

# WebInfo Lab1

PB18000239 何灏迪

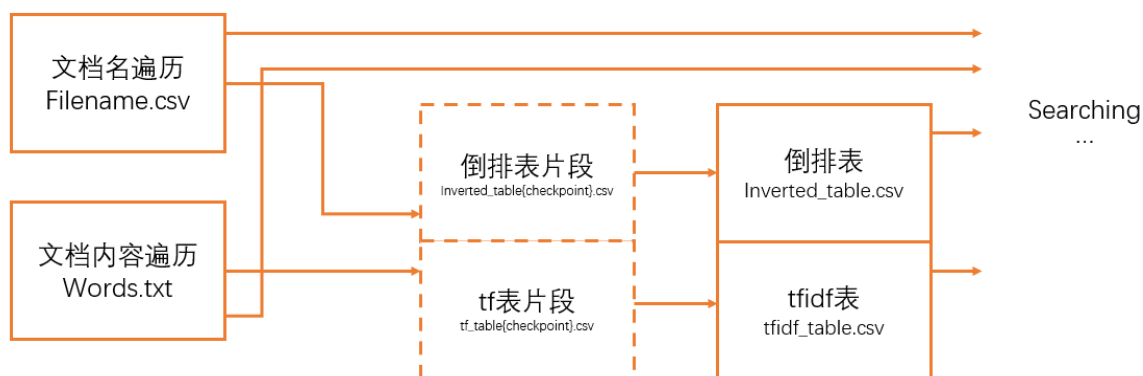
PB18000221 袁一玮

## 程序框架

### 初始化工作

文件初始化部分由 `dataLoader.py` 负责进行。

该过程具体解释如下：



程序的初始化部分分为如上图中的三个阶段，第一个阶段获取基本的文件名信息以及词频信息，第二个阶段在遍历文档的过程中生成倒排表和tf表，第三个阶段根据之前的片段合成最终的倒排表以及tfidf表。

后续的搜索过程需要基于 `Filename.csv`，`Words.txt`，`Inverted_table.csv`，`tfidf_table.csv` 四个表格进行，而中间的表格片段仅为中间结果，在已经存在上面四个文件的时候将不需要进行生成。

下面进行具体的解释。

### 文档名遍历

由于在过程中如果使用完整的路径对文档进行标识将会占用大量的空间，尤其是倒排表中需要针对每一个词存储在其中出现过的文件名。通过先对文档名进行遍历，令每一个文件对应一个id，并且在处理的过程中完全使用该id进行处理，可以有效地节省存储空间。

### 文档内容遍历

本次实验要求针对词频在前1000的词生成倒排索引表以及tfidf表，因此需要在后续工作开始之前先对词频数进行统计，从而获得这1000个词的列表。这个过程无法和后续的遍历一起进行，只能单独进行一次内容遍历，其耗时与生成索引表格的耗时相近，约需一小时。

### 倒排表片段与tf表片段

这两个表格的生成在同一次遍历过程中进行。

考虑到文档数量较大，如果一次性完成所有遍历之后存入文件中，一方面不必要的在运行过程中占用了许多空间，一方面如果出现错误情况需要全部重新进行。因此，程序设置了checkpoint机制，允许按默认10000个文件存储一个表格的方式进行生成。

该生成的过程需要使用之前生成的 `filename.csv` 以及 `words.txt`。由于此时尚没有某一个词出现在文档中次数的统计数据（`words.txt` 中统计的出现次数会重复统计一个词在文档中多次出现的情况），无法直接计算tfidf向量，只计算tf向量。

## 倒排表及tfidf表

基于倒排表片段，对各个倒排表片段的内容进行拼接后存储完整的倒排表。

利用完整的倒排表可以获得tfidf中需要的词频数据，从而可以帮助后续tfidf表格的计算。

tfidf表格同样由之前的tf表格片段和词频向量计算而成。这里理应进行归一化，但是由于在存储过程中希望节省空间，对于浮点数的输出进行了一定的长度限制，进行归一化后大部分数值都将小于1，在长度限制下有效位数更加受限，因此为保留更好的精度，不在tfidf表格生成的过程中进行归一化。这将在进行tfidf搜索的过程中耗费更多的时间（因为搜索过程中需要进行归一化操作），但是相对精度会更好。

## 搜索

### bool search

该部分由 `boolSearch.py` 进行

布尔检索想法很简单，只要把每一个元素放在倒排表中查询，再对各个 list 进行布尔操作。

刚开始尝试用 LL 文法表示 `a AND (b OR NOT c)` 这样的语句，但是操作符 `(、)` 行为比较复杂。故放弃 LL，转而使用数据结构中使用过的双栈表示法。

