12组（新冠疫情爬出地球）的信息检索项目总结报告

71119117 陈铎文 71119118丁一铭

1. 项目设计

前期工作：在古诗文网中爬取200首宋词，将其编号存入dataset中的200个txt文件中

1.项目图形界面

实现项目UI的搭建以及完成对搜索做出反应，提供给用户查询接口

1. 读取与分词模块

读取我们爬取的200宋词并对其进行分词，并构建相应的倒排索引

3.Bool模型模块

实现bool模型中简单的and or andnot 的简单查询以及如（xx and xx and xx）的复杂查询功能

4.RSV模型模块

实现概率模型的功能，处理需要的自然语言，按照RSV的大小返回所需要的结果

5.一元语言模型模块

读取用户的搜索，返回以一元语言模型匹配后的结果

6.二元语言模型模块

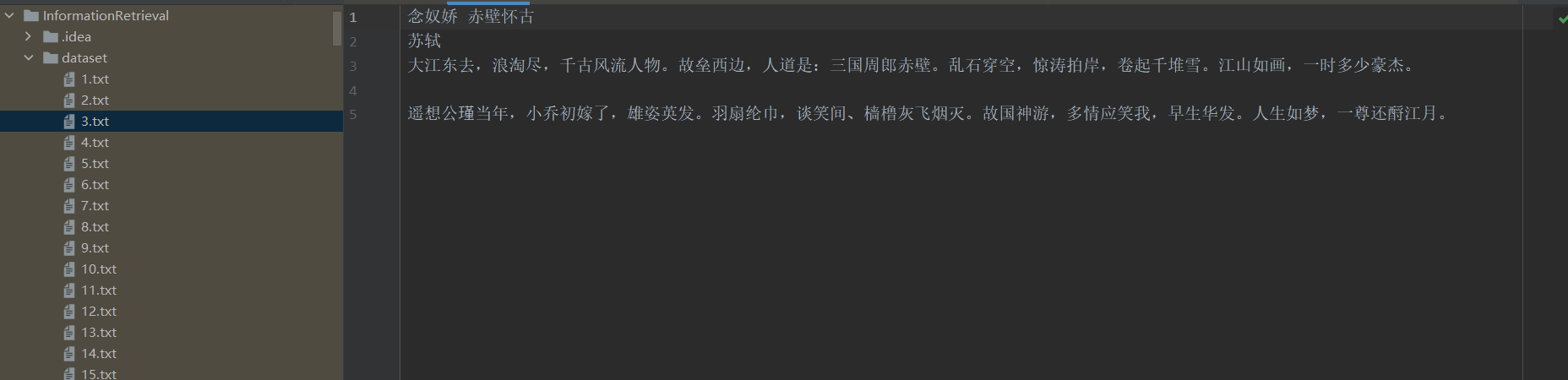
读取用户的搜索，返回以二元语言模型匹配后的结果

1. 项目实现

基于eBM(Extended Boolean Model)

