Report of Lab3

Yuxiang Chen 5110309783

October 21, 2012

Contents

1 The Purpose of Lab3 and My Preparation

It not hard to find that the purpose of lab3 is to let us know and command the theorem of indexing and searching the information on the Internet based on the use of lucene. And what we need to know first is the principle of Full-text Search. In this method, we need to firstly index the information we have got on the Internet, then search the index according to some key words. Although the principle sounds easy, it's a little hard for me to do it very fast, since I don't know what lucene is and how it works at first. Then I have to study on my own ...... And at last I find how I can accomplish my experiment.
The first part of the Full-text Search is indexing, and it is composed of the following 4 parts:
1. Find the document you want to index;
2. Pass the document into the Tokenizer;
3. Give the token you got at step 2 to the linguistic processor;
4. Pass the term to the indexer.
And the latter half – search can also be split into 4 parts:

1. The user input the query;
2. Parse and process the query;
3. Search the index and find the document agreed to the syntax tree;
4. Sort the result according to the relativity of the document, etc.

The second half of the experiment is mainly an extending on the basis of the first part, which acquires us to contain more information indexed and searched in our programs. The first needs an extra information 'site' and the second one needs the url of the picture to be searched, and the name and url of the page it belongs to.

And thus using the tools we have, it's time to achieve our goal.

2 The Main Part of the Experiment

In this experiment, I split the program into three parts: crawling to get web pages, setting up index, and searching according to the query.

2.1 The First Half of the Experiment

The first part is separated into three parts by me, which in my view makes the experiment more organized and easier to understand.

2.1.1 Crawl to Get Web Pages

In this part, I mainly used the program we wrote in lab2 to crawl web pages and store them in "F:\html". But I found that the speed to crawl web pages is a little slow, so I choose 'http://www.sjtu.edu.cn' to make it fast for us to get enough pages. If you want to crawl other pages, you may just change it in the following code.

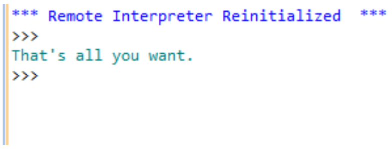
And here is the code:

```
from BeautifulSoup import BeautifulSoup
import urllib2
import re
import urlparse
import os
import urllib
import socket
import threading
import Queue
import time
import chardet
import sys
reload(sys)
```

```
sys.setdefaultencoding('utf8')


def valid_filename(s):
    import string
    valid_chars = "-_.() %s%s" % (string.ascii_letters, string.digits)
    s = ''.join(c for c in s if c in valid_chars)
    return s

def get_page(page):
    time.sleep(0.001)
    try:
        content=urllib2.urlopen(page,timeout=3).read()
        result = chardet.detect(content)['encoding']
        if result=='GB2312':
            content=content.decode('gbk').encode('utf8')
        return content
    except:
        #There is an error.#
        return []

def get_all_links(content, page):
    if content==[]:
        return []
    links = []
    tempset=set()
    soup=BeautifulSoup(content)
    for i in soup.findAll('a',{'href':re.compile(('^http|^/'))}):
        tempset.add(i['href'])
    for i in tempset:
        links.append(urlparse.urljoin(page,i))
    return links

def add_page_to_folder(page, content):
    folder = 'F:\\html'
    index_filename = 'F:\\html\index.txt'
    filename = valid_filename(page)
    index = open(index_filename, 'a')
    index.write(filename + ';' + page + '\n')
    index.close()
    if not os.path.exists(folder):
        os.mkdir(folder)
    f = open(os.path.join(folder, filename), 'w')
    f.write(content)
    f.close()

