Iran University Of science And Technology

(IUST)

حذف هرزنامه

استاد:

دکتر مینایی

دانشجو:

پگاه ترکمندی

شماره دانشجویی:

89551264

2014

```
            tmp = tmp + line[:end]
            b_start = 0
            b_end =1

    if(b_end == 1):
        tmp = re.findall(r"[\w']+",tmp)
        nd=0
        for i in range(0,len(tmp)):
            tmp[i]=tmp[i].lower()
            tmp[i] = porter.stem(tmp[i])
            if(tmp[i] in list_stop):
                continue
            m.append((tmp[i],[doc_id,i]))
            nd+=1
        num_doc.append([pp,nd])
        b_end=0
        tmp=str()
        doc_id = doc_id + 1
    if sys.getsizeof(m)>4*1024*1024:
        total_temp_files=total_temp_files+1
        fname='partial-temp-'+str(total_temp_files)+'.dat'
        print 'saving temporary file ',fname,'...'
        m=sort_data(m)
        write_file(fname,m,int(0),len(m))
        total_records=total_records+len(m)
        m=[]
    position=fr.tell()
    line=fr.readline()
    _num+=1
total_temp_files=total_temp_files+1
fname='partial-temp-'+str(total_temp_files)+'.dat'
print 'saving temporary file ',fname,'...'
m=sort_data(m)
write_file(fname,m,int(0),len(m))
```

اردیبهشت -1393

Create_black_list()

در ابتدا لیستی از کلماتی که در اسناد هرزنامه بیشتر مشاهده می شود را در فایلی ذخیره کرده و هنگام بررسی هر سند یا میل استفاده می شود.

:Inverted Index <=

(با استفاده از() build index merged_based و build index sorted_based() چون در حینی که term را به همراه posting list در هارد ذخیره کنم قبل merge() زدن مثل build index sorted_based() به شکل sort شده ذخیره میکنم بعد mergeمیزنم.)

def main_dictionary()

در این قسمت از اول corpus شروع به خواندن میکند هنگامی که تگ باز هر میل پیدا شد ادرس بایت خط قبل از خطی که شامل تگ شروع هر میل است را در یک لیست را به شکل زیر ذخیره میکنیم←

```
Position= f.tell()     //address har byte

 f.readline()
```

و از هر تگ باز تا تگ بسته تمام کلمات داخل میل را با استفاده از دو تابع lower(string) و porte.stem(string) به ریشه کلمه و حروف کوچیک تبدیل میکند و همراه با شماره میلی که شامل این رشته و جایگاهش در یک لیست ذخیره میکنیم ([(string,[(doc_id,position)])) در حین انجام این کار هنگامی که تگ باز هر میل پیدا شد ادرس بایت خط قبل از خطی که شامل تگ شروع هر میل است را در یک لیست ذخیره میکنیم وتا اخر corpus تا جایی که اندازه لیست کمتر از mg4 است به لیست اضافه میکنیم و زمانی که اندازه لیست در ROM بیشتر از اندازه در نظر گرفته شده باشد لیست را sort کرده و با این کار رشته های یکسان کنار هم قرار گرفته و شماره سندهایی که شامل این رشته هستند را در یک لیست قرار میدهیم که همان posting list هر term است

```
def sort_data(m):
    m.sort()
    for i in range(len(m)):
        j=1
        while( i+j < len(m) and m[i][0] == m[i+j][0]):
            m[i][1].extend(m[i+j][1])
            j=j+1
        m.__delslice__(i+1,i+j)
    return m
```

و در نهایت لیست به شکل رو به رو است ←

list=[(term,postinglist)] postinglist=[(doc_id,position)]

و در هارد به صورت باینری binary به شکل زیرذخیره میکنیم←

Len(list),len(term),term,len(posting_list),posting_list

```
def write_file(oname,m,s,l):
    with open(oname, 'wb') as f:
        f.write(struct.pack('i',l))
        for i in range(s,l):
            f.write(struct.pack('i',len(m[i][0])))
            if len(m[i][0])>0:
                f.write(''.join(chr(j) for j in map(ord,m[i][0])).encode('ascii'))
                f.write(struct.pack('i',len(m[i][1])))
                for j in range(len(m[i][1])):
                    f.write(struct.pack('i',m[i][1][j][0]))
                    f.write(struct.pack('i',m[i][1][j][1]))
```

در نهایت که تمام corpus را خواندیم و نتیجه را در چندین فایل ذخیره کردیم با مرج کردن هر دو فایل و تبدیل به یک فایل تمام InvertedIndex را که به طور جدا در هر فایل sort شده بود را sort میکنیم

def merge_files (file1, file2, output)

