Web信息处理与应用 实验报告

lab1

团队成员

- 何春望 PB17000075
- 吴健宗 PB17000082

实验内容

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**和**模糊语义匹配**的方式返回最相关的一系列邮件文档。

算法描述

- 精确查询（bool 查询）：
  对于给定的 bool 查询，如（term1 AND term2 OR NOT term3），根据倒排索引表，返回符合查询的所有文档。
- 模糊语义匹配：
  对文档集合建立 tf-idf 矩阵，对输入的语义查询 Q，计算 Q 的 tf-idf 向量 v，并找出 v 与文档集合中的文档相似度最高的文档子集。

优化内容

时间复杂度优化

- 跳表指针
  - bool 查询中，对AND，OR的处理，使用了**跳表指针**技术，减少了时间复杂度。

空间复杂度优化

- 减小 tf-idf **矩阵精度**
  - 对 tf-idf 矩阵的设置，从 numpy 默认的 float64 精度减小到 float16 精度。因为在文档数量，单词数量有限的情况下，tf-idf矩阵的可能值实际上是有限的、离散的，不需要很高的精度就可以达到筛选文档相似度最高的文档子集的目标。这样做减少了 tf-idf 矩阵存储的空间复杂度。
  - 对所有文档建立 tf-idf 矩阵时，如果使用 float16 的精度，只需要900MB-1000MB的空间，如果使用 float64，则需要四倍左右的空间。

> 注：实验结果表示，在倒排索引，tf-idf矩阵，单词索引三个数据库中，tf-idf 矩阵占用空间远大于其他两个，因此减少 tf-idf 矩阵的空间大小是最有必要的。

- **去邮件头**处理
  - 针对 enron dataset 的文档格式，我们做了特殊的"去邮件头"处理。具体是：对每一个邮件，先去掉一些对查询帮助不大的内容，如：邮件发送人，接受人，邮件版本(Mime-Version)，日期，邮件编号等。这样做可以减少内存消耗，并大大加快后期的处理速度，同时减小倒排索引表，tf-idf 矩阵等数据库的大小。
  - 经粗略统计，enron dataset 中，每条邮件大概有三分之二是"无效信息"，因此先把它们去掉是很有帮助的。

> 注：去邮件头处理在去除邮件编号，邮件版本等"冗余信息"的同时，貌似还去掉了日期，发件人等在实际应用中可能有效的信息。考虑到本实验的查询词是针对邮件正文进行查询，因此在本实验的查询词范围内，这些信息是无效的；如果需要查询这些信息，在去邮件头处理中不对它们做去除处理即可。

- 去停用词处理
  - 本实验中，我们采用了去停用词处理以减少倒排表和 tf-idf 文档的大小。
  - 我们同时采用了**两个停用词表**，一个是通用英文停用词表，包含 I, you, do 等常用停用词。另一个是针对 enron dataset 的特殊停用词表，包含 enron, com, html 等在该数据集中频繁出现的无效词语。

查询结果优化

- 词根化处理
  - 本实验中，我们调用了 nltk 库对邮件中的单词进行词根化。这样做有利于避免因为词性转化而导致搜索不到目标词的问题。如，用户需要搜索"message"，但某个文档中出现的是"messages"，即"message"的复数形式。如果不做词根化处理，搜索引擎则不会把他们识别为同一个词，导致遗漏的情况。
  - 另外，我们实现了三种不同的词根化方案，可以满足对构建时间和准确度不同的权衡与取舍要求，极大地提高了系统的灵活性。