```

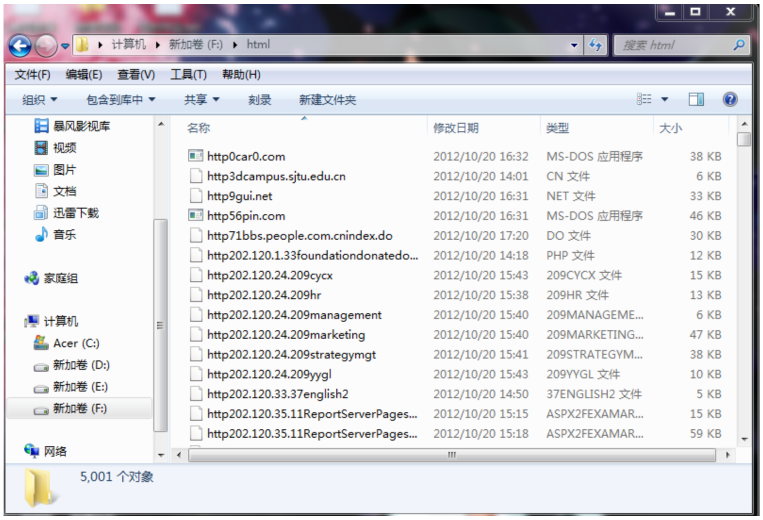
```
def working():
    page_num=0
    while page_num<task_per_thread:
        page = q.get()
        if page not in crawled:
            content = get_page(page)
            outlinks = get_all_links(content,page)
            if outlinks==[]:
                q.task_done()
                continue
            page_num+=1
            add_page_to_folder(page,content)
            for link in outlinks:
                q.put(link)
            if varLock.acquire():
                crawled.append(page)
                varLock.release()
                q.task_done()
            else:
                q.task_done()
        else:
            q.task_done()
    while q.empty()==False:
        q.get()
        q.task_done()

if not os.path.exists("F:\\html"):
    os.mkdir("F:\\html")
NUM = 100
task_per_thread=50
crawled = []
varLock = threading.Lock()
q = Queue.Queue()
q.put('http://www.sjtu.edu.cn')
for i in range(NUM):
    t=threading.Thread(target=working)
    t.setDaemon(True)
    t.start()
q.join()
print "That's␣all␣you␣want."
```

In this part, we need to analyze the coding method of the page, and I transform them all to 'utf8', which will make the following parts easier. And then I save the name of the page and its url to ”F:\html\index.txt”. And here are the pictures of the compiler and the file. And in the picture, plus one ”index.txt”, there are 5001 documents in total.



```
*** Remote Interpreter Reinitialized  ***
>>> 
That's all you want.
>>> 
```

Figure 1: the outcome of the compiler



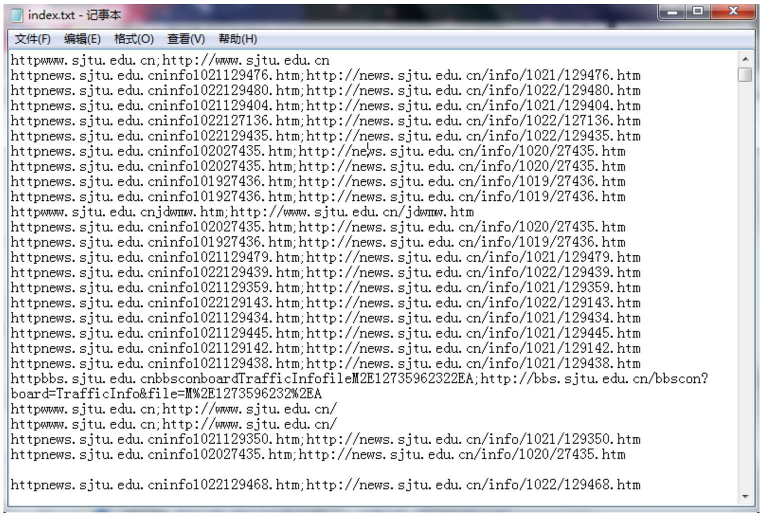
名称	修改日期	类型	大小
http0car0.com	2012/10/20 16:32	MS-DOS 应用程序	38 KB
http3dcampus.sjtu.edu.cn	2012/10/20 14:01	CN 文件	6 KB
http9gui.net	2012/10/20 16:31	NET 文件	33 KB
http56pin.com	2012/10/20 16:31	MS-DOS 应用程序	46 KB
http71bbs.people.com.cnindex.do	2012/10/20 17:20	DO 文件	30 KB
http202.120.1.33foundationdonatedo...	2012/10/20 14:18	PHP 文件	12 KB
http202.120.24.209cycx	2012/10/20 15:43	209CYCX 文件	15 KB
http202.120.24.209hr	2012/10/20 15:38	209HR 文件	13 KB
http202.120.24.209management	2012/10/20 15:40	209MANAGEME...	6 KB
http202.120.24.209marketing	2012/10/20 15:40	209MARKETING...	47 KB
http202.120.24.209strategymgt	2012/10/20 15:41	209STRATEGYM...	38 KB
http202.120.24.209yygl	2012/10/20 15:43	209YYGL 文件	10 KB
http202.120.33.37english2	2012/10/20 14:50	37ENGLISH2 文件	5 KB
http202.120.35.11ReportServerPages...	2012/10/20 15:15	ASPX2FEXAMAR...	15 KB
http202.120.35.11ReportServerPages...	2012/10/20 15:18	ASPX2FEXAMAR...	59 KB