### tfidf search

该部分由 `tfidfSearch.py` 进行

tfidf搜索较为简单，针对输入的搜索内容生成相应的tfidf向量后，调用 `numpy` 库进行向量距离的计算，通过距离排序获得最优的搜索结果。

搜索结果见末尾结果展示部分。

## 优化思路

## 文档处理

文档处理耗费大量的时间。无论是最基础的词频统计还是后续倒排表及tf表的建立，都需要经过文档处理的过程，需要重点优化该过程。

1. 针对分词问题，最初调用了 `nltk` 库中提供的分词接口，但由于英文分词本身没有太大的复杂度，为减少时间的开销决定直接使用 `re.split()` 完成分词。
2. 针对词干提取部分，使用了 `nltk` 库中的 `PorterStemmer`，该部分的耗时占据了文档处理中的最大部分。经考虑，由于词语常常重复出现，使用cache进行词干提取的缓存可以大幅度提升效率。每次获取词干时先在cache中进行查找，如之前无缓存记录再进行词干提取。
3. 在程序中尽可能的使用 `set`，`dict` 等存储各类词语列表（如停用词、出现次数最高的1000个词等），从而减少进行 `in` 判断的耗时。

# 存储优化

存储中，倒排表和tfidf表格占用了大量的空间，在这部分我们主要进行空间上的优化。

- 1. 倒排表优化空间有限。如前所述，通过为每个文件对应一个id可以有效地减少倒排表需要存储的字符串长度。
- 2. tf表及tfidf表部分优化空间较大。首先，该部分所有数据都为浮点数格式，如果直接进行存储，每个数最长可至十余字符，提供了很多不必要的精度。因此，在实际操作中仅保留三位小数。结合下面的第三点，这将减少至少50%的空间占用。
- 3. 针对tf表及tfidf表，为节省存储空间，并没有使用常规的存储格式。由于tf向量中含有大量的0值，表格中仅存储非0值向量所在维及其值，如[0.1, 0, 0, 0.3, 0, 0 0, 0] 将仅存储为 [0, 0.1, 3, 0.3]。通过这样的手段，空间占用被大幅度压缩，但对数据的“解码”过程带来了一定的时间损失。

# 断点加载

考虑到文档数量较大，如果一次性完成所有遍历之后存入文件中，一方面不必要的在运行过程中占用了许多空间，一方面如果出现错误情况需要全部重新进行。因此，程序设置了checkpoint机制，允许按默认10000个文件存储一个表格的方式进行生成。

# 结果展示

## bool search 结果展示

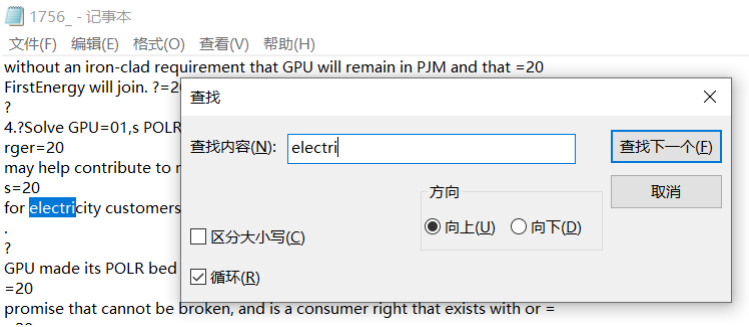
查询 `energy AND company AND opportunity AND NOT enron AND electricity`

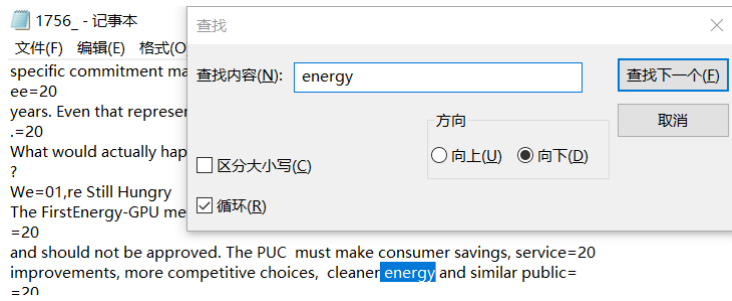
结果返回如下

```
(WebInfo) C:\Users\86185\Desktop\Files\aca\web\WebInfo-Lab\lab1>python src\boolSearch.py
filename and words list found.
Initialization Complete.
Loading inverted table..
Please input one space between all words(include parenthesis/brace and items).
(quit by input 'EXIT')Search for:energy AND company AND opportunity AND NOT enron AND electricity
Found in 861 files. First found in dasovich-j\notes_inbox\1756_

Please input one space between all words(include parenthesis/brace and items).
(quit by input 'EXIT')Search for:
```