实验结果

- 构建索引时间：3h
- 查询延迟：
  - bool 检索：<1s
  - 语义检索：约10s
- tf-idf矩阵占用空间： 约1GB

运行示例

build inv-index and tf-idf matrix

运行 `build_index_and_matrix.py` ，大约三个小时后，得到结果：

```
17385 517386 517387 517388 517389 517390 517392 517396 517400 517401
['pleas' 'power' 'energi' 'market' 'compani' 'time' 'thank' 'messag'
 'mail' 'price' 'corp' 'gas' 'forward' 'call' 'inform' 'day' 'trade'
 'origin' 'week' 'servic' 'meet' 'busi' 'deal' 'report' 'california'
 'develop' 'schedul' 'chang' 'attach' 'imag']
[[0.     0.     0.     ... 0.     0.347  0.347 ]
 [0.     0.     0.     ... 0.     0.     0.9297]
 [0.     0.     0.     ... 0.     0.     0.    ]
 [0.8584 0.8584 0.     ... 0.     0.     0.    ]
 [0.     0.     0.     ... 0.     0.     0.    ]]
build time used: 10771.067316293716 s
```

以下是文档频率排行前十的结果，可以看到，基本去除了冗余的**停用词**。

```
pleas 232694
power 60808
energi 63864
market 71675
compani 61716
time 125017
thank 191584
messag 129731
mail 86285
price 59774
```

bool search

- 选用查询词：power, company, customers, employees, president
- 查询用例
  > 注：查询结果中文档的排序是按照文件名字典序排序的

  - **power AND company AND customers AND employees AND president**
    执行结果：
    ```
    (base) D:\课程\Web信息处理与应用\lab\lab1\src>python bool_search.py
    Please input your bool query (words connected with AND, OR, NOT): power AND company AND customers AND employees AND president
    bool search time: 0.7252500057220459 s
    total number of results:  1816
    first 100 results:  [3654, 3655, 3915, 3931, 4228, 4249, 4384, 4402, 4424, 4440, 4510, 4546, 4591, 4759, 4765, 4810, 4812, 5084
    , 5089, 5189, 5190, 5335, 5370, 5387, 5507, 5508, 6115, 6130, 7014, 7685, 7701, 8799, 9239, 9272, 9963, 11171, 12396, 12576, 14
    252, 14700, 18316, 19336, 21796, 23443, 24294, 24512, 25752, 28501, 28509, 28513, 28563, 28938, 29004, 32542, 32825, 32879, 329
    11, 32984, 33011, 33268, 33273, 33589, 33689, 33777, 33784, 35935, 36868, 37393, 38937, 40063, 40175, 41785, 43247, 43472, 4400
    4, 44025, 44577, 44644, 44796, 44907, 46812, 49116, 49121, 49122, 49132, 49135, 49353, 49435, 49500, 49505, 49506, 49515, 49531
    , 49533, 51975, 52220, 52323, 52735, 53507, 54145]
    first ten file path:
    ('../dataset/maildir\\arnold-j\\all_documents\\605_',)
    ('../dataset/maildir\\arnold-j\\all_documents\\606_',)
    ('../dataset/maildir\\arnold-j\\all_documents\\846_',)
    ('../dataset/maildir\\arnold-j\\all_documents\\860_',)
    ('../dataset/maildir\\arnold-j\\deleted_items\\179_',)
    ('../dataset/maildir\\arnold-j\\deleted_items\\198_',)
    ('../dataset/maildir\\arnold-j\\deleted_items\\323_',)
    ('../dataset/maildir\\arnold-j\\deleted_items\\33_',)
    ('../dataset/maildir\\arnold-j\\deleted_items\\35_',)
    ('../dataset/maildir\\arnold-j\\deleted_items\\374_',)
    ```

