HW1 布尔检索

1.作业描述

给定文档集合，使用BSBI算法实现倒排索引的构建，并使用可变长编码压缩保存到磁盘，然后实现联合查询。最后额外选择实现一种编码方式（gamma或者delta编码）

要求：截止日期为10.8。请大家在截止日期前将代码（包含运行结果，测试内容不作要求），实验报告（可单独撰写，也可整合在jupyter notebook中）一起提交到邮箱nkulxb2022@163.com，命名方式为学号_姓名_hw1。

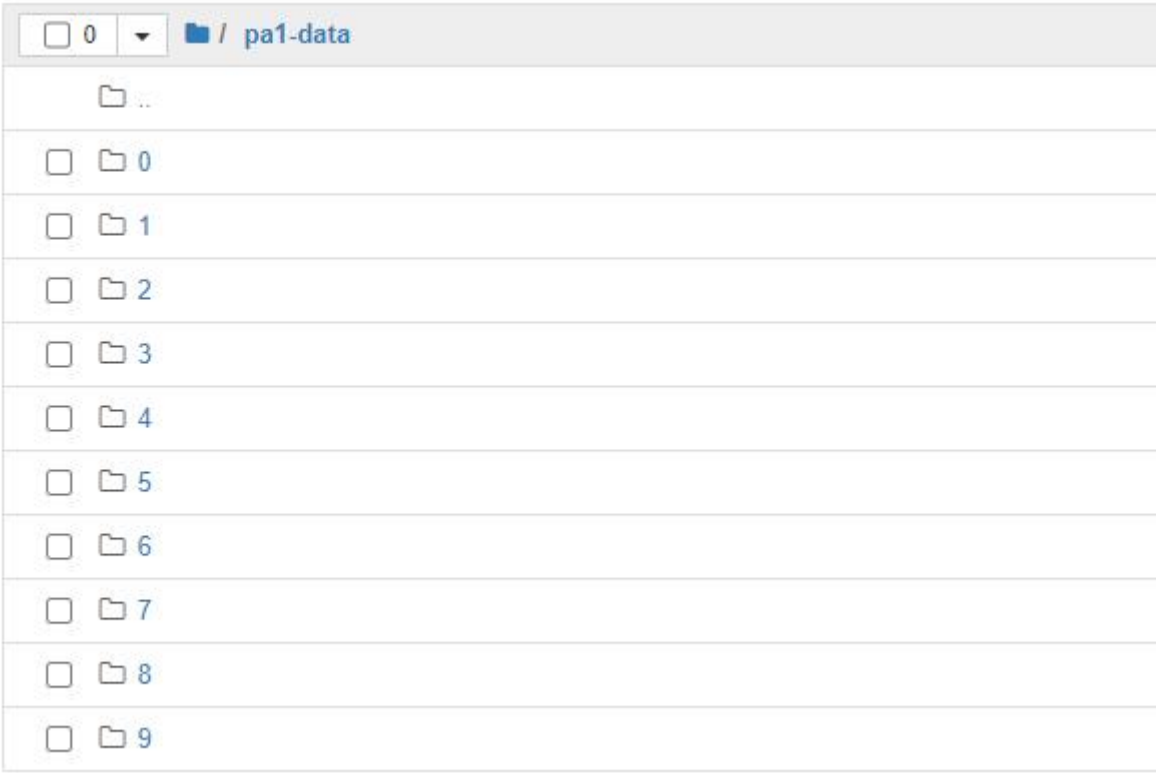
2.索引构建

倒排索引形式: 关键词---> [文档1,文档2 .....] ===》**termID---> [docID1, docID2.....]**

posting_list是倒排索引中的一条记录[docID1,docID2,.....]，以查询词term为key可以通过posting_dict中记录的位置长度等三元组信息在“索引”中找到term对应的posting_list，详情查看代码注释

BSBI算法：基于磁盘的外部排序算法

对于一个大的文档集合，将其均分为很多块（如下子目录）



对于不同的小块，我们依次进行如下流程：

```
1.词到id，文档到id的映射  字典{key:value}
    1.1实现对应代码的IdMap类
2.得到每个块的倒排索引形式 termID---> docID1, docID2.....
    2.1实现思路：先得到词-文档对：(termID,docID)，再将其“排序”汇总为  termID--->
```

```
docID1, docID2.....
    2.2主要对应代码为BSBIIndex类（继承新增的部分函数）
3.将每个块的倒排索引进行合并
    3.1需要注意的是我们合并的是已经保存在磁盘上的倒排索引，所以需要从索引文件中先读取出来，注意编码解码问题。
    3.2对应代码块：合并
```

3.索引压缩

实现可变长编码替换掉默认的UncompressedPostings类，代码补充在CompressedPostings类

4.布尔检索

思路：给定terms，获取posting_dict对应的value信息（三元组），根据三元组给定的信息从索引中找到对应的posting_list。根据与或非等查询条件将不同term查到的posting_list进行交并差操作，我们的作业只包含联合查询。

5.额外的编码方式

根据要求任选一种实现（gamma delta）

6.实验报告

不做单独要求，可以整合到jupyter notebook中，可以单独撰写。