实验一描述文档

实验目的

本实验要求以给定的邮件数据集为基础，实现一个邮件搜索引擎。对于给定的查询，能够以精确查询或模糊语义匹配的方式返回最相关的一系列邮件文档。

任务描述

本实验要求同时实现`bool检索`和`语义检索`。首先将每一封邮件视作一个文档，进行分词、词根化、去停用词处理。词根化的过程与停用词库不作硬性要求，可以以你认为合适的方式任意选取。完成上述初始化步骤后，你需要：

- 1. 对于经过预处理的文档集合$D=\{D_1,D_2,\ldots,D_N\}$, 根据倒排索引算法建立倒排索引表S，并以合适的方式存储生成的倒排索引文件。
  2. 对于给定的bool查询 Q_{bool} (Q_{bool}的书写规则以上课内容为准)，根据你生成的倒排索引表S, 返回符合查询规则 Q_{bool} 的文档集合$D_{bool}=\{D_1^{bool},D_2^{bool},\ldots,D_M^{bool}\}$
  3. 根据文档集合D, 计算每个文档的`tdf-idf`向量 $T=\{v_1,v_2,\ldots,v_N\}$，并将T以矩阵的形式存储。
  4. 对于给定的语义查询 $Q_{se}=\{word_1,word_2,\ldots,word_n\}$，其中每个$word_i$代表一个查询词， 计算$Q_{se}$的`tf-idf`向量$v_{qse}$，并根据$v_{qse}$ 与 T的相似度返回前10个最相关的文档集合 $D_{se}=\{D_1^{se},D_2^{se},\ldots,D_{10}^{se}\}$。

除此之外，可选做的内容包括：

- 1. 对你的倒排索引过程进行时间复杂度或者是空间复杂度的优化。
  2. 采用外部知识库 (例如同义词表) 优化你的索引效果。
  3. 采用`word2vec`等其他语义表征方式表征你的查询和文档，并选用合适的案例与`tf-idf`的结果进行对比分析。

对于实现的优化我们将视优化效果给予酌情加分。

数据集介绍

`Enron Email Datase`是目前在电子邮件相关研究中使用最多的公开数据集,其邮件数据是安然公司(Enron Corporation,原是世界上最大的综合性天然气和电力公司之一,在北美地区是头号天然气和电力批发销售商)150位高级管理人员的往来邮件。这些邮件在安然公司接受美国联邦能源监管委员会调查时被其公布到网上。

本地可用的数据介绍:

- 1. 数据内容为：Enron公司的150名员工每人有一个对应的文件夹，文件夹内是其文件的分组，里面存放他们的往来邮件，全部文件大小约1.7G。
  2. 每一封邮件是一个txt文档在该文档中,包含了每封邮件包括邮件头、正文在内的详细内容(不包含邮件附件)。如下图所示:

```
Message-ID: <7707807.1075852367961.JavaMail.evans@thyme>
Date: Mon, 10 Sep 2001 06:07:04 -0700 (PDT)
From: m..forney@enron.com
To: joe.capasso@enron.com
Subject: RE: Ercot Portal Mismatch
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: Forney, John M. </O=ENRON/OU=NA/CN=RECIPIENTS/CN=JFORNEY>
X-To: Capasso, Joe </O=ENRON/OU=NA/CN=RECIPIENTS/CN=Jcapasso>
X-cc:
X-bcc:
X-Folder: \JFORNEY (Non-Privileged)\Forney, John M.\Sent Items
X-Origin: FORNEY-J
X-FileName: JFORNEY (Non-Privileged).pst

Thanks -
Good job in documentation

JMF

 -----Original Message-----
From:   Capasso, Joe
Sent:   Saturday, September 08, 2001 2:58 AM
To: Forney, John M.; Saibi, Eric; McElreath, Alexander; Olinde Jr., Steve
Subject:    Ercot Portal Mismatch



  John,

I had a mismatch problem in the portal that I think you and the group should know about.

I completed a trade with Calpine tonight (09/08/01), I purchased 50mw for HE 3 & 4 and
quickly input it in the portal.  At approximately 1:50am, Calpine still did not have their side
of the trade input, so I gave Scott at Calpine a quick call.  He was in the process
of inputting the information in the system and I could see the Calpine entries turning from
red to black.

All of the entries matched up correctly except for the 300 interval and this is where
the problem begins.  Scott confirmed the trade verbally and I, also, had the same information
(i.e. - 50mw, South Zone, etc...).  Even though, everything was the same in both portals, it remained red
and unmatched.
```

3. 数据集已被上传到睿客网，下载网址为：链接：https://rec.ustc.edu.cn/share/fab7a6f0-1364-11eb-abcf-e1b50652fbb2 密码：1uka
4. 查询词表的下载链接：https://rec.ustc.edu.cn/share/c76b8710-1437-11eb-906b-05ad53331ecc 密码：im8t

提交要求

请以如下文件目录结构组织相关文件结构：

```
exp1/
|----src/
     |----bool_search.{FileSuffix}
     |----semantic_search.{FileSuffix}
|----dataset/
     |----{your dataset}
|----output/
     |----{your output files}
|----实验报告.pdf
|----README
```

其中，各目录/文件具体要求如下：

- `src`目录下放置你的源代码文件，其中`bool_search`为以bool检索方式的源代码文件；`semantic_search`为以语义检索方式；`.{FileSuffix}`根据你使用的编程语言的文件后缀而决定。如果存在相应的优化，请直接添加在原文件中。
- `dataset`目录下放置你的数据集文件。在提交时，可以将此文件夹置空。
- `output`目录放置你的输出文件，包括生成的`倒排表`，以及经过所有文档的`tf-idf`矩阵。如果采用了其他的语义表征方式，也请生成对应的矩阵并在`README`中注明。
- `实验报告.pdf`应包含对你采用的算法以及所作优化的描述，同时请用相关的实验数据证明你的优化是有效的。同时，请在实验报告中展示你认为最具有代表性的运行示例。如果是多人组队，请在实验报告中注明所有组成员的学号和姓名。
- `README`文件中包含你的源代码的运行环境，编译运行方式，以及对关键函数的说明。同时，对于所作要求之外的文件，也请在`README`中注明这些文件的含义。