信息系统检索

作业一：布尔检索

姓名：蒋浩南　　学号：2012948

一、实验目的：

根据给定文档集合，利用 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。并额外实现 gamma 编码。

二、实验思路：

1. BSBI 算法实现倒排索引表构建

（1） IdMap 辅助类，实现对 BSBIndexl 类所储存的 term 和 doc 的 str 和 id 转换。

（2） BSBIIndex 类下的 parse_block 函数，将每个输入路径作为子块，构建 termID 和 docID 对。实现为两层遍历，第一层遍历根据 doc 划分 docid，添加，第二层遍历该 doc，提取词项，添加 term，同时添加 termID 和 docID 对。

（3） InvertedIndexWriter(InvertedIndex)中的 append，根据要求的格式添加，词项和对应的倒排列表。

（4） BSBIIndex(BSBIIndex)下的 invert_write 函数，将解析得到的 td_pairs 转换成倒排表。遍历生成的 td_pairs,需去重。如果不是上一个 termid 且不是初始 termid，添加上一 termid 的倒排列表。是上一个 termid，直接将 docid 添加到倒排列表。

（5） InvertedIndexIterator(InvertedIndex)，根据格式实现简单的迭代。

（6） 合并：BSBIIndex(BSBIIndex)中的 merge，类似于 td_pairs 的合并。

2.布尔联合检索

（1）实现 InvertedIndex 的子类 InvertedIndexMapper，它能够找到对应 terms 在索引文件中位置并取出它的倒排记录表。

（2）sorted_intersect。经典的线性取交集，i 和 j 分别为两个列表的遍历，只有当两个列表都没遍历完的时候继续循环。分三种情况，其中两种为一大一小时，小的自增；而相等时在结果中添加，都自增。

（3）利用 sorted_intersect 和 InvertedIndexMapper 来实现 retrieve 函数。第一个 token，将该 token 的对应 id 的 index_mapper 加入。非第一个，新加入要与之前加入的取交集。

3.索引压缩

可变长编码:

对输入的数字，用一个字节存七位有效位。不断循环右移七位，知道输入的数为零。

4.额外编码方式

实现 gamma-encoding，同理。对 int 的编码，当为 0 和 1 时我们直接赋值为 0，返回。对于 gamma-encoding，分为两部分，首先去掉首位 1,计算去掉首位 1 后的位数，即 int(log(gap, 2))，int(log(gap, 2))个 1 加 0 为前半部分。去掉首位 1 后的剩余部分为第二部分，即 bin(gap)[3:]，即去掉其中'ob1'

三、说明及注意事项~~

1.对于验证。稍微更改一下 assert，改为 set(my_results) - set(reference_results)为空，且 set(reference_results) - set(my_results)为空；与原来的断言等价。因为有

个别元素的前后顺序不一样（二分人工查出来的 55555）

2. 注意建立 InvertedIndexMapper 时，一定要记得参数 postings_encoding=self.postings_encoding。更新编解码方式。(找错误找的心痛)