可以看到，对五个词用AND连起来查询，一共用了0.72s，查询到1816个结果。

下面是第一个结果的部分内容(../dataset/maildir/arnold-j/all_documents/605_)

```
Message-ID: <19189634.1075857603655.JavaMail.evans@thyme>
Date: Mon, 14 May 2001 01:30:00 -0700 (PDT)
From: ann.schmidt@enron.com
Subject: Enron Mentions - 05/12/01 - 05/13/01
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: Ann M Schmidt
X-To:
X-cc:
X-bcc:
X-Folder: \John_Arnold_Jun2001\Notes Folders\All documents
X-Origin: Arnold-J
X-FileName: Jarnold.nsf

As Final Exams Begin, Power Is a Big Question
The New York Times, 05/13/01


British Telecom
The Times of London, 05/12/01


Houston needs to think small about future technology
Houston Chronicle, 05/13/01


Panel plots new course for area's future / Education, economics, quality of
life top group's list of needed improvements
Houston Chronicle, 05/13/01


MSEB not to pick up 15 pc in DPC after phase II completion
Press Trust of India Limited, 05/13/01


Enron plans to pull out of Gulf gas project: MEED
Agence France-Presse, 05/13/01


SMALL BUSINESS / Pleasure cruisin' / Yacht fleet owner offers customers what
amounts to limo service on the lake
Houston Chronicle, 05/13/01


More power to reform agenda
The Economic Times, 05/13/01


India Power Min: New Power Deal With Enron Unit Possible
Dow Jones International News, 05/12/01


India: Talks begin on Dabhol issue
Business Line (The Hindu), 05/12/01
```
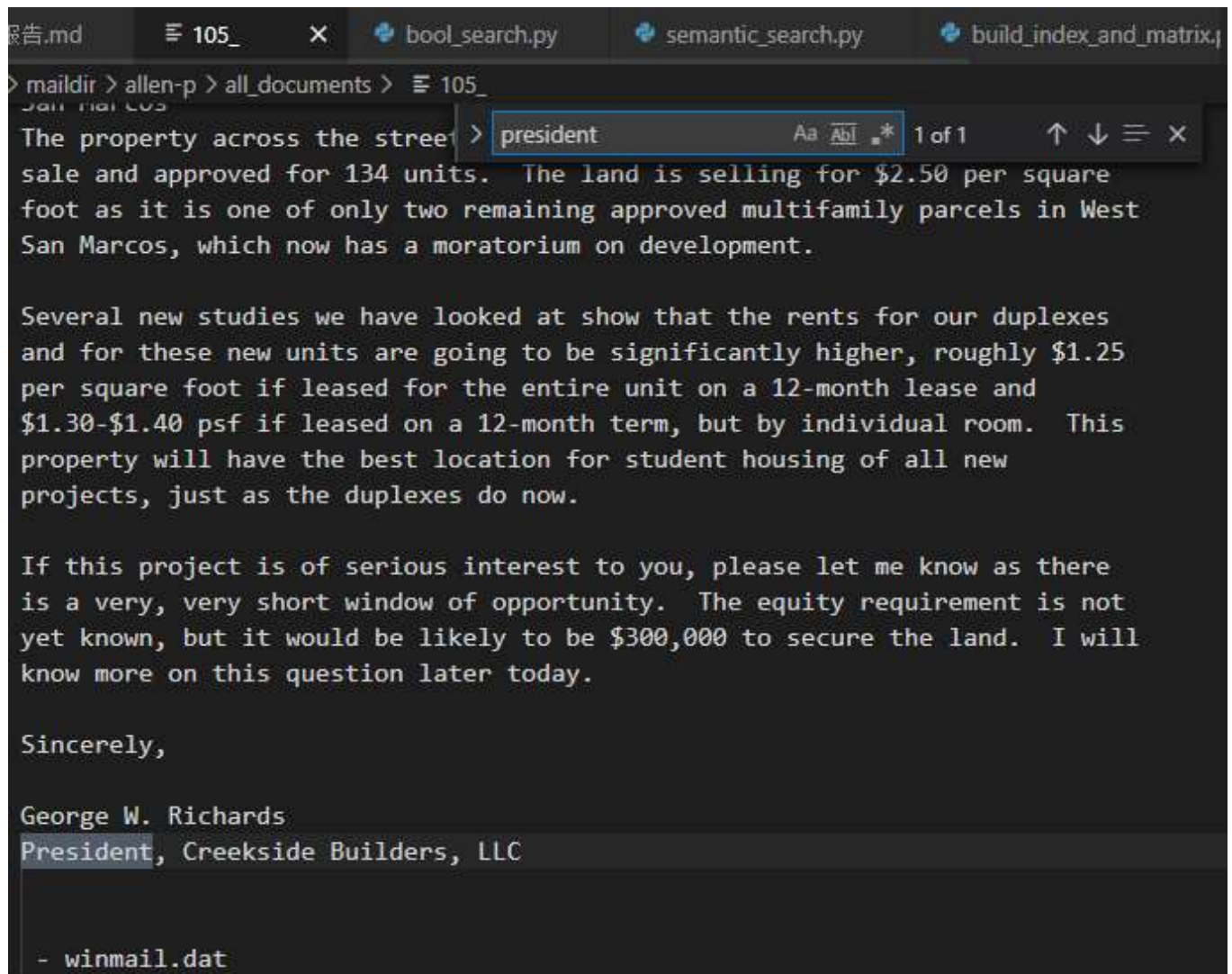
这是一封长邮件，里面涵盖了全部五个搜索词，正因为搜索词要求五个词都在，所以几乎只有这种长邮件才满足要求。