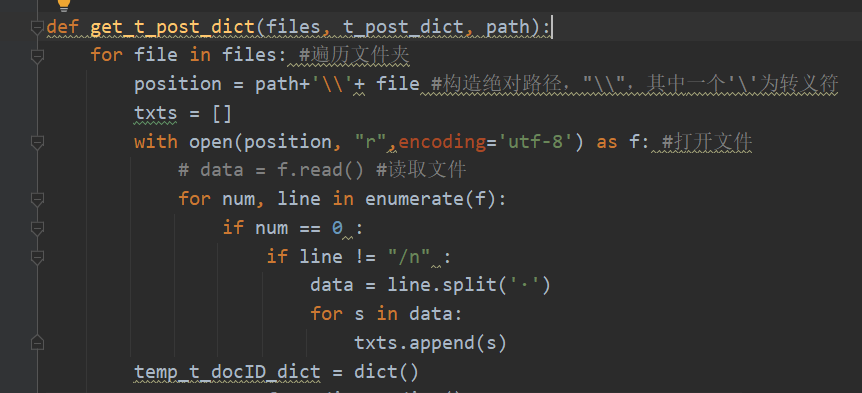
1. 获取宋词文档

手工获取宋词内容：



1. 分别提取词题与诗词

循环遍历文件，用列表存放文本：

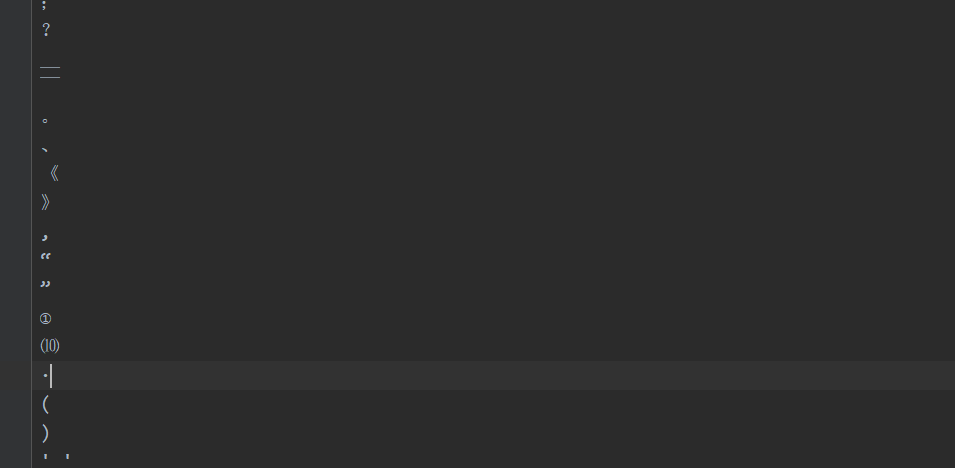


1. 使用Hanpl分词

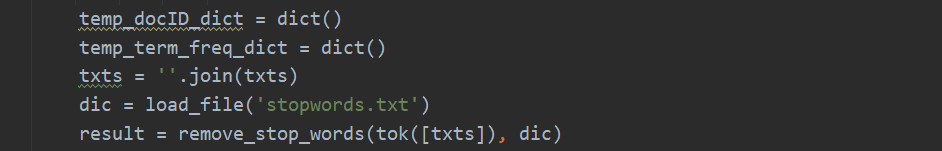
调用hanlp库：

import hanlp

载入停用词字典：



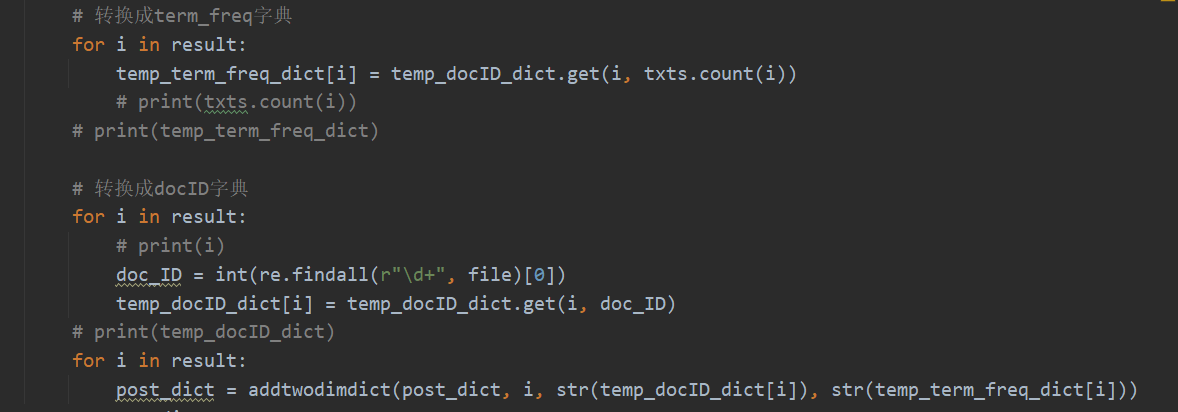