排名靠前的 `dasovich-j\notes_inbox\1756_` 内容较长，截图如下：





## tfidf 搜索结果展示

```
filename and words list found.
Initialization Complete.
(quit by input 'EXIT')Search for:power business company meeting

Most related files:
    kean-s\all_documents\9_
    kean-s\archiving\untitled\2317_
    kean-s\calendar\untitled\9_
    kean-s\discussion_threads\9_
    griffith-j\all_documents\402_
    griffith-j\discussion_threads\376_
    griffith-j\newpower\2_
    allen-p\all_documents\148_
    allen-p\all_documents\230_
    allen-p\all_documents\238_

(quit by input 'EXIT')Search for:EXIT
```

将排名靠前的几篇展示如下：

- kean-s\all\_documents\9\_ 内容如下，出现了 meeting 与 company 两个关键词。

```
Message-ID: <3571459.1075846140064.JavaMail.evans@thyme>
Date: Wed, 19 Mar 1997 23:30:00 -0800 (PST)
From: steven.kean@enron.com
Subject: Washington DC
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: Steven J Kean
X-To:
X-cc:
X-bcc:
X-Folder: \Steven_Kean_Dec2000_1\Notes Folders\All documents
X-Origin: KEAN-S
X-FileName: skean.nsf
```

pena - 1:30; Spurling @11:15 and D'Amato @12:00

11:15 D'Amato  
12:00 Spurling  
1:30 Pena meeting - Tentative  
3:30 JaneI Guerrero

Flying up on company plane.

- kean-s\archiving\untitled\2317\_ 内容如下：(之后的几个均与上一封完全相同)

Message-ID: <30010916.1075846270126.JavaMail.evans@thyme>  
Date: wed, 19 Mar 1997 23:30:00 -0800 (PST)  
From: steven.kean@enron.com  
Subject: Washington DC  
Mime-Version: 1.0  
Content-Type: text/plain; charset=us-ascii  
Content-Transfer-Encoding: 7bit  
X-From: Steven J Kean  
X-To:  
X-cc:  
X-bcc:  
X-Folder: \Steven\_Kean\_Dec2000\_1\Notes Folders\Archiving\Untitled  
X-Origin: KEAN-S  
X-FileName: skean.nsf

pena - 1:30; Spurling @11:15 and D'Amato @12:00

11:15 D'Amato  
12:00 Sperling  
1:30 Pena meeting - Tentative  
3:30 Janel Guerrero

Flying up on company plane.

- griffith-j\all\_documents\402\_ 内容如下, 出现了 power, company, business 三个关键词。

Message-ID: <12055679.1075849642682.JavaMail.evans@thyme>  
Date: Tue, 10 Apr 2001 06:44:00 -0700 (PDT)  
From: registrations@newpower.com  
To: john.griffith@enron.com  
Subject: Re: <<01100134337001A>> NewPower Application  
Mime-Version: 1.0  
Content-Type: text/plain; charset=us-ascii  
Content-Transfer-Encoding: 7bit  
X-From: "Registrations" <registrations@newpower.com>  
X-To: john.griffith@enron.com  
X-cc:  
X-bcc:  
X-Folder: \John\_Griffith\_Nov2001\Notes Folders\All documents  
X-Origin: GRIFFITH-J  
X-FileName: jgriffit.nsf

The New Power Company  
The smarter way to power your home

registrations@newpower.com - Business Hours 7:00 a.m. to 9:00 p.m. EST

```
filename and words list found.
Initialization Complete.
(quit by input 'EXIT')Search for:electricity document plans

Most related files:
    horton-s\all_documents\95_

    horton-s\discussion_threads\90_

    bass-e\inbox\55_

    mann-k\all_documents\2542_

    mann-k\sent\1905_

    mann-k\_sent_mail\2457_

    zipper-a\sent_items\101_

    nemec-g\all_documents\157_

    nemec-g\sent\154_

    allen-p\all_documents\148_

(quit by input 'EXIT')Search for:
```

