信息检索与数据挖掘实验报告

基本信息：

姓名： 遒沐一

班级： 2018级数据班

学号： 201800130020

实验内容： Inverted Index and Retrieval Model

实验任务：

- 通过tweets数据集构建倒排索引表
- 实现布尔检索模型
- 实现简单查询和复杂查询

实验数据：

实验数据如下：（部分截图）

{"userName": "Mariah Peoples", "clusterNo": 82, "text": "House may kill Arizona-style immigration bill, Rep. Rick Rand says: The House is unlikely to pass the \"Ari... http://tinyurl.com/4jrjcdz", "timeStr": "Sun Jan 23 00:02:37 +0000 2011", "tweetId": "28
{"userName": "servando", "clusterNo": 35, "text": "Mourners recall Sarge Shriver's charity, idealism \n (AP): AP - R. Sargent Shriver was always an optimist, pio... http://bit.ly/gqMcdG", "timeStr": "Sun Jan 23 00:07:48 +0000 2011", "tweetId": "2896709
{"userName": "Heide Eversoll", "clusterNo": 60, "text": "Bass Fishing Techniques: 2 Fantastic Tips To Improve Your Casting Skills", "timeStr": "Sun Jan 23 00:10:05 +0000 2011", "tweetId": "28967672074993664", "errorCode": "200", "textCleaned": " ", "re
{"userName": "Ailsa Hung", "clusterNo": 97, "text": "#Financial Aid | Proper Method Of Getting Financial Aid For Education http://ping.fm/BK0R3 #applying-for-financial-aid financial-aid-essay #", "timeStr": "Sun Jan 23 00:11:03 +0000 2011", "tweetId
{"userName": "Brothy", "clusterNo": 89, "text": "Supreme Court: NASA's intrusive background checks OK http://bit.ly/hjgy9", "timeStr": "Sun Jan 23 00:13:18 +0000 2011", "tweetId": "28968479176531969", "errorCode": "200", "textCleaned": " ", "relev
{"userName": "Rich", "clusterNo": 78, "text": "The McDonalds music to fireworks is an all time low.", "timeStr": "Sun Jan 23 00:13:42 +0000 2011", "tweetId": "28968581949558787", "errorCode": "200", "textCleaned": " ", "relevance": 1}
{"userName": "Hiding in the Burgh", "clusterNo": 35, "text": "@alyce Very sweet and quiet, if not polished - Bono & Hansard at Sgt Shriver's funeral 2day: http://youtu.be/Bf14XBbcVZg (when was ...cont'd", "timeStr": "Sun Jan 23 00:17:03 +0000 2011
{"userName": "Gareth", "clusterNo": 86, "text": "So, Avon&Somerset Police have charged Vincent Tabak with the murder of Jo Yeates. I really hope they're right, otherwise his life is ruined.", "timeStr": "Sun Jan 23 00:26:18 +0000 2011", "tweetId": "28
{"userName": "Eric W Belko, Sr.", "clusterNo": 41, "text": "Hawaii Gov Waffles on Obama\u2019s Birth Certificate \u2013 Patriot Update http://t.co/1UxYa0r via @AddThis", "timeStr": "Sun Jan 23 00:31:35 +0000 2011", "tweetId": "28973080491589632"
{"userName": "Tommy McGregor", "clusterNo": 35, "text": "Ive never retweeted myself but wanted to pass on to @atu2 RT @tommymcgregor: I Want Bono To Sing At My Funeral! http://bit.ly/i0KdEn", "timeStr": "Sun Jan 23 00:38:40 +0000 2011", "t
{"userName": "Chai and News", "clusterNo": 13, "text": "OPRAH'S FAMILY SECRET | Weekly World News: The Weekly World News will not wait for the Big O to divulge her litt... http://bit.ly/id1qiO", "timeStr": "Sun Jan 23 00:38:50 +0000 2011", "tweetI
{"userName": "World News", "clusterNo": 73, "text": "Iran nuclear talks end with no agreement; US officials say six powers aligned - Washington Post: Fox News (blog)... http://bit.ly/e78uRg", "timeStr": "Sun Jan 23 00:44:48 +0000 2011", "tweetId": "28
{"userName": "Martell Thornton", "clusterNo": 26, "text": "Are Jobs Really Obama's Focus?: More job losses and unemployment figures seemingly steady at over nine percent t... http://bit.ly/el4klF", "timeStr": "Sun Jan 23 00:46:29 +0000 2011", "tweetI
{"userName": "Glenn Ellis", "clusterNo": 83, "text": "Will the cyber worm turn?: ... any kind of computer network. Manufacturers of industrial controllers and automat... http://bit.ly/gyBgeW", "timeStr": "Sun Jan 23 00:47:28 +0000 2011", "tweetId": "28
{"userName": "Glenn Ellis", "clusterNo": 83, "text": "Will the cyber worm turn?: ... any kind of computer network. Manufacturers of industrial controllers and automat... http://bit.ly/gyBgeW", "timeStr": "Sun Jan 23 00:47:28 +0000 2011", "tweetId": "28
{"userName": "gossiphut", "clusterNo": 96, "text": "James Franco Wigs Out at Sundance: Just Jared here, blogging from inside the decked out Bing Bar at the 2011... http://dlvr.it/DtfS9", "timeStr": "Sun Jan 23 00:50:22 +0000 2011", "tweetId": "289778
{"userName": "A person", "clusterNo": 86, "text": "Vincent Tabak charged with the murder of Joanna Yeates - http://newzfor.me/?cbqb", "timeStr": "Sun Jan 23 00:53:41 +0000 2011", "tweetId": "28978641706684416", "errorCode": "200", "textCleaned
{"userName": "Ladyjones74", "clusterNo": 32, "text": "Bachmann To Give Her Own State Of The Union Rebuttal http://huff.to/gMfRXk via @huffingtonpost WHY WHY WHY!!!", "timeStr": "Sun Jan 23 00:53:45 +0000 2011", "tweetId": "28978659108855
{"userName": "trevreynoNews", "clusterNo": 86, "text": "Sun: Man charged with killing Jo Yeates: POLICE tonight have charged 32-year-old Vincent Tabac with the murder http://bit.ly/ieVzdg", "timeStr": "Sun Jan 23 00:53:53 +0000 2011", "tweetId": "
{"userName": "DESIMAG.CO.UK", "clusterNo": 86, "text": "Man charged for Jo Yeaters murder - WWW.DESIMAG.CO.UK", "timeStr": "Sun Jan 23 00:53:55 +0000 2011", "tweetId": "28978703102902273", "errorCode": "200", "textCleaned": " ", "relevance