使用hanlp的tok格式对文本进行分词：



1. 构建词典

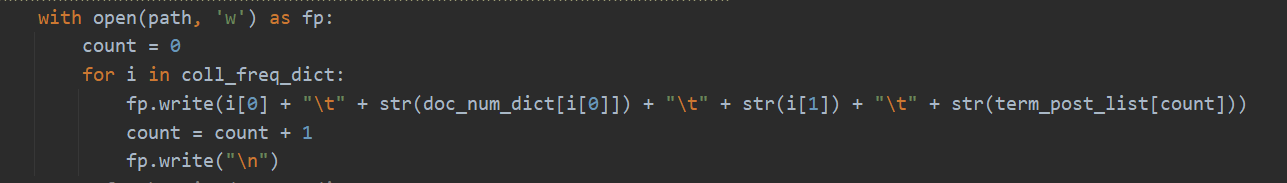
分别用两个字典来存储term\_freq与docID

再使用二维字典来存储term与<docID,term\_freq >

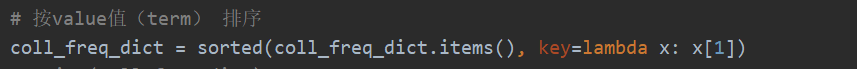


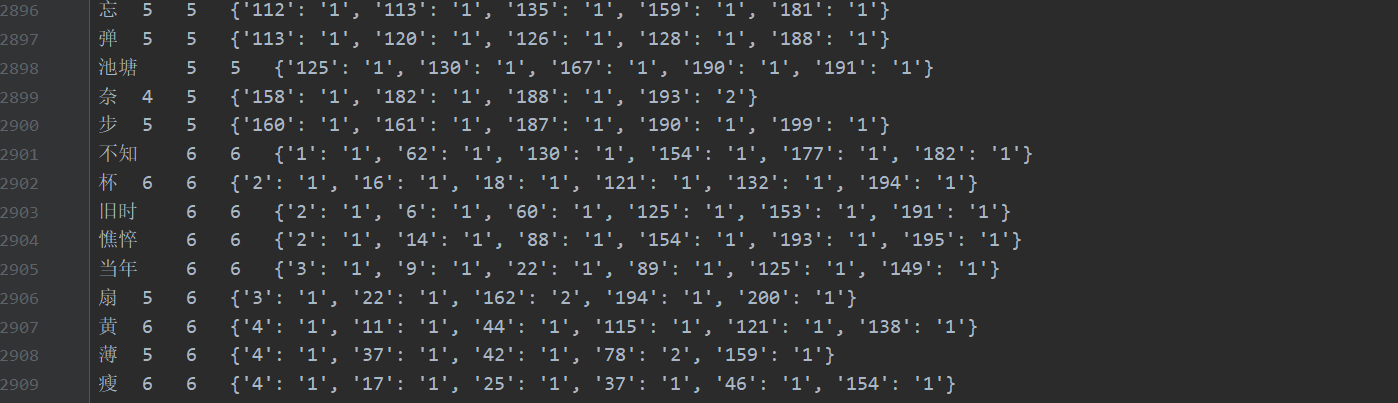
1. 构建记录表

