实验(一)——邮件搜索引擎

> Member：童云林 PB17071444，杨朝辉 PB17071433
> Method：基于倒排表的布尔检索，TF-iDF向量空间模型语义检索
> Platform：Python 3.7.6
> Date：2020-11-22

实验算法

倒排表查询

- Inverted Index 构造
    1. 检索每篇文档，获得 < 词项，文档ID > 对，并写入临时索引
    2. 对临时索引中的词项进行排序
    3. 遍历临时索引，对于相同词项的文档ID进行合并
- Bool Retrieval
    1. 对检索命令拆分为str类型列表并解析：将其中的词条作同样的词根化处理，将整个中缀表达式按照优先级为{'not': 3, 'and': 2, 'or': 1}转换为后缀表达式
    2. 根据以上后缀表达式各个词条对应的倒排表中的文档ID集合做集合运算
    3. 返回结果各文档ID对应的文件名称（绝对路径名，默认前十个结果显示）

语义查询

- TF-iDF 矩阵构造
    1. 在扫描处理所有文档时得到DF向量、TF矩阵
    2. 按照如下公式构建TF-iDF矩阵

$$W_{t,d}=\begin{cases}(1+\lg(tf_{t,d}))\log\left(\frac{N}{df_t}\right), tf_{t,d}>0\\ 0, tf_{t,d}=0\end{cases}$$

- 语义查询（VSM）
    1. 对查询命令按照（ascii码排序的）词条表做向量化，得到query_arr
    2. 将query_arr、tfidf各列向量作模归一化处理，query_arr与tfidf各列作点积，结果按照降序排列取前10结果对应的下标（文档ID）
    3. 返回结果各文档ID对应的文件名称（绝对路径名，默认前十个结果显示）

问题与优化

- 运行时间

  由于文档扫描处理、倒排表构建、TF-iDF矩阵构建过程均比较耗时，故而采用文档频率前1000的词条作为待检索对象，同时程序运行中目测基于所有文档（50多万个邮件文件）构建倒排表及TF-iDF矩阵需要6~7个小时运行，故而最终只根据前50000个文档（考虑到的时间原文档数量的十分之一）实现该搜索引擎。

- 存储空间

  按照矩阵大小计得基于所有文档所有词条生成的对象所需存储空间数十GB，故而采用稀疏矩阵存储TF、TF-iDF矩阵等，倒排表由于长度限于1000，直接以Python中的字典数据存储数据；他们都作为程序中DocumentProcessed类的属性，检索器所依赖的该类实例通过二进制序列化在内存与硬盘间进行存取，最终生成于 ./output 目录下基于前50000个文档的对象文件仅 170MB。

结果显示

布尔检索

1. 该结果表明 England 该词并不在文档频率前1000之列，故而没有检索到可用文档。

```
Input your query command: team and financial and company and meeting and market and Enron and price and england
There is no document satisfying your query requirement!
```

2. 该结果显示了 37 个满足检索要求的文档，实际上布尔表达式各个单词都取自 E:\VSCode\Python\webinfo-assignment\exp1\dataset\maildir\brawner-s\deleted_items\243 文件中，结果也确实返回了该文档名称。

```
Input your query command: team and financial and company and meeting and market and Enron and price and performance
Results (absolute file path) is/are:
E:\VSCode\Python\webinfo-assignment\exp1\dataset\maildir\arnold-j\notes_inbox\67
E:\VSCode\Python\webinfo-assignment\exp1\dataset\maildir\arnold-j\notes_inbox\68
E:\VSCode\Python\webinfo-assignment\exp1\dataset\maildir\brawner-s\deleted_items\243
E:\VSCode\Python\webinfo-assignment\exp1\dataset\maildir\arnold-j\deleted_items\69
E:\VSCode\Python\webinfo-assignment\exp1\dataset\maildir\arnold-j\inbox\54
E:\VSCode\Python\webinfo-assignment\exp1\dataset\maildir\brawner-s\sent_items\74
E:\VSCode\Python\webinfo-assignment\exp1\dataset\maildir\arnold-j\_sent_mail\622
E:\VSCode\Python\webinfo-assignment\exp1\dataset\maildir\arnold-j\deleted_items\702
E:\VSCode\Python\webinfo-assignment\exp1\dataset\maildir\arnold-j\deleted_items\198
E:\VSCode\Python\webinfo-assignment\exp1\dataset\maildir\arnold-j\deleted_items\439
There is/are still 27 results not shown
```

语义检索

为表示检索结果的准确性，选取邮件内容较少的某邮件 E:\VSCode\Python\webinfo-assignment\exp1\dataset\maildir\baughman-d\all_documents\190，将其中的两段话作为检索命令，检索结果中该文档名称排名确实靠前。

```
188
189
190
193
194
195
196
197
198
199
200
201

Two of TradersNews Energy's?three new hourly  indexes,?the TVA and ComEd
hubs, are attached. We will distribute these two indexes to you at no charge
via e-mail  during the beta testing period, which?concludes March 19. On
that  date,?we will formally launch the indexes on our Web site, at:?
www.tradersnewsenergy.com.

If you are interested in receiving the ERCOT hourly  index during this month
of testing,?e-mail Suzanna Strangmeier, at suzanna.strangmeier@ipgdirect.com,
 or call her at 713/647-7325.
```

```
Input your query command: Two of TradersNews Energy's?three new hourly  indexes,?the TVA and ComEd hubs, are attached. W
e will distribute these two indexes to you at no charge via e-mail  during the beta testing period, which?concludes Marc
h 19. On that  date,?we will formally launch the indexes on our Web site, at:? www.tradersnewsenergy.com. If you are int
erested in receiving the ERCOT hourly  index during this month of testing,?e-mail Suzanna Strangmeier, at suzanna.strang
meier@ipgdirect.com, or call her at 713/647-7325
Results (absolute file path) is/are:
E:\VSCode\Python\webinfo-assignment\exp1\dataset\maildir\baughman-d\discussion_threads\187
E:\VSCode\Python\webinfo-assignment\exp1\dataset\maildir\baughman-d\all_documents\190
E:\VSCode\Python\webinfo-assignment\exp1\dataset\maildir\baughman-d\power\cinergy_index\27
E:\VSCode\Python\webinfo-assignment\exp1\dataset\maildir\baughman-d\discussion_threads\182
E:\VSCode\Python\webinfo-assignment\exp1\dataset\maildir\baughman-d\all_documents\185
E:\VSCode\Python\webinfo-assignment\exp1\dataset\maildir\baughman-d\power\cinergy_index\30
E:\VSCode\Python\webinfo-assignment\exp1\dataset\maildir\baughman-d\all_documents\197
E:\VSCode\Python\webinfo-assignment\exp1\dataset\maildir\baughman-d\discussion_threads\194
E:\VSCode\Python\webinfo-assignment\exp1\dataset\maildir\baughman-d\power\cinergy_index\26
E:\VSCode\Python\webinfo-assignment\exp1\dataset\maildir\baughman-d\discussion_threads\211
```