3.注意对于自己构建的路径在验证比较前要将 pa1-data\\处理掉。

4.可以直接看源文件中的注释。

四、实验结果

1.对未压缩的索引的验证

测试dev queries（提前构建好的查询与结果）是否正确

```
]:
for i in range(1, 9):
    with open('dev_queries/query.' + str(i)) as q:
        query = q.read()
        my_results = [os.path.normpath(path) for path in BSBI_instance.retrieve(query)]
        #print(my_results)
        print("len(my_results): ",len(my_results))
        with open('dev_output/' + str(i) + '.out') as o:
            reference_results = [os.path.normpath(x.strip()) for x in o.readlines()]
            print("len(reference_results): ",len(reference_results))
            # print(reference_results)
            set1=set(my_results) - set(reference_results)
            set2=set(reference_results) - set(my_results)
            assert set1==set() and set2==set(), "Results DO NOT match for query: "+query.strip()
        # assert my_results == reference_results, "Results DO NOT match for query: "+query.strip()
    print("Results match for query:", query.strip())

    #稍微更改一下assert，改为set(my_results) - set(reference_results)为空，且set(reference_results) - set(my_results)为空；与原来的断言等价
    #因为有及其个别元素的前后顺序不一样
    #经验证与人工比对，结果正确
```

```
<class '__main__.UncompressedPostings'>
len(my_results):  12409
```

```
<class '__main__.UncompressedPostings'>
len(my_results):  12409
len(reference_results):  12409
Results match for query: we are
<class '__main__.UncompressedPostings'>
len(my_results):  6094
len(reference_results):  6094
Results match for query: stanford class
<class '__main__.UncompressedPostings'>
len(my_results):  22335
len(reference_results):  22335
Results match for query: stanford students
<class '__main__.UncompressedPostings'>
len(my_results):  63
len(reference_results):  63
Results match for query: very cool
<class '__main__.UncompressedPostings'>
len(my_results):  81770
len(reference_results):  81770
Results match for query: the
<class '__main__.UncompressedPostings'>
len(my_results):  66675
len(reference_results):  66675
Results match for query: a
<class '__main__.UncompressedPostings'>
len(my_results):  81770
len(reference_results):  81770
Results match for query: the the
<class '__main__.UncompressedPostings'>
len(my_results):  4232
len(reference_results):  4232
Results match for query: stanford computer science
```

2. 对压缩索引的验证

```
        print( Results match for query: , query.strip())
'''


for i in range(1, 9):
    with open('dev_queries/query.' + str(i)) as q:
        query = q.read()
        my_results = [os.path.normpath(path) for path in BSBI_instance_compressed.retrieve(query)]
        #print(my_results)
        print("len(my_results): ",len(my_results))
        with open('dev_output/' + str(i) + '.out') as o:
            reference_results = [os.path.normpath(x.strip()) for x in o.readlines()]
            print("len(reference_results): ",len(reference_results))
            # print(reference_results)
            set1=set(my_results) - set(reference_results)
            set2=set(reference_results) - set(my_results)
            assert set1==set() and set2==set(), "Results DO NOT match for query: "+query.strip()
        # assert my_results == reference_results, "Results DO NOT match for query: "+query.strip()
    print("Results match for query:", query.strip())

    #稍微更改一下assert，改为set(my_results) - set(reference_results)为空，且set(reference_results) - set(my_results)为空；与原来的断言等价
    #因为有及其个别元素的前后顺序不一样
    #经验证与人工比对，结果正确
```

Out[65]:
```
'\nfor i in range(1, 9):\n    with open(\'dev_queries/query.\' + str(i)) as q:\n        query = q.read()\n        my_results = [os.path.n
ormpath(path) for path in BSBI_instance_compressed.retrieve(query)]\n        with open(\'dev_output/\' + str(i) + \'.out\') as o:\n
reference_results = [os.path.normpath(x.strip()) for x in o.readlines()]\n            assert my_results == reference_results, "Results DO
NOT match for query: "+query.strip()\n        print("Results match for query:", query.strip())\n'

<class '__main__.CompressedPostings'>
len(my_results):  12409
len(reference_results):  12409
Results match for query: we are
<class '__main__.CompressedPostings'>
len(my_results):  6094
len(reference_results):  6094
Results match for query: stanford class
<class '__main__.CompressedPostings'>
len(my_results):  22335
len(reference_results):  22335
Results match for query: stanford students
<class '__main__.CompressedPostings'>
len(my_results):  63
len(reference_results):  63
Results match for query: very cool
<class '__main__.CompressedPostings'>
len(my_results):  81770
len(reference_results):  81770
Results match for query: the
<class '__main__.CompressedPostings'>
len(my_results):  66675
len(reference_results):  66675
Results match for query: a
<class '__main__.CompressedPostings'>
len(my_results):  81770
len(reference_results):  81770
Results match for query: the the
<class '__main__.CompressedPostings'>
len(my_results):  4232
len(reference_results):  4232
Results match for query: stanford computer science
```

验证通过