<term, doc\_num,coll\_freq>与<term\_freq, docID>一一对应



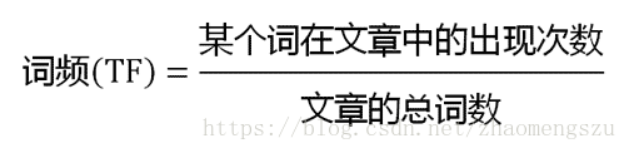
按照term的coll\_freq来排序

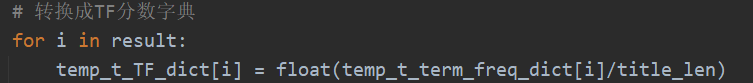


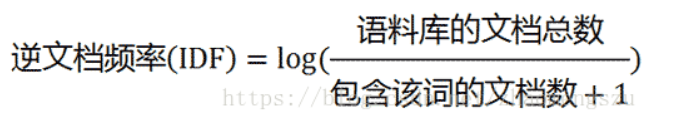


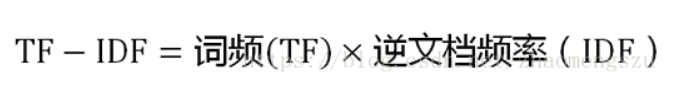
1. 构建含评分的记录表

使用TF-IDF评分法：