---

实验分析：

首先我们对数据分析，可以很明显的感觉到数据是json格式的，那么我们可以以json类型读出数据，然后提取想要的数据（文档id+文档内容）然后就是对于文档倒排索引和布尔检索的分析：

文档倒排索引：文档倒排索引是通过判断单词是否在文档中出现过，从而建立单词与他们对应的文档集合的索引表。

倒排索引通常通过以下几步完成:

- **(1)** 对文档进行预处理：单词变小写、去除停词、单词还原、去除标点等
- **(2)** 分割出每个文档的单词，并进行判断：
    - 若他们未在索引中出现过，那么加入这个单词作为key，并将它对应的文档列表设为仅含当前文档
    - 若他们在索引中出现过，那么将单词对应的文档列表增加一项（当前文档id）
- **(3)** 对每个单词的文档列表进行去重和排序操作（实际上倒排索引不需要这个步骤，但为了之后的操作我们对列表进行排序）

布尔检索模型：这里主要阐述and、or、not三个操作

- **AND :** and操作接收两个参数也就是 A and B，他会对A、B以及他们对应的文档列表进行一个交操作，这里介绍一种比较普及的算法：首先我们需要要求两个列表是有序（由小到大）的，然后我们需要两个指针分别指向两个文档的开始，接下来我们将通过这两个指针对文档列表进行遍历操作，遵循下列规则：
    - 如果A、B指针对应的值是相等的那么将这个值加入结果列表，并将A、B指针均向后移一位
    - 若A指针对应的值比B指针对应的值大则将B的指针向后移一位（由于有序我们可以很容易的想到在B指针之后还可能存在与A现在对应值相同的值）
    - 若B指针对应的值比A指针对应的值大则将A的指针向后移一位
    - A、B有一个遍历完则直接结束
- **OR :** or操作接收两个参数也就是 A or B，他会对A、B以及他们对应的文档列表进行一个并操作，这里同样介绍一种比较普遍的算法：首先我们需要要求两个列表是有序（由小到大）的，然后我们需要两个指针分别指向两个文档的开始，接下来我们将通过这两个指针对文档列表进行遍历操作，遵循下列规则：
    - 如果A、B指针对应的值是相等的那么将这个值加入结果列表，并将A、B指针均向后移一位
    - 若A指针对应的值比B指针对应的值大则将B的指针向后移一位、并将B对应的值加入结果列表
    - 若B指针对应的值比A指针对应的值大则将A的指针向后移一位、并将A对应的值加入结果列表
    - A、B有一个遍历完则将另一个剩下的所有元素加入结果列表
- **NOT :** not操作接收一个参数也就是 not A，他会对A以及它对应的文档列表与文档列表集合（All）的全集进行一个非操作，这里介绍一种比较普遍的算法：首先我们需要要求两个列表（A列表与全集列表）是有序（由小到大）的，然后我们需要两个指针分别指向两个文档的开始，接下来我们将通过这两个指针对文档列表进行遍历操作，遵循下列规则：
    - 如果A、All指针对应的值是相等的那么将A、All指针均向后移一位