horton-s\all\_documents\95\_ 内容如下, 出现了关键词 document。

```
Message-ID: <11695282.1075844934299.JavaMail.evans@thyme>
Date: Wed, 23 Feb 2000 08:34:00 -0800 (PST)
From: ruth.mann@enron.com
To: caroline.barnes@enron.com, frank.bay@enron.com, lynn.blair@enron.com,
    sharon.brown@enron.com, john.buchanan@enron.com,
    janet.butler@enron.com, deb.cappiello@enron.com,
    alma.carrillo@enron.com, scott.coburn@enron.com,
    janet.cones@enron.com, bill.cordes@enron.com,
    shelley.corman@enron.com, lisa.costello@enron.com,
    larry.derooin@enron.com, rick.dietz@enron.com, dari.dornan@enron.com,
    diane.eckels@enron.com, george.fastuca@enron.com,
    anne.jolibois@enron.com, bob.hall@enron.com, mary.hamilton@enron.com,
    glen.hass@enron.com, robert.hayes@enron.com, rod.hayslett@enron.com,
    theresa.hess@enron.com, tamara.hopkins@enron.com,
    stanley.horton@enron.com, steve.hotte@enron.com,
    martha.janousek@enron.com, steven.january@enron.com,
    tammy.jaquet@enron.com, robert.kilmer@enron.com,
    frazier.king@enron.com, dan.kirtane@enron.com, terry.lehn@enron.com,
    teb.lokey@enron.com, dorothy.mccoppin@enron.com,
    mike.mcgowan@enron.com, gerry.medeles@enron.com,
    mary.miller@enron.com, jan.moore@enron.com, sheila.nacey@enron.com,
    ray.neppl@enron.com, virginia.o'neill@enron.com,
    zelda.paschal@enron.com, maria.pavlou@enron.com,
    peggy.phillips@enron.com, janet.place@enron.com, jenny.rub@enron.com,
    patti.rumler@enron.com, sharon.solon@enron.com,
    staci.holtzman@enron.com, cindy.stark@enron.com,
    james.studebaker@enron.com, dee.svatos@enron.com,
    john.tsucalas@enron.com, julia.white@enron.com,
    ricki.winters@enron.com
```

Subject: GISB Highlights  
Mime-Version: 1.0  
Content-Type: text/plain; charset=us-ascii  
Content-Transfer-Encoding: 7bit  
X-From: Ruth Mann  
X-To: Caroline Barnes, Frank Bay, Lynn Blair, Sharon Brown, John Buchanan, Janet Butler, Deb Cappiello, Alma Carrillo, Scott Coburn, Janet Cones, Bill Cordes, Shelley Corman, Lisa Costello, Larry DeRoin, Rick Dietz, Dari Dornan, Diane Eckels, George Fastuca, Anne Jolibois, Bob M Hall, Mary Lou Hamilton, Glen Hass, Robert Hayes, Rod Hayslett, Theresa Hess, Tamara Hopkins, Stanley Horton, Steve Hotte, Martha Janousek, Steven January, Tammy Jaquet, Robert Kilmer, Frazier King, Dan Kirtane, Terry Lehn, Teb Lokey, Dorothy McCoppin, Mike McGowan, Gerry Medeles, Mary Kay Miller, Jan Moore, Sheila Nacey, Ray Nepp1, Virginia O'Neill, Zelda Paschal, Maria Pavlou, Peggy Phillips, Janet Place, Jenny Rub, Patti Rumler, Sharon Solon, Staci Holtzman, Cindy Stark, James Studebaker, Dee Svatos, John Tsucalas, Julia White, Ricki Winters  
X-cc:  
X-bcc:  
X-Folder: \Stanley\_Horton\_1\Notes Folders\All documents  
X-Origin: HORTON-S  
X-FileName: shorton.nsf

This document is from Theresa Hess and Tammy Hopkins.

bass-e\inbox\55\_ 内容如下, 出现了关键词 plan:

Message-ID: <14068132.1075840320520.JavaMail.evans@thyme>  
Date: wed, 30 Jan 2002 08:53:42 -0800 (PST)  
From: timothy.blanchard@enron.com  
To: eric.bass@enron.com  
Subject: Super Bowl  
Mime-Version: 1.0  
Content-Type: text/plain; charset=us-ascii  
Content-Transfer-Encoding: 7bit  
X-From: Blanchard, Timothy </O=ENRON/OU=NA/CN=RECIPIENTS/CN=TBLANCHA>  
X-To: Bass, Eric </O=ENRON/OU=NA/CN=RECIPIENTS/CN=Ebass>  
X-cc:  
X-bcc:  
X-Folder: \ExMerge - Bass, Eric\Inbox  
X-Origin: BASS-E  
X-FileName: eric bass 6-25-02.PST

what's the plan?



```
filename and words list found.
Initialization Complete.
(quit by input 'EXIT')Search for:FOR THOSE OF YOU WHO HAVE RESERVED CABARET TICKETS - SHOW ME THE MONEY!!!!!!Make Checks
payable to Enron Corp. and forward to me.DEADLINE IS Friday 5/26.I still have a few tickets left, so if you want to have
some fun ... come to

Most related files:
  cash-m\all_documents\268_
  cash-m\connect_deletes\2_
  lenhart-m\all_documents\1941_
  lenhart-m\discussion_threads\379_
  lenhart-m\sent\1940_
  allen-p\all_documents\148_
  allen-p\all_documents\230_
  allen-p\all_documents\238_
  allen-p\all_documents\269_
  allen-p\all_documents\360_
```