که دو اشاره گر به اول هر دو فایل داریم و term و هر دو را داخل ROM اورده و مقایسه میکنیم هر کدام که رشته کوچکتری داشت تا زمانی که کوچکتر باشد داخل فایل نهایی مینویسیم و زمانی که رشته هر دو برابر بود posting list هر دو را یکی کرده و داخل فایل نهایی مینویسیم و این کار را تا زمانی انجام میدیم که هر دو فایل را کامل خوانده باشیم و تمامی محتوا هر دو را داخل فایل نهایی بریزیم و اینقد این تابع را به ازای هر دو فایل صدا میکنیم و تبدیل به یک فایل میکنیم که در نهایت یک فایل داریم که شامل inverted index نهایی از corpus است.

حال برای اینکه هنگام جست و جوی blacklist_word کل این فایل را جست و جو نکنیم و یا اینکه سایز فایل به حدی بزرگ باشد که داخل ROM جا نشود از per term استفاده میکنیم که این فایل نهایی را به چندین فایل به اندازه حدودی 2-3 mg ذخیره میکنیم def unmerged_files() و term اول هر فایل را به همرا شماره فایل در یک فایل ذخیره میکنیم که موقع جست و جو برای blacklist_word اول این فایل را جست و جو کرده و تعیین میکنیم که در کدام فایل قرار دارد.

def write_start_unmergefile()

حال ما دارای چند فایل هستیم که در فاز 2 از آن ها استفاده میکنیم:

1) unmerge-inverted-index-i:

چندین فایل با این اسم داریم که شامل invertedindex هستند

2) size-unmerge-inverted-indexsize-unmerge-inverted-index:

شامل تعداد فایل های invertedindex است

3) unmerge-inverted-index:

شامل شروع هر فایل invertedindex است (PerTerm)

4) number_of_document:

شامل ادرس بایت خط قبل از شروع هر سند است

با استفاده از تابع rankBM25_DocumentAtAtime_WithHeap(q,k)

میل هایی که مرتبط هستند را به ترتیب ویژگی هرزنامه بودنشان را بدست می اوریم ولی این تابع فقط میل هایی را که شامل تعداد بیشتری از term های blacklist_word باشد را در نظر می گیرد و اینکه term های blacklist_word پشت سر هم باشد را بررسی

نمیکند در نتیجه با استفاده از تابع nextphrase این شرط را نیز بررسی کرده و ارزش بیشتری به میل هایی که شامل این شرط باشد را میدهیم

و در نهایت میل هایی را که هرزنامه تشخیص داده شده اند را به کاربر نشان میدهیم

**** با استفاده از تابع p= f.tell() ادرس بایت خط قبل شروع هر سند را در هارد ذخیره کرده بودیم که خیلی این کار در مواقعی که میخواهیم فقط یکسری از اسناد را بررسی کنیم مطلوب است چون با دستور f.seek(p) می توانیم بدون اینکه از اول corpusشروع به خواندن کنیم مستقیما به خط شروع هر میل برویم.