- **power AND company OR customers AND employees OR president**
  执行结果：

```
(base) D:\课程\Web信息处理与应用\lab\lab1\src>python bool_search.py
Please input your bool query (words connected with AND, OR, NOT): power AND company OR customers AND employees OR president
bool search time: 0.8492903709411621 s
total number of results:  37912
first 100 results:  [6, 11, 14, 15, 49, 109, 146, 154, 278, 337, 394, 404, 409, 484, 503, 507, 525, 616, 631, 632, 646, 656, 709, 712, 732, 749, 768, 771, 773, 784, 786, 804, 817, 923, 935, 941, 943, 950, 1017, 1047, 1048, 1053, 1064, 1094, 1126, 1136, 1183, 1196, 1199, 1212, 1263, 1267, 1316, 1351, 1370, 1423, 1432, 1441, 1447, 1467, 1470, 1478, 1505, 1522, 1614, 1622, 1786, 1852, 1856, 1929, 1936, 1949, 1951, 1971, 2032, 2040, 2042, 2043, 2084, 2120, 2242, 2252, 2253, 2305, 2404, 2449, 2456, 2457, 2494, 2587, 2595, 2751, 2817, 2821, 2838, 2895, 2914, 2916, 2917, 2935]
first ten file path:
('../dataset/maildir\\allen-p\\all_documents\\105_',)
('../dataset/maildir\\allen-p\\all_documents\\10_',)
('../dataset/maildir\\allen-p\\all_documents\\112_',)
('../dataset/maildir\\allen-p\\all_documents\\113_',)
('../dataset/maildir\\allen-p\\all_documents\\146_',)
('../dataset/maildir\\allen-p\\all_documents\\1_',)
('../dataset/maildir\\allen-p\\all_documents\\234_',)
('../dataset/maildir\\allen-p\\all_documents\\241_',)
('../dataset/maildir\\allen-p\\all_documents\\355_',)
('../dataset/maildir\\allen-p\\all_documents\\408_',)
```

讲其中两个AND换成OR，搜索结果数量增加为37912个

```
president                    Aa Abl .*  1 of 1    ↑ ↓ ≡ ×
```

The property across the stree sale and approved for 134 units. The land is selling for $2.50 per square foot as it is one of only two remaining approved multifamily parcels in West San Marcos, which now has a moratorium on development.

Several new studies we have looked at show that the rents for our duplexes and for these new units are going to be significantly higher, roughly $1.25 per square foot if leased for the entire unit on a 12-month lease and $1.30-$1.40 psf if leased on a 12-month term, but by individual room. This property will have the best location for student housing of all new projects, just as the duplexes do now.

If this project is of serious interest to you, please let me know as there is a very, very short window of opportunity. The equity requirement is not yet known, but it would be likely to be $300,000 to secure the land. I will know more on this question later today.