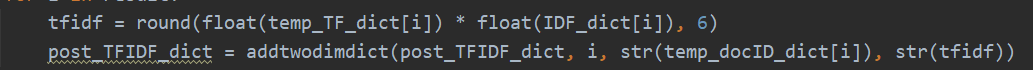




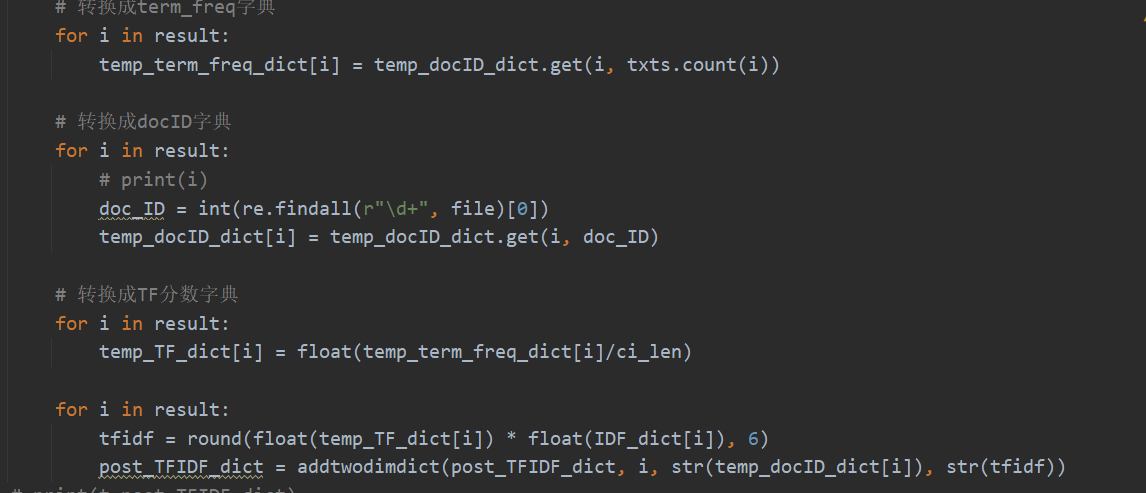




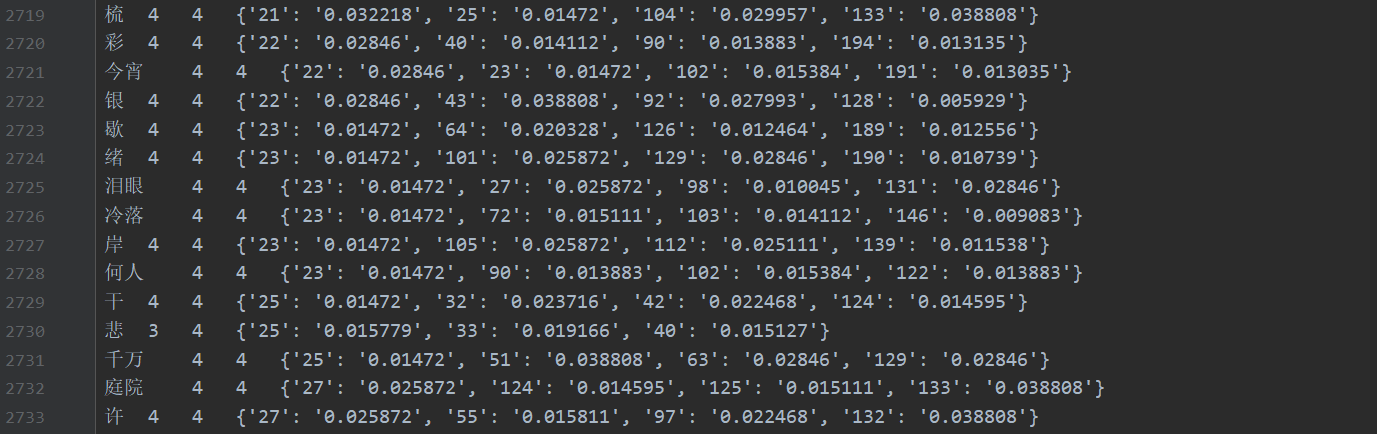
对分数保留六位小数：



用二维字典存储：

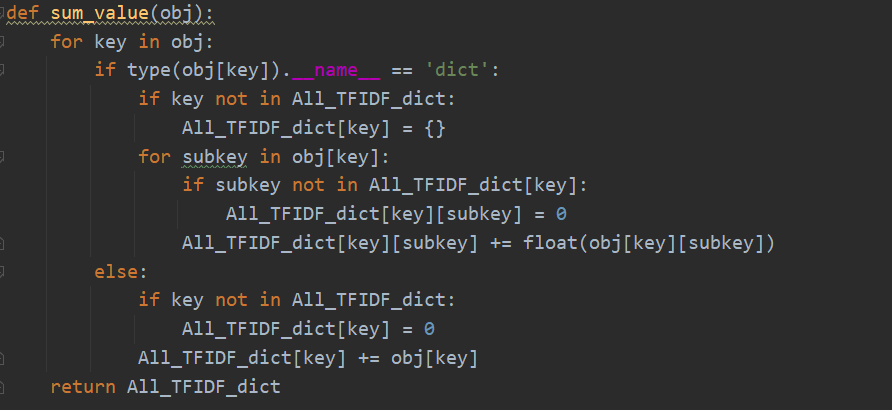


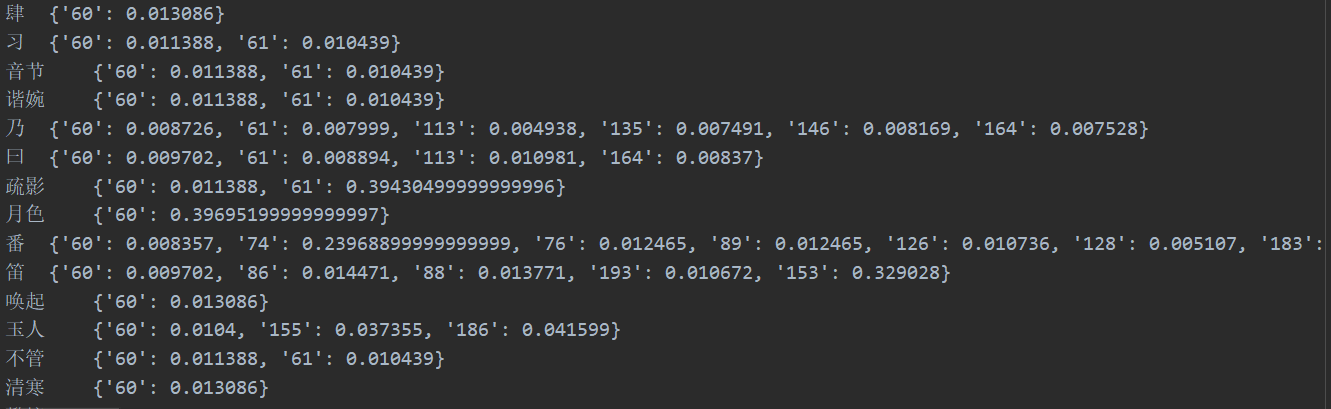
按照term的coll\_freq来排序



1. 合并词题与诗词的记录表