Figure 2: the screenshot of F:\html

2.1.2 Set up the Index

It is really a little annoying in this part, since there're so many things to do. In this part, we have to set up an index to save the name,path,title,url and contents of the page we have crawled. Meanwhile, I also give the Chinese words a segmentation using ICTCLAS50. And though it's quite complex, I feel it is hard to describe how to achieve them, so I will only display the code and you may find how it works. By the way, in order to find url easily, I make a dictionary using 'index.txt' we saved in the first part.

```
import sys, os, lucene, threading, time,chardet,urllib2
from datetime import datetime
from BeautifulSoup import BeautifulSoup
from ctypes import *

```



```
httpwww.sjtu.edu.cn;http://www.sjtu.edu.cn
httpnews.sjtu.edu.cninfo1021129476.htm;http://news.sjtu.edu.cn/info/1021/129476.htm
httpnews.sjtu.edu.cninfo1022129480.htm;http://news.sjtu.edu.cn/info/1022/129480.htm
httpnews.sjtu.edu.cninfo1021129404.htm;http://news.sjtu.edu.cn/info/1021/129404.htm
httpnews.sjtu.edu.cninfo1022127136.htm;http://news.sjtu.edu.cn/info/1022/127136.htm
httpnews.sjtu.edu.cninfo1022129435.htm;http://news.sjtu.edu.cn/info/1022/129435.htm
httpnews.sjtu.edu.cninfo102027435.htm;http://news.sjtu.edu.cn/info/1020/27435.htm
httpnews.sjtu.edu.cninfo102027435.htm;http://news.sjtu.edu.cn/info/1020/27435.htm
httpnews.sjtu.edu.cninfo101927436.htm;http://news.sjtu.edu.cn/info/1019/27436.htm
httpnews.sjtu.edu.cninfo101927436.htm;http://news.sjtu.edu.cn/info/1019/27436.htm
httpwww.sjtu.edu.cnjdwmw.htm;http://www.sjtu.edu.cn/jdwmw.htm
httpnews.sjtu.edu.cninfo102027435.htm;http://news.sjtu.edu.cn/info/1020/27435.htm
httpnews.sjtu.edu.cninfo101927436.htm;http://news.sjtu.edu.cn/info/1019/27436.htm
httpnews.sjtu.edu.cninfo1021129479.htm;http://news.sjtu.edu.cn/info/1021/129479.htm
httpnews.sjtu.edu.cninfo1022129439.htm;http://news.sjtu.edu.cn/info/1022/129439.htm
httpnews.sjtu.edu.cninfo1021129359.htm;http://news.sjtu.edu.cn/info/1021/129359.htm
httpnews.sjtu.edu.cninfo1022129143.htm;http://news.sjtu.edu.cn/info/1022/129143.htm
httpnews.sjtu.edu.cninfo1021129434.htm;http://news.sjtu.edu.cn/info/1021/129434.htm
httpnews.sjtu.edu.cninfo1021129445.htm;http://news.sjtu.edu.cn/info/1021/129445.htm
httpnews.sjtu.edu.cninfo1021129142.htm;http://news.sjtu.edu.cn/info/1021/129142.htm
httpnews.sjtu.edu.cninfo1021129438.htm;http://news.sjtu.edu.cn/info/1021/129438.htm
httpbbs.sjtu.edu.cnbbsconboardTrafficInfofileM2E12735962322EA;http://bbs.sjtu.edu.cn/bbscon?
board=TrafficInfo&file=M%2E1273596232%2EA
httpwww.sjtu.edu.cn;http://www.sjtu.edu.cn/
httpwww.sjtu.edu.cn;http://www.sjtu.edu.cn/
httpnews.sjtu.edu.cninfo1021129350.htm;http://news.sjtu.edu.cn/info/1021/129350.htm
httpnews.sjtu.edu.cninfo102027435.htm;http://news.sjtu.edu.cn/info/1020/27435.htm

httpnews.sjtu.edu.cninfo1022129468.htm;http://news.sjtu.edu.cn/info/1022/129468.htm
```