Sincerely,

George W. Richards
President, Creekside Builders, LLC

- winmail.dat

查看第一个文档，其中含有 president 这个词，结果正确

- **power AND company AND NOT (customers OR employees OR president)**
  执行结果：

```
(base) D:\课程\Web信息处理与应用\lab\lab1\src>python bool_search.py
Please input your bool query (words connected with AND, OR, NOT): power AND company AND NOT (customers OR employees OR president)
bool search time: 1.18554306603027344 s
total number of results:  8172
first 100 results:  [154, 484, 773, 817, 943, 1212, 1622, 1929, 2084, 2242, 2252, 2253, 2595, 2895, 3155, 3287, 3491, 3658, 4230, 4
282, 4297, 4300, 4397, 4517, 4592, 4822, 4974, 5187, 5227, 5391, 5511, 5743, 6323, 6352, 6523, 6630, 6675, 6835, 6850, 7314, 7676,
7920, 8006, 8015, 8016, 8018, 8029, 8060, 8084, 8095, 8124, 8128, 8139, 8171, 8180, 8182, 8323, 8354, 8488, 8558, 8610, 8617, 8625,
 8740, 8778, 8820, 9034, 9042, 9049, 9136, 9154, 9264, 9265, 9267, 9314, 9332, 9457, 9486, 9487, 9488, 9492, 9515, 9518, 9552, 9553
, 9554, 9556, 9558, 9559, 9578, 9582, 9583, 9638, 9639, 9643, 9646, 9655, 9676, 9729, 9738]
first ten file path:
('../dataset/maildir\\allen-p\\all_documents\\241_',)
('../dataset/maildir\\allen-p\\all_documents\\540_',)
('../dataset/maildir\\allen-p\\deleted_items\\238_',)
('../dataset/maildir\\allen-p\\deleted_items\\344_',)
('../dataset/maildir\\allen-p\\deleted_items\\52_',)
('../dataset/maildir\\allen-p\\discussion_threads\\427_',)
('../dataset/maildir\\allen-p\\sent\\195_',)
('../dataset/maildir\\allen-p\\sent\\473_',)
('../dataset/maildir\\allen-p\\sent_items\\104_',)
('../dataset/maildir\\allen-p\\sent_items\\248_',)
```

含有 power 和 company 的同时不含有后面三个词，查询结果有8172个

```
> maildir > allen-p > all_documents >  ☰ 241_
Message-ID: <3536013.1075855670762.JavaMail.evans@thyme>
Date: Thu, 22 Jun 2000 00:26:00 -0700 (PDT)
From: phillip.allen@enron.com
To: scott.carter@chase.com
Subject: Re: The New Power Company
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: Phillip K Allen
X-To: "Carter, Scott" <Scott.Carter@chase.com> @ ENRON
X-cc:
X-bcc:
X-Folder: \Phillip_Allen_Dec2000\Notes Folders\All documents
X-Origin: Allen-P
X-FileName: pallen.nsf

Scott,

 I emailed your question to a friend that works for the new company.  I think
I know the answer to your questions but I want to get the exact details from
him.  Basically, they will offer energy  online at a fixed price or some
price that undercuts the current provider.  Then once their sales are large
enough they will go to the wholesale market to hedge and lock in a profit.
The risk is that they have built in enough margin to give them room to manage
the price risk.  This is my best guess.  I will get back to you with more.

Phillip
```

第一个文档中，含有 Power, Company, 同时不含有其他三个词

- **NOT power AND NOT company AND NOT customers AND NOT employees AND NOT president**

  执行结果：

```
(base) D:\课程\Web信息处理与应用\lab\lab1\src>python bool_search.py
Please input your bool query (words connected with AND, OR, NOT): NOT power AND NOT company AND NOT customers AND NOT employees AND
 NOT president
bool search time: 2.1534695625305176 s
total number of results:  373787
first 100 results:  [3, 4, 5, 7, 8, 9, 10, 12, 13, 16, 17, 18, 19, 21, 22, 23, 24, 25, 26, 27, 29, 30, 31, 32, 33, 34, 35, 36, 37,
38, 39, 40, 41, 42, 43, 44, 45, 47, 48, 50, 51, 54, 55, 56, 59, 60, 61, 62, 63, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77,
 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 11
0, 111, 112, 113, 114, 115, 116, 117]
first ten file path:
('../dataset/maildir\\allen-p\\all_documents\\102_',)
('../dataset/maildir\\allen-p\\all_documents\\103_',)
('../dataset/maildir\\allen-p\\all_documents\\104_',)
('../dataset/maildir\\allen-p\\all_documents\\106_',)
('../dataset/maildir\\allen-p\\all_documents\\107_',)
('../dataset/maildir\\allen-p\\all_documents\\108_',)
('../dataset/maildir\\allen-p\\all_documents\\109_',)
('../dataset/maildir\\allen-p\\all_documents\\110_',)
('../dataset/maildir\\allen-p\\all_documents\\111_',)
('../dataset/maildir\\allen-p\\all_documents\\114_',)
```