用二维字典存储：





1. 设计查询算法

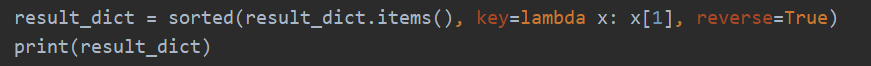
基本算法：

构建链表数据结构：

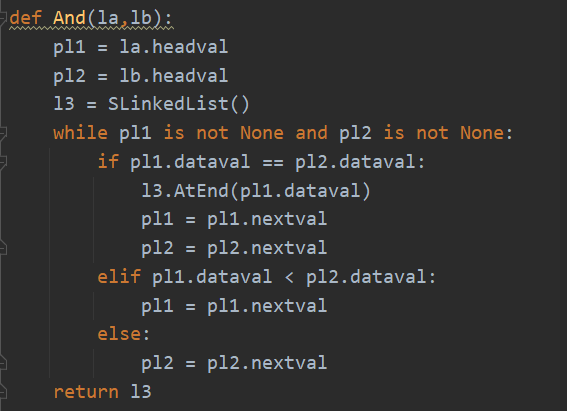
DocID均以链表形式存储：



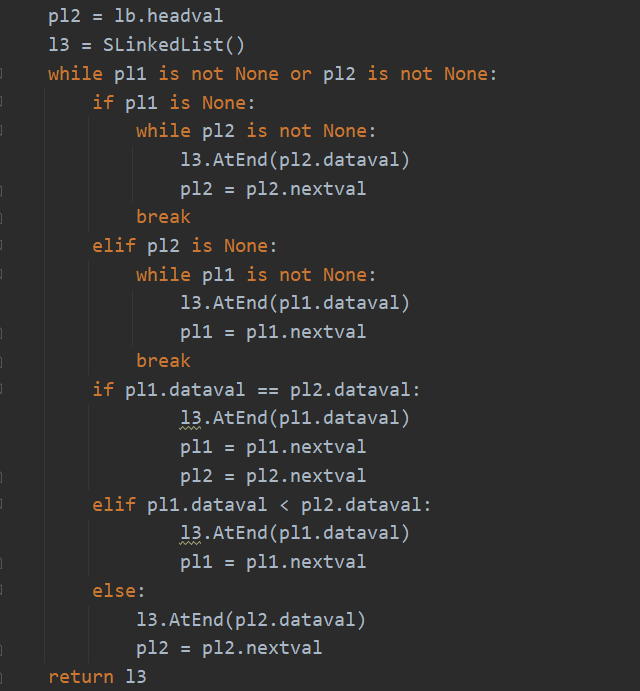
并按分数由高到低排序：



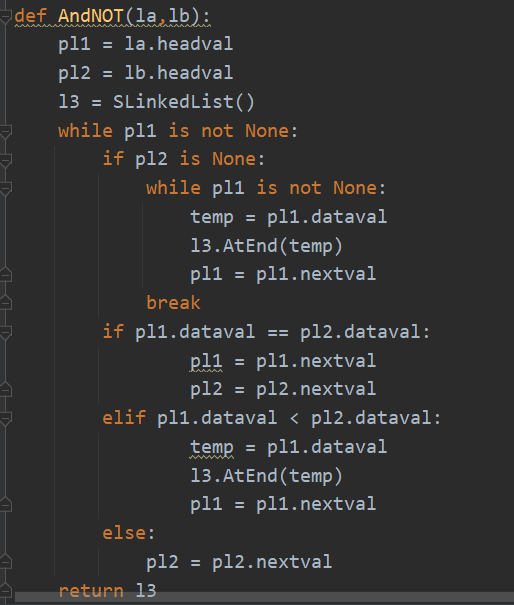
AND：



OR:



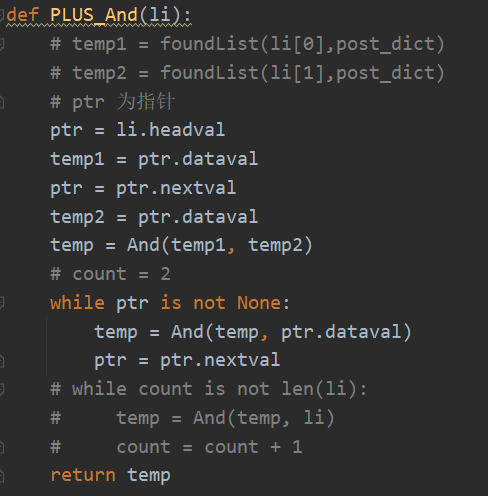
ANDNOT：



扩展算法：

PLUS\_AND：

在内部循环调用AND算法：



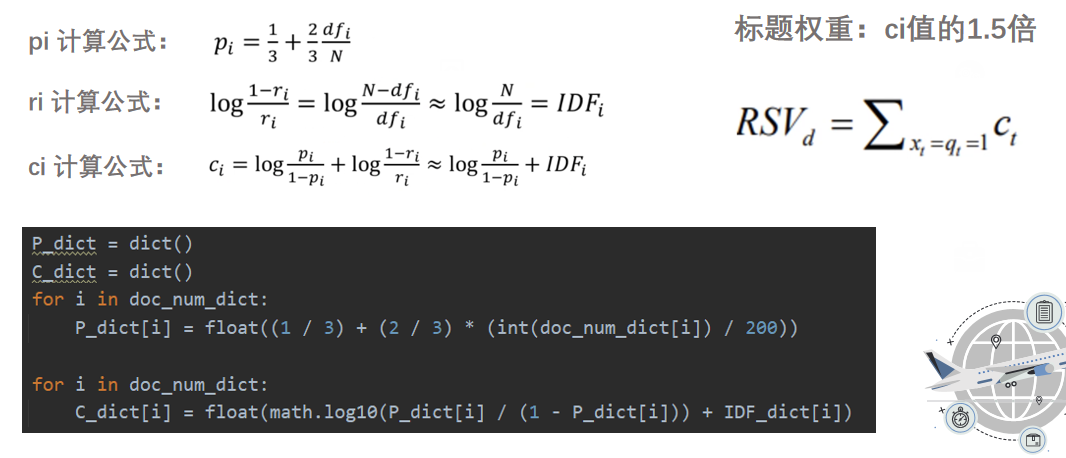
PLUS\_OR：

在内部循环调用OR算法：

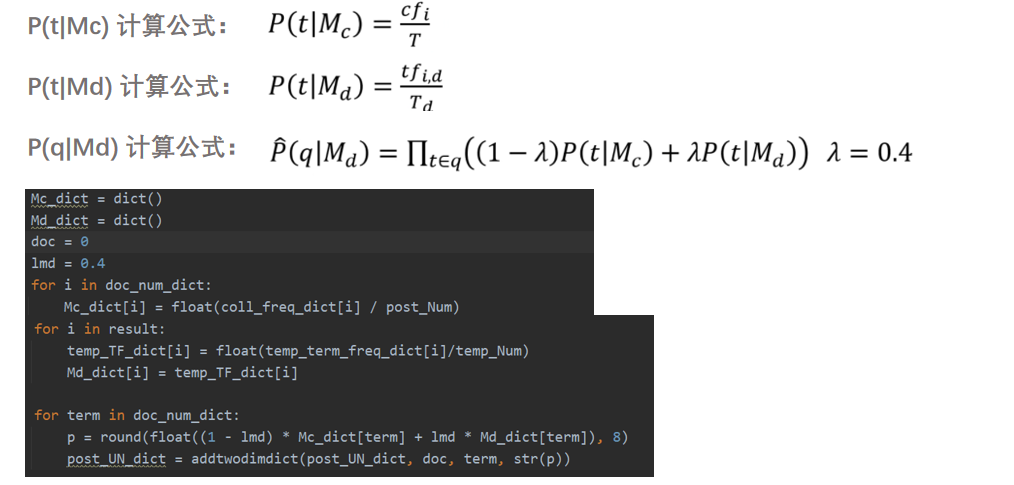


基于PL(Probabilistic Language Model)