Figure 3: the screenshot of index.txt

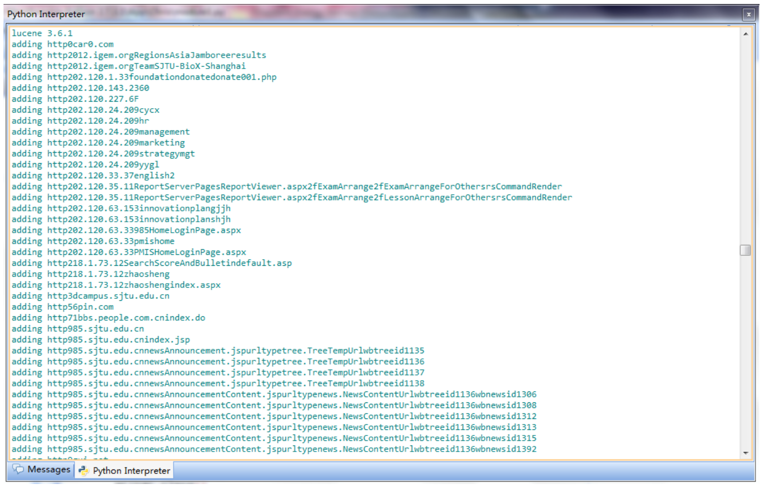
```python
"""
This class is loosely based on the Lucene (java implementation) demo class
org.apache.lucene.demo.IndexFiles.  It will take a directory as an argument
and will index all of the files in that directory and downward recursively.
It will index on the file path, the file name and the file contents.
The
resulting Lucene index will be placed in the current directory and called
'index'.
"""

class IndexFiles(object):
    """Usage: python IndexFiles <doc_directory>"""

    def __init__(self, root, storeDir, analyzer):

        if not os.path.exists(storeDir):
            os.mkdir(storeDir)
        store = lucene.SimpleFSDirectory(lucene.File(storeDir))
        writer = lucene.IndexWriter(store, analyzer, True,
                                    lucene.IndexWriter.MaxFieldLength.LIMITED)
        writer.setMaxFieldLength(1048576)
        self.indexDocs(root, writer)
        print 'optimizing index.',
        writer.optimize()
        writer.close()
        print 'done'

    def indexDocs(self, root, writer):
        for root, dirnames, filenames in os.walk(root):
            for filename in filenames:
                if filename.endswith(’.txt’):
                    continue
                print ”adding”, filename
                try:
                    path = os.path.join(root, filename)
                    file = open(path)
                    buf = file.read()
                    contents=buf
                    result = chardet.detect(buf)[’encoding’]
                    if result==’GB2312’:
                        contents = buf.decode(’gbk’).encode(’utf8’)
                    file.close()
                    soup=BeautifulSoup(contents)
                    url=mydict[filename]
                    title=str(soup.head.title.string).decode(’utf8’)
                    contents=’’.join(soup.findAll(text=True))
                    doc = lucene.Document()
                    doc.add(lucene.Field(”name”, filename,
                                         lucene.Field.Store.YES,
                                         lucene.Field.Index.NOT_ANALYZED))
                    doc.add(lucene.Field(”path”, path,
                                         lucene.Field.Store.YES,
                                         lucene.Field.Index.NOT_ANALYZED))
                    doc.add(lucene.Field(”url”, url,
                                         lucene.Field.Store.YES,
                                         lucene.Field.Index.NOT_ANALYZED))
                    doc.add(lucene.Field(”title”, title,
                                         lucene.Field.Store.YES,
                                         lucene.Field.Index.NOT_ANALYZED))
                    if len(contents) > 0:
                        dll=cdll.LoadLibrary(”F:\\ICTCLAS50_Windows_32_C\ICTCLAS
                        dll.ICTCLAS_Init(c_char_p(”F:\\ICTCLAS50_Windows_32_C”))
                        strlen = len(c_char_p(contents).value)
                        t =c_buffer(strlen*6)
                        bSuccess = dll.ICTCLAS_ParagraphProcess
                        (c_char_p(contents),c_int(strlen),t,c_int(0),0)
                        contents=t.value.decode(’gbk’).encode(’utf8’)
                        ##list=t.value.split()
                        ##print ’ ’.join(list)
                        dll.ICTCLAS_Exit()
                        doc.add(lucene.Field(”contents”, contents,
                                             lucene.Field.Store.NO,
                                                                  lucene.Field.Index.ANALYZED))
                    else:
                        print "warning:␣no␣content␣in␣%s" % filename
                    writer.addDocument(doc)
                except Exception, e:
                    print "Failed␣in␣indexDocs:", e

if __name__ == '__main__':
##    if len(sys.argv) < 2:
##        print IndexFiles.__doc__
##        sys.exit(1)
    lucene.initVM()
    print 'lucene', lucene.VERSION
    start = datetime.now()
    dic= open('F:\\html\index.txt')
    d = dic.readlines()
    dic.close()
    mydict = {}
    for word in d:
        key = word.split(';')[0]
        value = word.split(';')[1]
        mydict[key] = value
    try:
##        IndexFiles(sys.argv[1], "index", lucene.SimpleAnalyzer(lucene.Version.
        IndexFiles('F:\\html', "F:\\index", lucene.SimpleAnalyzer(lucene.Version
        end = datetime.now()
        print end - start
    except Exception, e:
        print "Failed:␣", e
```