使用文档的原文片段进行搜索，可以得到正确结果。 `cash-m\all_documents\268_` 的内容如下：

```
Message-ID: <11424325.1075860483092.JavaMail.evans@thyme>
Date: Wed, 24 May 2000 08:08:00 -0700 (PDT)
From: cheryl.arguijo@enron.com
To: bcc@enron.com
Subject: Please send out below message - Thanks!
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: Cheryl Arguijo
X-To: BCC
X-CC:
X-bcc:
X-Folder: \Michelle_Cash_Dec2000\Notes Folders\All documents
X-Origin: Cash-M
X-FileName: mcash.nsf
```

PLEASE RESPOND TO IRIS WASER!!

FOR THOSE OF YOU WHO HAVE RESERVED CABARET TICKETS - SHOW ME THE MONEY!!!!!!

Make Checks payable to Enron Corp. and forward to me - Iris Waser at EB846a -  
DEADLINE IS Friday 5/26.

I still have a few tickets left, so if you want to have some fun ... come to  
the CABARET on July 8 (Saturday matinee).


Member ticket prices are \$37.00. Give me a call at 39012 or send your check  
to EB846a.

## 优化效果展示

在这里仅展示最关键的两个优化：

### tf向量存储的空间优化

不使用优化:

 tf_table10000.csv	2020/12/3 17:20	Microsoft Excel ...	20,667 KB
---	-----------------	---------------------	-----------

使用之前所述优化方法后:

 tf_table10000.csv	2020/12/3 17:14	Microsoft Excel ...	3,524 KB
---	-----------------	---------------------	----------

可见使用的存储空间仅为原先的1/6.

如果不进行优化, 最终生成的tfidf表格将达到近1GB.

## 遍历过程中使用cache实现词干提取的时间优化

未使用cache进行优化时:

```
Start without checkpoint.
1000 files have been visited, time cost: 28.5625
Cache length: 0
2000 files have been visited, time cost: 18.390625
Cache length: 0
3000 files have been visited, time cost: 18.625
Cache length: 0
4000 files have been visited, time cost: 21.8125
Cache length: 0
5000 files have been visited, time cost: 47.59375
Cache length: 0
6000 files have been visited, time cost: 23.21875
Cache length: 0
7000 files have been visited, time cost: 17.34375
Cache length: 0
8000 files have been visited, time cost: 14.765625
Cache length: 0
9000 files have been visited, time cost: 28.484375
Cache length: 0
10000 files have been visited, time cost: 19.453125
Cache length: 0
```

使用cache优化后:

```
filename and words list found.
Start without checkpoint.
1000 files have been visited, time cost: 10.5
Cache length: 12896
2000 files have been visited, time cost: 5.296875
Cache length: 13383
3000 files have been visited, time cost: 4.96875
Cache length: 14895
4000 files have been visited, time cost: 6.109375
Cache length: 20874
5000 files have been visited, time cost: 7.4375
Cache length: 32031
6000 files have been visited, time cost: 5.25
Cache length: 33454
7000 files have been visited, time cost: 5.234375
Cache length: 34778
8000 files have been visited, time cost: 4.875
Cache length: 35177
9000 files have been visited, time cost: 7.109375
Cache length: 38590
10000 files have been visited, time cost: 5.1875
Cache length: 39698
output/inverted_table10000.csv has been saved.
output/tf_table10000.csv has been saved.
```

这仅是开始阶段的10000个文件的情况，考虑当过程逐渐进行，更大比例的词将能在cache中直接获取，其效率将更进一步的提升。仅最开始的10000个文件，性能已经达到原先的4倍。

## 总结

---

1. 观察tfidf搜索的结果，不难发现结果往往具有文档短小的特点，最夸张者会出现文档中并不出现任何词频在前1000词内的词语，其tfidf向量即为[0, ..., 0]。这样的文档仍会被认为是相关性靠前的文档。

对搜索词本身较短的情况，对其提取tfidf向量的实际意义有待进一步考虑。如果要提升tfidf搜索的效果，可能更好的方式是直接将文档在搜索词对应维的tfidf值相加，之后进行排序。