中华古诗词检索系统实验报告

author：史恩扬 1711373

实验环境

Python 3.7

设计思路

将构建的dict封装在MyDictionary类中，其中的成员变量包括古诗储存地址rootdir，古诗数目EMAIL_NUM，字典长度DICTIONARY_LEN，古诗字典emails，单词字典dictionary。

利用python提供的json，jieba，collection包对于古诗进行预处理，通过json包读入json文件，通过jieba包将中文文段分词，并通过collection包进行词汇统计。

利用python提供的pyecharts包完成统计数据的可视化。

利用python中的dict数据结构构建单词字典，作者字典以及邮件字典，dict数据结构本身自带哈希表，便于之后的查找。

古诗字典中的每一元素对应一个古诗，包括该古诗所在json文件的地址以及为该古诗在json文件中的索引值，是为了之后查询时便于获得邮件原文。

单词字典中的每一元素对应一个单词（在此处通过给单词加后缀区分单词所在位置），包括该元素对应的一个列表，其中储存了该单词出现的古诗id。

作者字典中的每一元素对应一位作者，包括其所在简介的json文件名，索引以及作品数目。用于最终对作品数目统计。

利用python提供的tkinter包构建一个简单的UI界面，可完成检索的功能。

利用python提供的shelve包完成对初始化后的类内成员的保存。

在init_dictionary函数中遍历所有json文件，通过visit_file函数遍历各json文件中的古诗的各个部分，并将预处理后的单词集合通过create_dictionary函数将其插入字典中。最终将该类写入硬盘中，方便重复使用该词典。

在查询过程中，UI的初始化函数中通过读入的方式生成字典。通过按不同的按钮设置command的值，包括与命令，或命令，非命令，清除命令，等于命令。通过执行calculate函数进行计算。

通过获取选取的索引部位以及查询的信息获取对应的词条。并且与之前的生成的临时结果进行command对应的操作，并将结果更新到result中。

当执行等于命令时，通过获取的古诗id，在古诗字典中查询其所在的json文件和索引值，将其读出并显示在界面上（如果结果过多，只显示前十个）。

MyDictionary类

init(self, root, init_type)

初始化MyDictionary类

root为古诗储存的根地址

init_type为初始化方式，0表示并未创建过字典，则需要重新创建字典，1表示已经创建过字典，则直接读取之前保存的类

load__dic(self)

读取之前创建过的类

write_dic(self)

写出已创建的类

init_dictionary(self)

初始化字典，遍历所有json文件，并通过visit_file读取其中的古诗

visit_file(self, fp)

fp为json文件地址

遍历文件中的古诗，及其各个部分，进行预处理，并通过create_dictionary函数插入字典

create_dictionary(self, words, t)

words为统计后的字符集合

t为字符所在位置

查询词条是否已创建，若已创建则插入，否则，创建新词条

analysis(self)

将作者的作品数目排序，并将结果可视化

MyUI类

calculate(self)

获取选取的索引部位以及查询的信息获取对应的词条，并根据command值进行不同的操作，并更新result

c_AND(self)

更新result值并设置command值为1

c_OR(self)

更新result值并设置command值为2

c_NOT(self)

更新result值并设置command值为3

c_EQUAL(self)