So after this part, we can get the index set up in F:\index, making the last part easily to be done. And I use 'SimpleAnalyzer' in the indexing and searching part, which will make the whole experiment more execute.
And below are the pictures of the outcome in the interpreter and the index in file "F:\index".

2.1.3 Search According to the Query

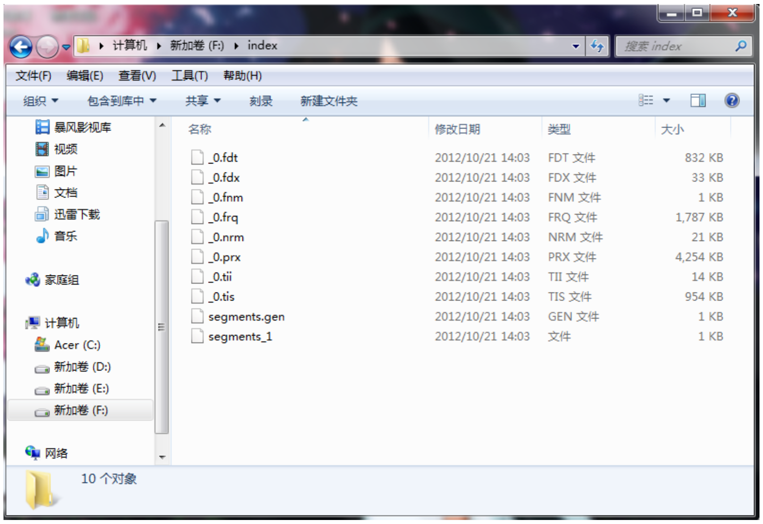
Things get really easy when they come to the last part. In this searching part, we only need to change the example a little bit, adding the outcome of title and url, which we have already made an index in the second part.
And that are the codes:

```
from lucene import \
    QueryParser, IndexSearcher, SimpleAnalyzer, SimpleFSDirectory, File, \
```



```
lucene 3.6.1
adding http0car0.com
adding http2012.igem.orgRegionsAsiaJamboreeresults
adding http2012.igem.orgTeamSJTU-BioX-Shanghai
adding http202.120.1.33foundationdonatedonate001.php
adding http202.120.143.2360
adding http202.120.227.6F
adding http202.120.24.209cycx
adding http202.120.24.209hr
adding http202.120.24.209management
adding http202.120.24.209marketing
adding http202.120.24.209strategymgt
adding http202.120.24.209yygl
adding http202.120.33.37english2
adding http202.120.35.11ReportServerPagesReportViewer.aspx2fExamArrange2fExamArrangeForOthersrsCommandRender
adding http202.120.35.11ReportServerPagesReportViewer.aspx2fExamArrange2fLessonArrangeForOthersrsCommandRender
adding http202.120.63.153innovationplangjjh
adding http202.120.63.153innovationplanshjh
adding http202.120.63.33985HomeLoginPage.aspx
adding http202.120.63.33pmishome
adding http202.120.63.33PMISHomeLoginPage.aspx
adding http218.1.73.12SearchScoreAndBulletindefault.asp