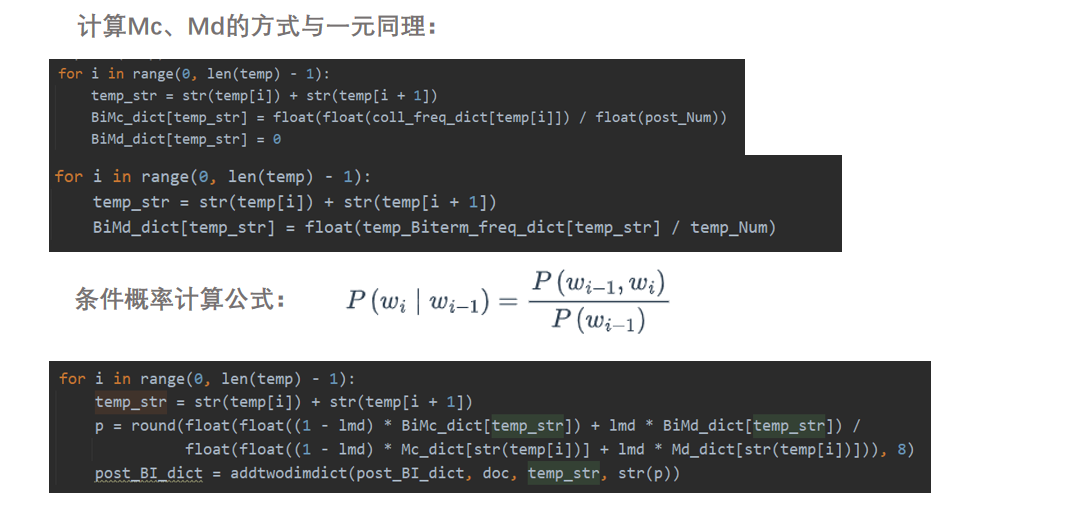
1.概率模型



2.一元语言模型



3.二元语言模型



4.top-k算法



遇到问题：

在编写布尔查询的相关操作时，发现结果与预期有较大偏差：

经过复查发现 and, or, andnot 方法中的比较函数比较的是地址的值而不是实际的值（文档编号）导致结果出现偏差，经过对值进行类型转换，解决了该问题。

1. 项目检索结果

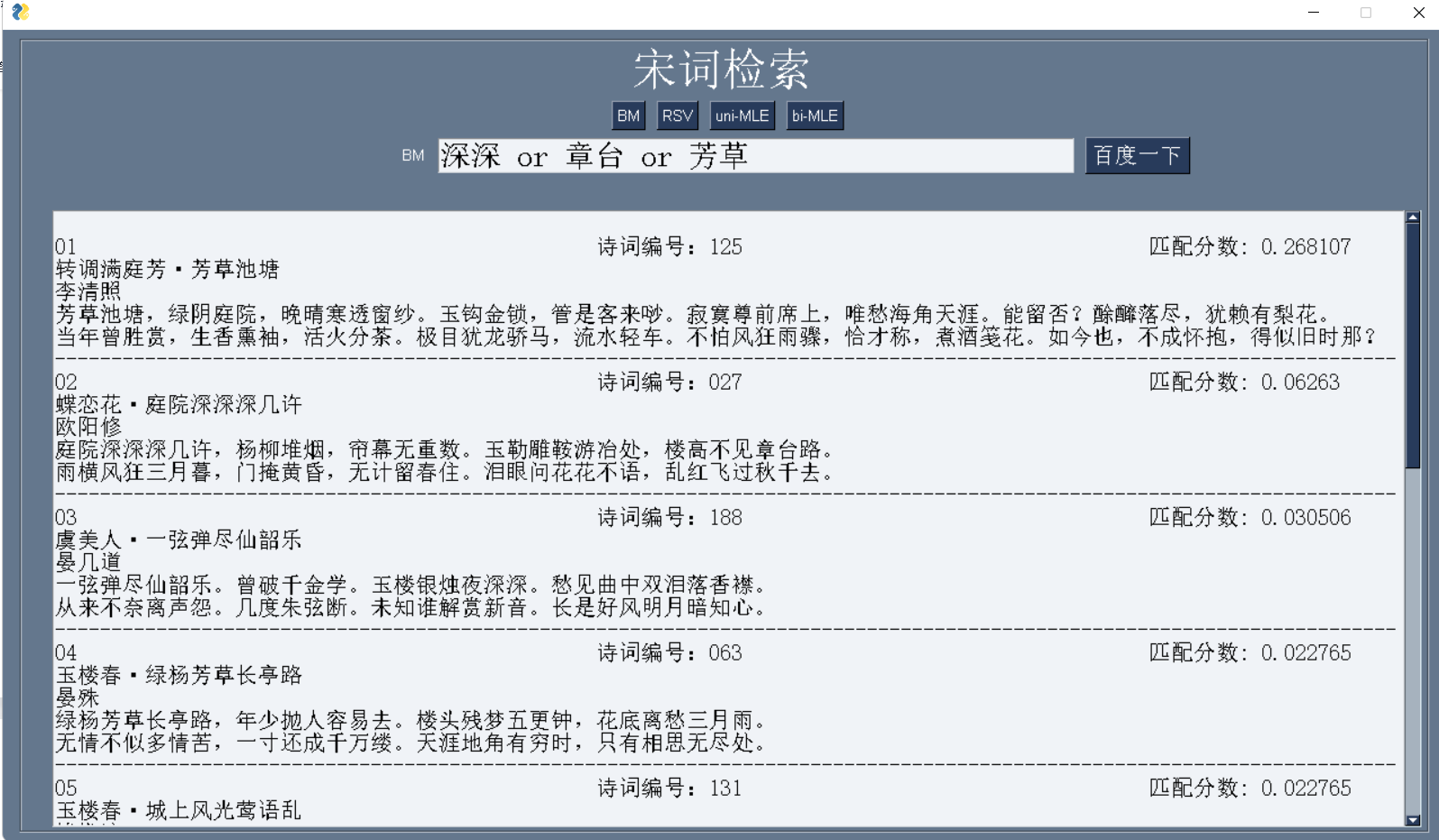
对每一种检索方式我们各寻找了10次查询，其中我们对RSV,一元语言模型，二维语言模型中进行了相同查询以对比查询结果。

BM:

一 and 曲 and 新词



深深 or 章台 or 芳草



（结果较多，按照匹配分数进行排序）

识 or 面 andnot 话



危楼 or 细细 andnot 欢



啦啦 and 啦（不存在的分词）



啦啦啦（格式有误）



深深深 and 丙辰 or 东南



东南 or 大 or学



春 or 花 or 秋月



起来 or 唯



RSV

残酒



秋千



识面话



丙辰中秋欢



啦啦啦



临江危楼



物是人非



细细危楼



曲



人人昨日



uni-gram

残酒



秋千

识面话

丙辰中秋欢

啦啦啦



临江危楼

物是人非



细细危楼



曲



人人昨日



bi-gram

残酒



秋千



识面话

丙辰中秋欢

啦啦啦



临江危楼



物是人非



曲



人人昨日



经过我们的测试，可以发现RSV搜索的结果与uni-gram的搜索结果基本一致，与bi-gram存在一定程度的差异。