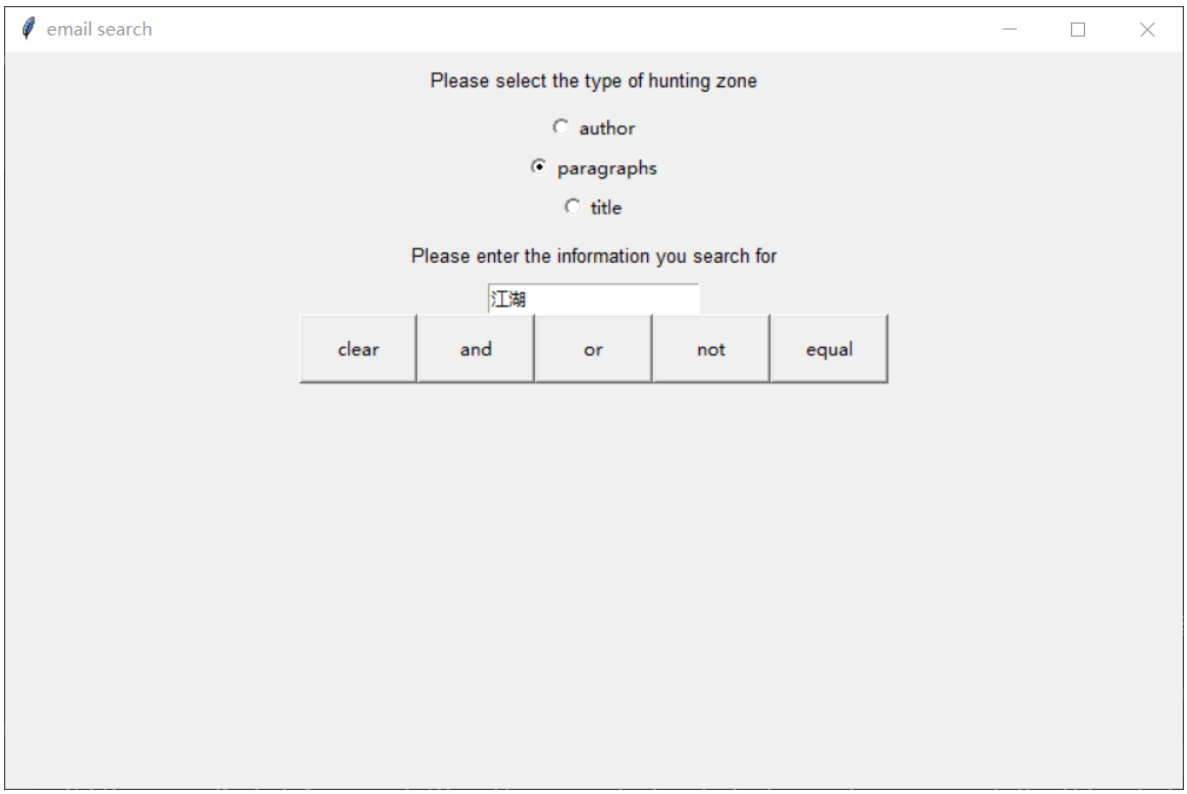
更新result值并通过古诗id，在古诗字典中查询其所在的json文件和索引值，将其读出并显示在界面上（如果结果过多，只显示前十个）。