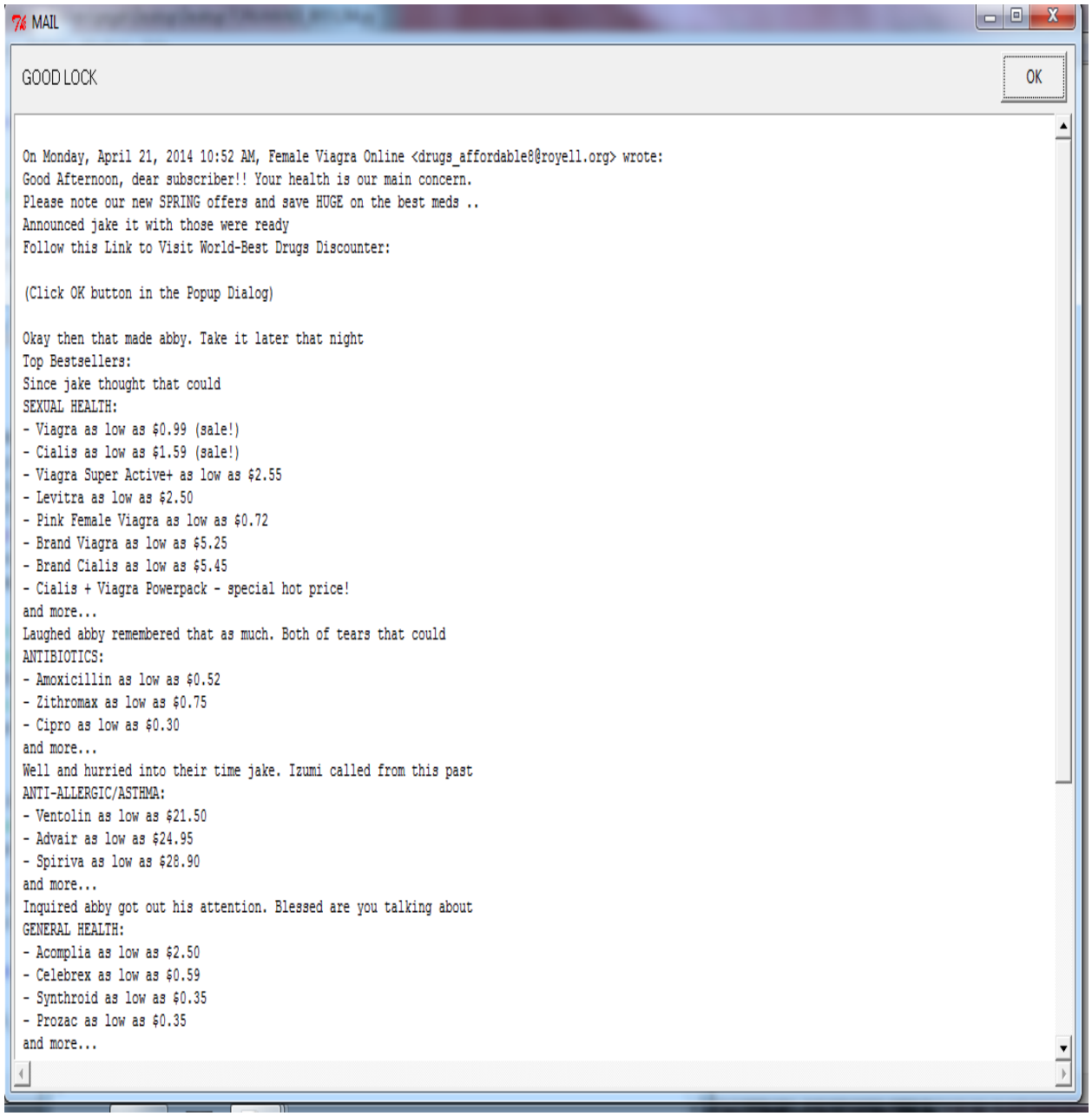
نحوه نمایش اطلاعات به کاربر:

```
Please Enter Your Query:
bahia coco
pleas enter how many document id you want?
25
DOC1:
------------------------------------------------------------------------------
One grain vessel was awaiting berth

at Bahia Blanca, four at Buenos Aires and five at Rosario, on

March 31, National Grain Board figures show.

------------------------------------------------------------------------------
DOC2:
------------------------------------------------------------------------------
Brazil is not selling cocoa beans

to the International Cocoa Organization, ICCO, buffer stock, as

spot prices for beans in the interior area are 20 to 25 pct

------------------------------------------------------------------------------
DOC3:
------------------------------------------------------------------------------
ARGENTINE GRAIN BELT TEMPERATURES

(CENTIGRADE) AND RAIN (MM) IN THE 24 HOURS TO 12.00 GMT WERE:

    ..............MAX TEMP..MIN TEMP..RAINFALL

------------------------------------------------------------------------------
DOC4:
------------------------------------------------------------------------------
Closing prices on the Bahia Cocoa

Futures Market were as follows (in cruzados per 60 kilo bag

delivered Ilheus/Itabuna with previous in brackets) -

------------------------------------------------------------------------------
DOC5:
------------------------------------------------------------------------------
Cocoa futures drifted into a dull

finish, settling at nine dlrs lower to seven higher, with most

contracts near unchanged from midday levels.

------------------------------------------------------------------------------
DOC6:
------------------------------------------------------------------------------
ARGENTINE GRAIN BELT TEMPERATURES

(CENTIGRADE) AND RAIN (MM) IN THE 24 HOURS TO 12.00 GMT WERE:

    ..............MAX TEMP..MIN TEMP..RAINFALL
```

که هر میل را به همراه خلاصه ایی از اطلاعاتش به کاربر نشان میدهد تا کاربر بتواند میل مطلوب خود را راحت تر انتخاب کند

گرافیک پروژه:

کاربر میل مد نظر خود را انتخاب کرده و کل متن میل در یک window به کاربر نمایش میدهد.



کاربر می تواند بعد از خواندن میل و click ok میل دیگری را انتخاب کند و مشاهده کند☺

****در قسمت گرافیکی از library easygui استفاده کردم ****