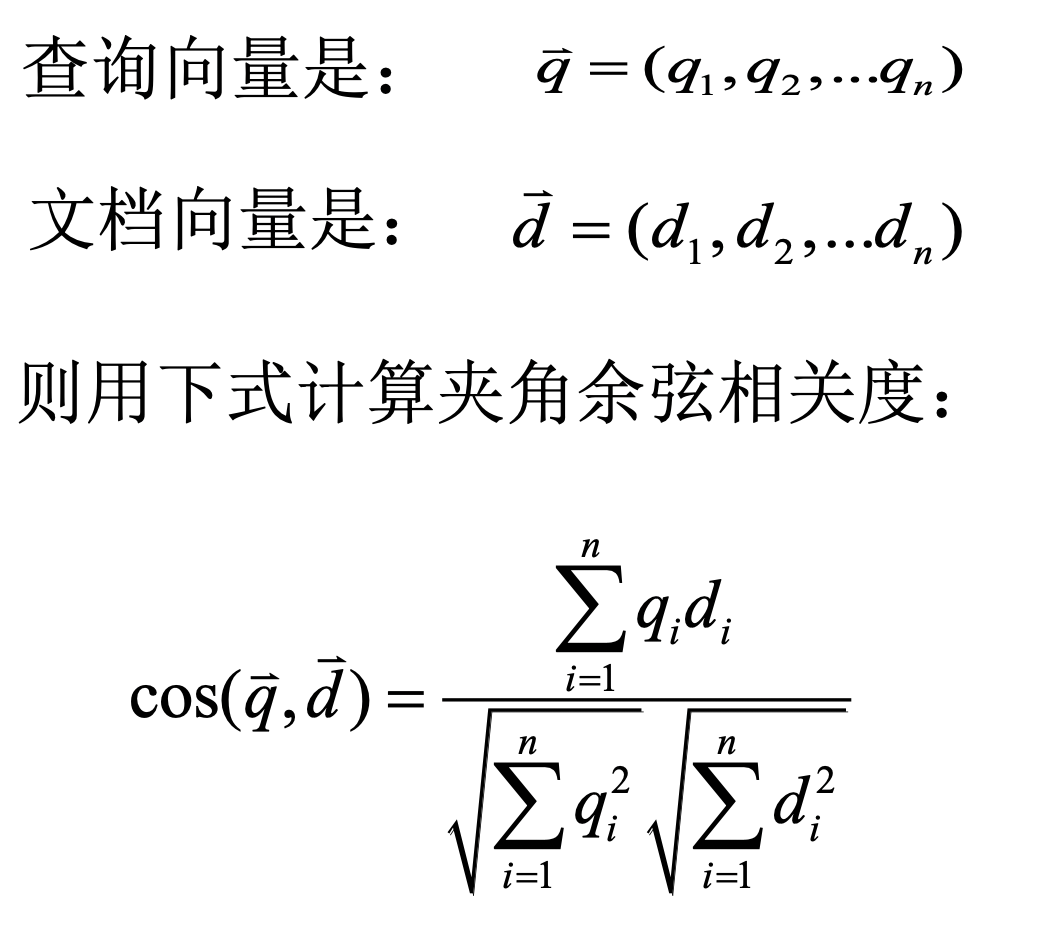
spider.py是爬取《基督山伯爵》117章（实际上那个网站少了一章，实为116）的代码，需要install beautifulsoup4这个库，与信息检索系统无关；

doc文件夹中存储的就是txt文本，共116个；

chinesestoptxt.txt是jieba分词时要忽略的字眼，比如“的”。

main.py上半部分是分词，需要install jieba库。

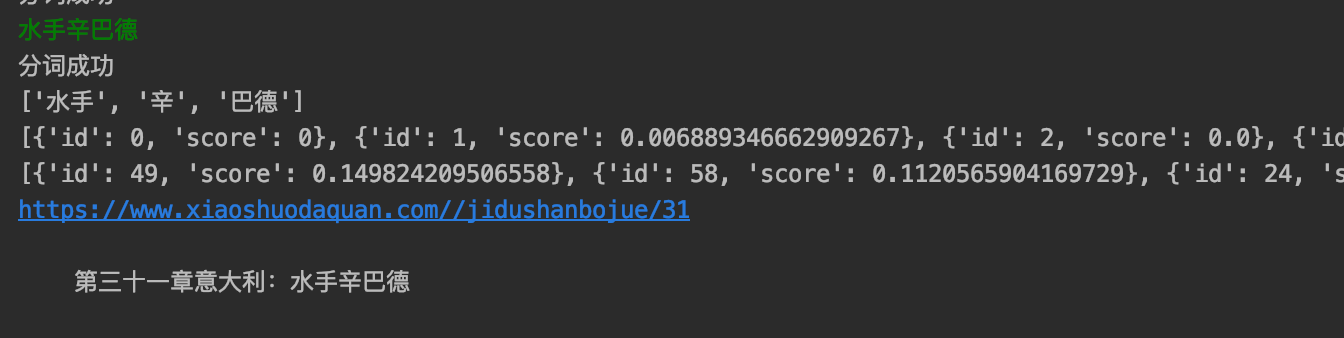
main部分是输入串及计算相似度值，最后储存在W数组里并排序。其中cos计算用的此公式：