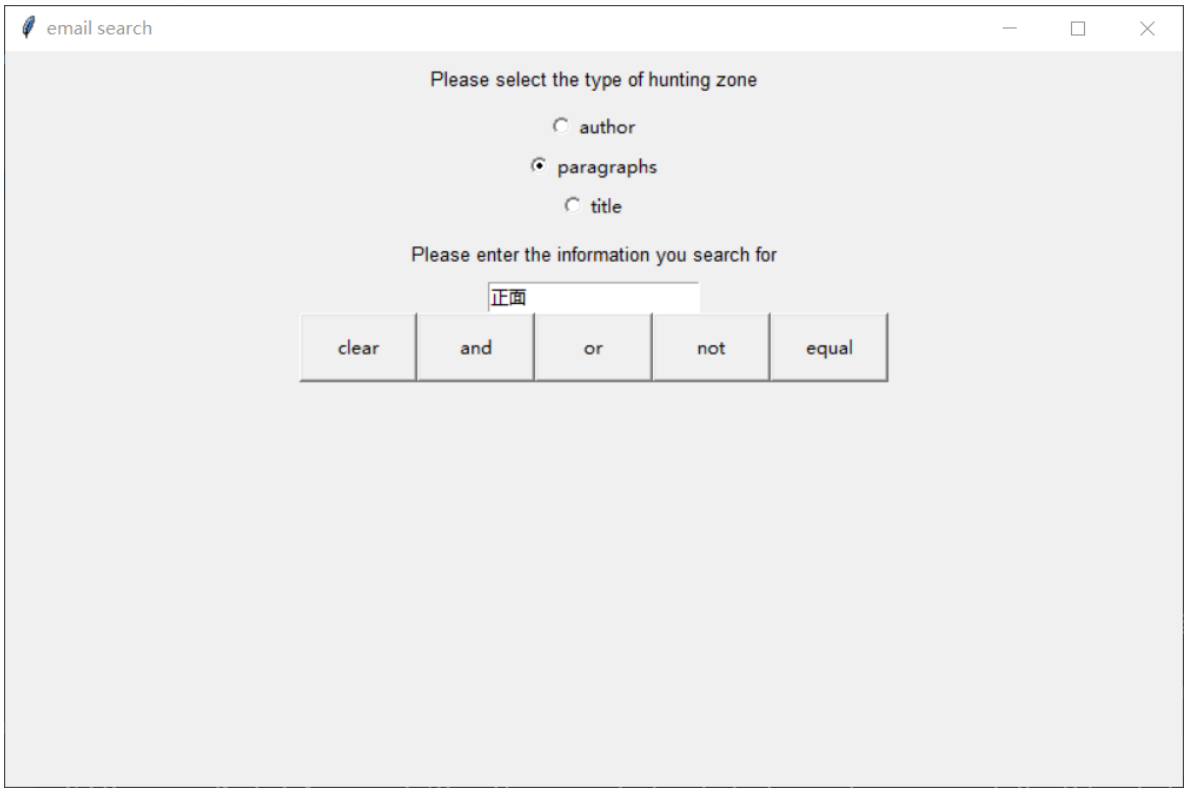
c_CLEAR(self)