搜索不含有这五个词的结果，一共有373787个

- **power OR company OR customers OR employees OR president**

  执行结果：

```
(base) D:\课程\Web信息处理与应用\lab\lab1\src>python bool_search.py
Please input your bool query (words connected with AND, OR, NOT): power OR company OR customers OR employees OR president
bool search time: 0.9252369403839111 s
total number of results:  143614
first 100 results:  [1, 2, 6, 11, 14, 15, 20, 28, 46, 49, 52, 53, 57, 58, 64, 108, 109, 136, 146, 147, 152, 154, 174, 176, 194, 207
, 242, 244, 246, 263, 268, 272, 273, 276, 277, 278, 279, 280, 281, 282, 292, 293, 295, 297, 304, 305, 306, 308, 313, 314, 319, 320,
 337, 338, 352, 360, 361, 370, 388, 390, 394, 398, 399, 404, 409, 410, 423, 424, 427, 430, 439, 441, 445, 470, 479, 484, 485, 490,
502, 503, 507, 509, 518, 525, 533, 554, 561, 582, 616, 617, 624, 625, 631, 632, 633, 635, 640, 642, 645, 646]
first ten file path:
('../dataset/maildir\\allen-p\\all_documents\\100_',)
('../dataset/maildir\\allen-p\\all_documents\\101_',)
('../dataset/maildir\\allen-p\\all_documents\\105_',)
('../dataset/maildir\\allen-p\\all_documents\\10_',)
('../dataset/maildir\\allen-p\\all_documents\\112_',)
('../dataset/maildir\\allen-p\\all_documents\\113_',)
('../dataset/maildir\\allen-p\\all_documents\\118_',)
('../dataset/maildir\\allen-p\\all_documents\\125_',)
('../dataset/maildir\\allen-p\\all_documents\\143_',)
('../dataset/maildir\\allen-p\\all_documents\\146_',)
```

至少包含一个搜索词的结果，有143614个

semantic search

- 选用查询词：opportunity, management, address, offer, price
- 查询用例
  - **opportunity**

    执行结果：

```
(base) D:\课程\Web信息处理与应用\lab\lab1\src>python semantic_search.py
Please input your semantic query (words seperated with space): opportunity
('../dataset/maildir\\zipper-a\\sent_items\\63_',) cosine similarity:  0.7962848297213623
('../dataset/maildir\\arnold-j\\deleted_items\\396_',) cosine similarity:  0.6300832925036747
('../dataset/maildir\\mcconnell-m\\_sent_mail\\79_',) cosine similarity:  0.4908396946564885
('../dataset/maildir\\mcconnell-m\\all_documents\\210_',) cosine similarity:  0.4908396946564885
('../dataset/maildir\\mcconnell-m\\sent\\80_',) cosine similarity:  0.4908396946564885
('../dataset/maildir\\zipper-a\\deleted_items\\36_',) cosine similarity:  0.46243093922651934
('../dataset/maildir\\ybarbo-p\\inbox\\184_',) cosine similarity:  0.35290889132821074
('../dataset/maildir\\cuilla-m\\sent_items\\115_',) cosine similarity:  0.2904245709123758
('../dataset/maildir\\arnold-j\\sent_items\\578_',) cosine similarity:  0.27835497835497836
('../dataset/maildir\\cuilla-m\\sent_items\\116_',) cosine similarity:  0.2679166666666667
semantic search time:  10.424912452697754 s
```