adding http218.1.73.12zhaosheng
adding http218.1.73.12zhaoshengindex.aspx
adding http3dcampus.sjtu.edu.cn
adding http56pin.com
adding http71bbs.people.com.cnindex.do
adding http985.sjtu.edu.cn
adding http985.sjtu.edu.cnindex.jsp
adding http985.sjtu.edu.cnnewsAnnouncement.jspurltypetree.TreeTempUrlwbtreeid1135
adding http985.sjtu.edu.cnnewsAnnouncement.jspurltypetree.TreeTempUrlwbtreeid1136
adding http985.sjtu.edu.cnnewsAnnouncement.jspurltypetree.TreeTempUrlwbtreeid1137
adding http985.sjtu.edu.cnnewsAnnouncement.jspurltypetree.TreeTempUrlwbtreeid1138
adding http985.sjtu.edu.cnnewsAnnouncementContent.jspurltypenews.NewsContentUrlwbtreeid1136wbnewsid1306
adding http985.sjtu.edu.cnnewsAnnouncementContent.jspurltypenews.NewsContentUrlwbtreeid1136wbnewsid1308
adding http985.sjtu.edu.cnnewsAnnouncementContent.jspurltypenews.NewsContentUrlwbtreeid1136wbnewsid1312
adding http985.sjtu.edu.cnnewsAnnouncementContent.jspurltypenews.NewsContentUrlwbtreeid1136wbnewsid1313
adding http985.sjtu.edu.cnnewsAnnouncementContent.jspurltypenews.NewsContentUrlwbtreeid1136wbnewsid1315
adding http985.sjtu.edu.cnnewsAnnouncementContent.jspurltypenews.NewsContentUrlwbtreeid1136wbnewsid1392
```

Figure 4: the outcome in interpreter



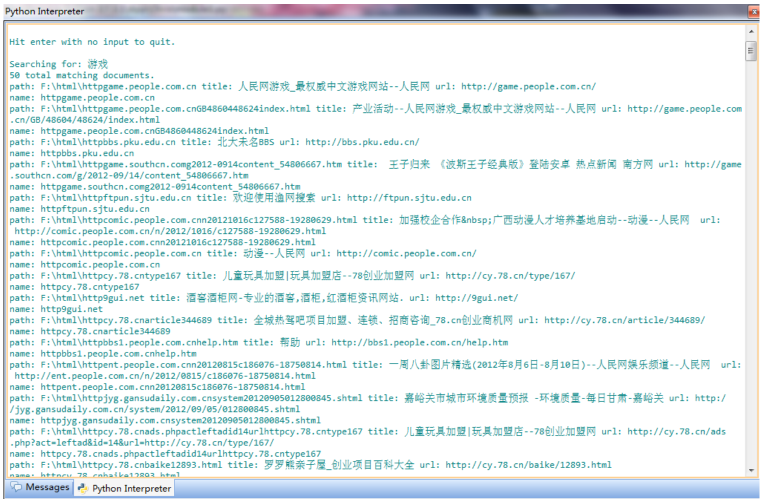
名称	修改日期	类型	大小
_0.fdt	2012/10/21 14:03	FDT 文件	832 KB
_0.fdx	2012/10/21 14:03	FDX 文件	33 KB
_0.fnm	2012/10/21 14:03	FNM 文件	1 KB
_0.frq	2012/10/21 14:03	FRQ 文件	1,787 KB
_0.nrm	2012/10/21 14:03	NRM 文件	21 KB
_0.prx	2012/10/21 14:03	PRX 文件	4,254 KB
_0.tii	2012/10/21 14:03	TII 文件	14 KB
_0.tis	2012/10/21 14:03	TIS 文件	954 KB
segments.gen	2012/10/21 14:03	GEN 文件	1 KB
segments_1	2012/10/21 14:03	文件	1 KB