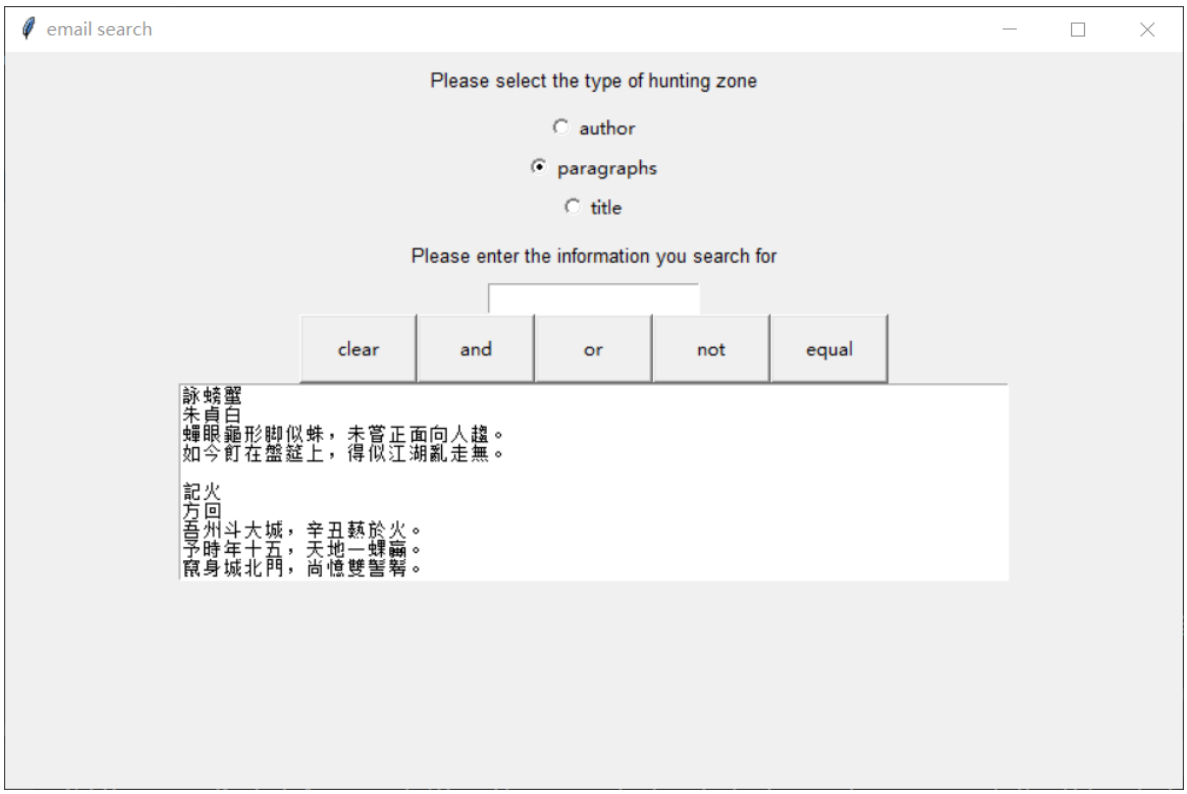
清空result，msg，text并设置command值为-1

实验结果

注意：程序内的路径为相对路径，如果要重新生成dictionary，必须将邮件放在同级文件夹下。


email search
Please select the type of hunting zone
author
paragraphs
title
Please enter the information you search for
江湖
clear
and
or
not
equal


email search
Please select the type of hunting zone
author
paragraphs
title
Please enter the information you search for
正面
clear
and
or
not
equal



搜索正文包含"江湖"且包含"正面"的古诗，结果如下

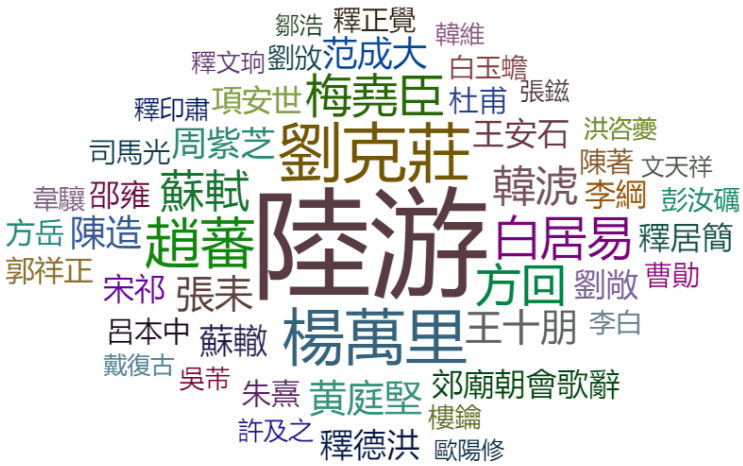
詠螃蟹
朱貞白
蟬眼龜形脚似蛛，未嘗正面向人趨。
如今飣在盤筵上，得似江湖亂走無。

記火
方回
吾州斗大城，辛丑蕘於火。
予時年十五，天地一蜾蠃。
竄身城北門，尚憶雙髻髻。
秋暑七月半，洶湧沸炎堁。
日午饑無食，枝頭得梨果。
不謂心尚孩，析薪未克荷。
先君儒起家，負冤踣奔播。
赭垣獲苟免，小宅亦云頗。
先祖之舊居，竟弗脫此禍。
二三叔父家，貲産素不夥。
焚如既已酷，生理各坎軻。
老者漸喪亡，賴嘗教敕我。
田屋悉破散，江湖走一舸。
掇科歘歷久，身幸青紫裹。
中年營一區，稍於己意可。
紫陽正面南，烏聊在其左。
前榮藝花木，後圃蒔菜蔬。
軍興盜賊起，萬卷破甑墮。
乙未九月災，天特赦么麼。
巋然七十翁，吾其理歸柁。

詠蟹
李貞白
蟬眼龜形脚似蛛，未曾正面向人趨。
如今飣在盤筵上，得似江湖亂走無。

文件 | E:/code/information_search/hw2/code/render.html

作品数目TOP50



对于作品数目前50名的作者统计（将光标移至作者名字后即可查看具体数目）

实验中的问题

在作业初期选择在词条内部存储该词所在的位置信息，发现占用大量内存，导致内存溢出，故而选则在词条处区分。

pyecharts包更新后模块使用方法改变，重新查阅资料后完成数据可视化。