输出十个匹配度最高的结果

查看第一个结果，短邮件，含有opportunity

- **opportunity, management**

  执行结果：

分别展现第一和第十的结果

```
> maildir > lavorato-j > all_documents >  ≡ 698_
Message-ID: <16305534.1075845545222.JavaMail.evans@thyme>
Date: Wed, 27 Dec 2000 23:42:00 -0800 (PST)
From: john.lavorato@enron.com
To: rick.buy@enron.com, ted.murphy@enron.com
Subject:
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: John J Lavorato
X-To: Rick Buy, Ted Murphy
X-cc:
X-bcc:
X-Folder: \John_Lavorato_Oct2001\Notes Folders\All documents
X-Origin: LAVORATO-J
X-FileName: jlavora.nsf

How's that for VAR management.

Lavo
```

```
Message-ID: <10905805.1075843958489.JavaMail.evans@thyme>
Date: Mon, 7 May 2001 00:11:00 -0700 (PDT)
From: mike.mcconnell@enron.com
To: morten.pettersen@enron.com
Subject: Re: Tokyo opportunities
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: Mike McConnell
X-To: Morten E Pettersen
X-cc:
X-bcc:
X-Folder: \Mark_McConnell_June2001\Notes Folders\Sent
X-Origin: MCCONNELL-M
X-FileName: mmcconn.nsf

I look forward to it.
m
```

分别是含有一个 management 和一个 opportunities 的邮件

含有 management 的邮件排行高于含有 opportunities 的邮件的原因是 management 的文档频率低于 opportunity，搜索引擎认为频率越低的词含有更丰富的信息。

另外，在排行第十的结果中， opportunity 以复数形式存在，这里体现了**去词根化**的结果，可以将一个词的不同形式都识别出来。

- **opportunity, management, address**

  执行结果：

  ```
  (base) D:\课程\Web信息处理与应用\lab\lab1\src>python semantic_search.py
  Please input your semantic query (words seperated with space): opportunity management address
  ('../dataset/maildir\\lavorato-j\\all_documents\\698_',) cosine similarity: 1.0770101925254814
  ('../dataset/maildir\\lavorato-j\\deleted_items\\190_',) cosine similarity: 1.0770101925254814
  ('../dataset/maildir\\lavorato-j\\sent\\574_',) cosine similarity: 1.0770101925254814
  ('../dataset/maildir\\lavorato-j\\sent_items\\48_',) cosine similarity: 1.0770101925254814
  ('../dataset/maildir\\carson-m\\all_documents\\125_',) cosine similarity: 0.9885496183206107
  ('../dataset/maildir\\carson-m\\all_documents\\61_',) cosine similarity: 0.9885496183206107
  ('../dataset/maildir\\carson-m\\discussion_threads\\83_',) cosine similarity: 0.9885496183206107
  ('../dataset/maildir\\carson-m\\sent\\123_',) cosine similarity: 0.9885496183206107
  ('../dataset/maildir\\carson-m\\sent\\59_',) cosine similarity: 0.9885496183206107
  ('../dataset/maildir\\carson-m\\_sent_mail\\118_',) cosine similarity: 0.9885496183206107
  semantic search time: 12.275743246078491 s
  ```