Figure 5: the screenshot of F:\index

```
    VERSION, initVM, Version


"""
This script is loosely based on the Lucene (java implementation) demo class
org.apache.lucene.demo.SearchFiles.  It will prompt for a search query, then it
will search the Lucene index in the current directory called 'index' for the
search query entered against the 'contents' field.  It will then display the
'path' and 'name' fields for each of the hits it finds in the index.  Note that
search.close() is currently commented out because it causes a stack overflow in
some cases.
"""
def run(searcher, analyzer):
    while True:
        print
        print "Hit enter with no input to quit."
        command = raw_input("Query:")
        if command == '':
            return
        print
        print "Searching for:", command
        query = QueryParser(Version.LUCENE_CURRENT, "contents",
                            analyzer).parse(command)
        scoreDocs = searcher.search(query, 50).scoreDocs
        print "%s total matching documents." % len(scoreDocs)

        for scoreDoc in scoreDocs:
            doc = searcher.doc(scoreDoc.doc)
            print 'path:', doc.get("path"), 'title:', doc.get("title"),'url:', d


if __name__ == '__main__':
    STORE_DIR = "F:\\index"
    initVM()
    print 'lucene', VERSION
    directory = SimpleFSDirectory(File(STORE_DIR))
    searcher = IndexSearcher(directory, True)
    analyzer = SimpleAnalyzer(Version.LUCENE_CURRENT)
    run(searcher, analyzer)
    searcher.close()
```

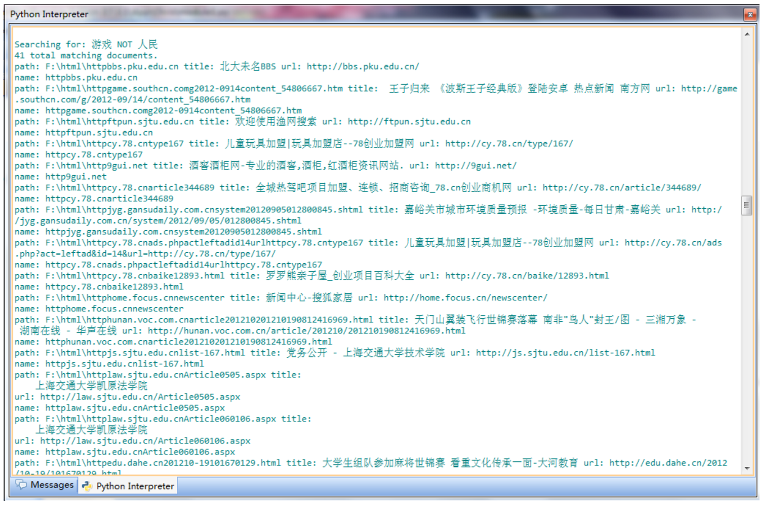
Then we can search the information in the pages we have crawled. These are the screenshots:

And thus, at last I finished the experiment working so hard, but luckily, this time, I was not troubled with the code problem as often as I used to be, thanks