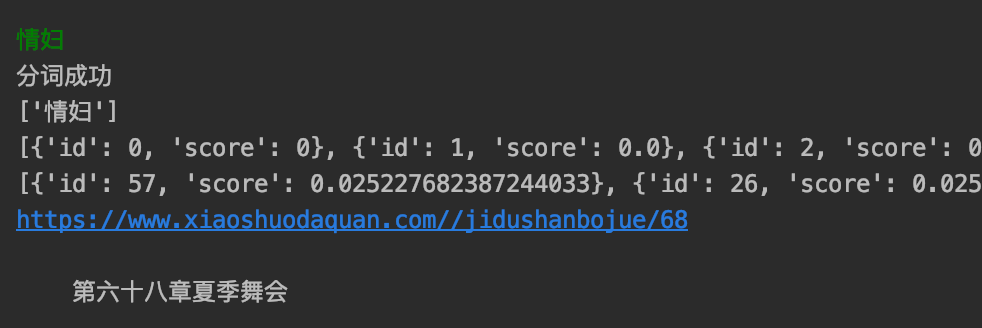


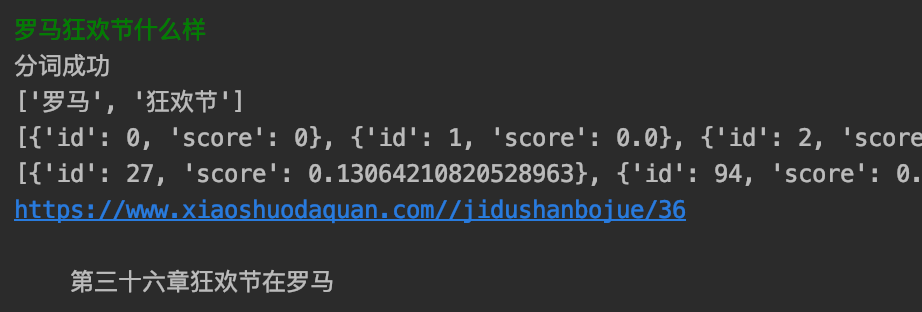
jieba库import成功就可以直接run这个程序，会给10次输入机会，10次之后就退出了。

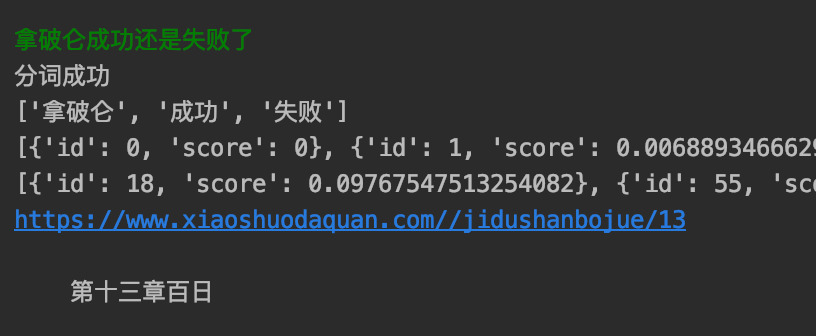
程序演示结果举例：

绿色是输入；然后将输入分词；之后是按照相似度排序先后的权值数组W；最后给出原链接url以及在我本地的txt文件的标题。我们选取的是相似度最大的那篇文章。

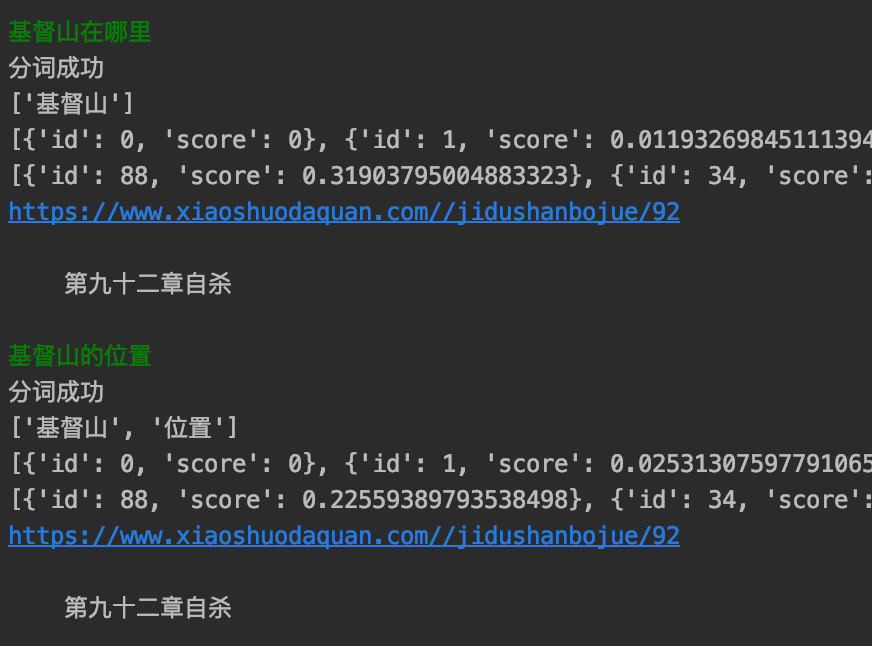








根据我的记忆，上面搜索是准确的。也有不准的，比如：



进去看了一下，原因是后期大仲马直接以“基督山”称呼主角……目前实现的这个算法中次数是会影响到匹配程度的，导致那一章“基督山”作为人名而不是地名出现了好多次，排序就靠前了。