  展现第一和第十的结果

  ```
  > maildir > lavorato-j > all_documents >  ≡ 698_
  Message-ID: <16305534.1075845545222.JavaMail.evans@thyme>
  Date: Wed, 27 Dec 2000 23:42:00 -0800 (PST)
  From: john.lavorato@enron.com
  To: rick.buy@enron.com, ted.murphy@enron.com
  Subject:
  Mime-Version: 1.0
  Content-Type: text/plain; charset=us-ascii
  Content-Transfer-Encoding: 7bit
  X-From: John J Lavorato
  X-To: Rick Buy, Ted Murphy
  X-cc:
  X-bcc:
  X-Folder: \John_Lavorato_Oct2001\Notes Folders\All documents
  X-Origin: LAVORATO-J
  X-FileName: jlavora.nsf


  How's that for VAR management.


  Lavo
  ```

```
> maildir > carson-m > _sent_mail > ≡ 118_
  Message-ID: <18922134.1075858201276.JavaMail.evans@thyme>
  Date: Thu, 8 Jun 2000 06:58:00 -0700 (PDT)
  From: mike.carson@enron.com
  To: james.mairet@kwikset.com
  Subject: RE: address..
  Mime-Version: 1.0
  Content-Type: text/plain; charset=us-ascii
  Content-Transfer-Encoding: 7bit
  X-From: Mike Carson
  X-To: "Mairet, James" <James.Mairet@Kwikset.com>
  X-cc:
  X-bcc:
  X-Folder: \Mike_Carson_Dec2000\Notes Folders\'sent mail
  X-Origin: Carson-M
  X-FileName: mcarson2.nsf


  Heather
```

- **opportunity, management, address, offer**

  执行结果：

```
(base) D:\课程\Web信息处理与应用\lab\lab1\src>python semantic_search.py
Please input your semantic query (words seperated with space): opportunity management address offer
('../dataset/maildir\\lavorato-j\\sent_items\\48_',) cosine similarity: 1.0770101925254814
('../dataset/maildir\\lavorato-j\\deleted_items\\190_',) cosine similarity: 1.0770101925254814
('../dataset/maildir\\lavorato-j\\sent\\574_',) cosine similarity: 1.0770101925254814
('../dataset/maildir\\lavorato-j\\all_documents\\698_',) cosine similarity: 1.0770101925254814
('../dataset/maildir\\carson-m\\discussion_threads\\83_',) cosine similarity: 0.9885496183206107
('../dataset/maildir\\ruscitti-k\\discussion_threads\\4_',) cosine similarity: 0.9885496183206107
('../dataset/maildir\\ruscitti-k\\all_documents\\262_',) cosine similarity: 0.9885496183206107
('../dataset/maildir\\carson-m\\all_documents\\125_',) cosine similarity: 0.9885496183206107
('../dataset/maildir\\dasovich-j\\sent\\77_',) cosine similarity: 0.9885496183206107
('../dataset/maildir\\carson-m\\all_documents\\61_',) cosine similarity: 0.9885496183206107
semantic search time: 12.262116193771362 s
```

- **opportunity, management, address, offer, price**

  执行结果：

```
(base) D:\课程\Web信息处理与应用\lab\lab1\src>python semantic_search.py
Please input your semantic query (words seperated with space): opportunity management address offer price
('../dataset/maildir\\lavorato-j\\sent_items\\48_',) cosine similarity: 1.0770101925254814
('../dataset/maildir\\lavorato-j\\sent\\574_',) cosine similarity: 1.0770101925254814
('../dataset/maildir\\lavorato-j\\all_documents\\698_',) cosine similarity: 1.0770101925254814
('../dataset/maildir\\lavorato-j\\deleted_items\\190_',) cosine similarity: 1.0770101925254814
('../dataset/maildir\\dean-c\\deleted_items\\104_',) cosine similarity: 1.0666666666666667
('../dataset/maildir\\zufferli-j\\sent_items\\173_',) cosine similarity: 1.0666666666666667
('../dataset/maildir\\giron-d\\_sent_mail\\341_',) cosine similarity: 1.0666666666666667
('../dataset/maildir\\giron-d\\sent\\342_',) cosine similarity: 1.0666666666666667
('../dataset/maildir\\giron-d\\sent\\341_',) cosine similarity: 1.0666666666666667
('../dataset/maildir\\giron-d\\discussion_threads\\338_',) cosine similarity: 1.0666666666666667
semantic search time: 10.777692556381226 s
```

总结

通过本实验，我们实现了一个bool检索和语义检索系统，加深了对倒排索引，tf-idf语义检索的理解。在优化系统过程中，进一步加深了对搜索系统的理解。

感谢助教和老师，提供了优质的学习资源！