- 若A指针对应的值比All对应的值大那么将All对应的值加入结果列表、并将All的指针向后移一位
- 若All指针对应的值比A对应的值大则将A的指针向后移一位
- 若A先遍历完那么将All剩下的加入（A一定先遍历完或者同时遍历完）

PS：and not等就是将All变为 and前面那个（或者 and (not)），这里不做赘述、or not先当与or (not)即限制性not再执行or，在实现代码中我军采用了比较简单的处理方式，也就是将not单独看待

为了应对复杂查询我们引入了后缀表达式，这是由于计算机无法通过中缀表达判断出操作数与操作符的对应关系，我们遍历分割后的操作列表并通过如下方法去构建后缀表达式（首先建立两个栈结果栈（ans）与符号栈（op）：

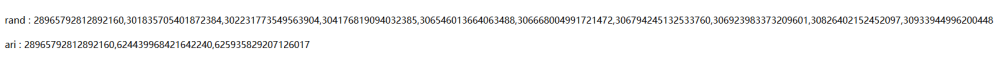
- 如果当前item不是操作符则直接加入ans栈
- 除左括号之外均要弹出一个运算符号(op栈非空则弹出并加入ans栈）），且不能弹出左括号,若为not不执行这个弹出操作，最后将当前操作符加入ans栈
- 如果当前字符为一切其他操作符；若op栈空，直接入op栈，若为右括号则弹出并加入ans栈，直到遇到左括号或op栈空，此时op栈不空再弹一个并加入ans栈
- 表达式遍历完了，但是栈中还有操作符不满足弹出条件，把栈中的东西全部弹出

PS:这么操作的原因是通过不断的分析得出的，这里给一个例子：

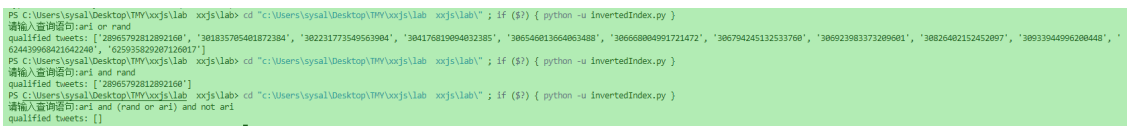
["ari","and","(","rand","or","rand",")"]->['ari', 'rand', 'rand', 'or', 'and'] 我们会发现操作符只取他前面两个对应的列表（或者就是列表），若有not则只看他前面那个，就可以得到最后的结果并且与我们对中缀的理解相同。

实验结果与分析：

测试数据：



测试检索与结果：



这里选用了统计出的两个比较短的word以及他们对应的文档列表（见测试数据）

其次通过两个简单查询（一个and和一个or）以及一个复杂查询（and or not与括号的配合）总共三个query来检测程序的正确性

比较明显的是第三个测试数据（挑选了必为空值的复杂查询）：通过括号内的or以及外部的and、and not来判断程序对复杂查询的适应性，实际上速度还是蛮快的，结果也是完全符合预期结果的

收获与反思：本次实验遇到了许多困难，一开始对于后缀表达式的错误理解导致复杂查询直接崩掉，最后总结出了上面叙述过的构建方法，还有一开始对于数据结构的选择也进行了多次的试错，不过在努力之后还是达到了比较满意的效果，对于倒排索引和布尔查询也有了更加深入的理解，但仍有一丝丝遗憾，由于自己的电脑出了问题，实验室电脑又没有nltk的环境因此本来想做的单词还原并没有实现、其次对于查询的优化也并没有思考到一个比较好的解决方案因此也没有实现。

PS:实验代码注释较多这里不做分析

Code