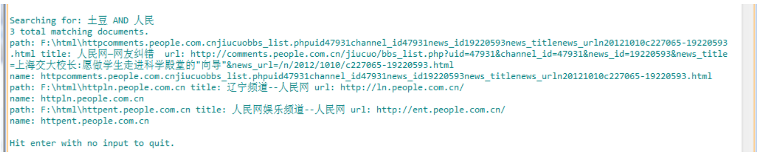
```
Hit enter with no input to quit.

Searching for: 游戏
50 total matching documents.
path: F:\html\httpgame.people.com.cn title: 人民网游戏_最权威中文游戏网站--人民网 url: http://game.people.com.cn/
name: httpgame.people.com.cn
path: F:\html\httpgame.people.com.cnGB4860448624index.html title: 产业活动--人民网游戏_最权威中文游戏网站--人民网 url: http://game.people.com
.cn/GB/48604/48624/index.html
name: httpgame.people.com.cnGB4860448624index.html
path: F:\html\httpbbs.pku.edu.cn title: 北大未名BBS url: http://bbs.pku.edu.cn/
name: httpbbs.pku.edu.cn
path: F:\html\httpgame.southcn.comg2012-0914content_54806667.htm title:  王子归来 《波斯王子经典版》登陆安卓 热点新闻 南方网 url: http://game
.southcn.com/g/2012-09/14/content_54806667.htm
name: httpgame.southcn.comg2012-0914content_54806667.htm
path: F:\html\httpftpun.sjtu.edu.cn title: 欢迎使用渔网搜索 url: http://ftpun.sjtu.edu.cn
name: httpftpun.sjtu.edu.cn
path: F:\html\httpcomic.people.com.cnn20121016c127588-19280629.html title: 加强校企合作&nbsp;广西动漫人才培养基地启动--动漫--人民网  url:
 http://comic.people.com.cn/n/2012/1016/c127588-19280629.html
name: httpcomic.people.com.cnn20121016c127588-19280629.html
path: F:\html\httpcomic.people.com.cn title: 动漫--人民网 url: http://comic.people.com.cn/
name: httpcomic.people.com.cn
path: F:\html\httpcy.78.cntype167 title: 儿童玩具加盟|玩具加盟店--78创业加盟网 url: http://cy.78.cn/type/167/
name: httpcy.78.cntype167
path: F:\html\http9gui.net title: 酒窖酒柜网-专业的酒窖,酒柜,红酒柜资讯网站. url: http://9gui.net/
name: http9gui.net
path: F:\html\httpcy.78.cnarticle344689 title: 全城热驾吧项目加盟、连锁、招商咨询_78.cn创业商机网 url: http://cy.78.cn/article/344689/
name: httpcy.78.cnarticle344689
path: F:\html\httpbbs1.people.com.cnhelp.htm title: 帮助 url: http://bbs1.people.com.cn/help.htm
name: httpbbs1.people.com.cnhelp.htm
path: F:\html\httpent.people.com.cnn20120815c186076-18750814.html title: 一周八卦图片精选(2012年8月6日-8月10日)--人民网娱乐频道--人民网  url:
 http://ent.people.com.cn/n/2012/0815/c186076-18750814.html
name: httpent.people.com.cnn20120815c186076-18750814.html
path: F:\html\httpjyg.gansudaily.com.cnsystem20120905012800845.shtml title: 嘉峪关市城市环境质量预报 -环境质量-每日甘肃-嘉峪关 url: http:/
/jyg.gansudaily.com.cn/system/2012/09/05/012800845.shtml
name: httpjyg.gansudaily.com.cnsystem20120905012800845.shtml
path: F:\html\httpcy.78.cnads.phpactleftadid14urlhttpcy.78.cntype167 title: 儿童玩具加盟|玩具加盟店--78创业加盟网 url: http://cy.78.cn/ads
.php?act=leftad&id=14&url=http://cy.78.cn/type/167/
name: httpcy.78.cnads.phpactleftadid14urlhttpcy.78.cntype167
path: F:\html\httpcy.78.cnbaike12893.html title: 罗罗熊亲子屋_创业项目百科大全 url: http://cy.78.cn/baike/12893.html
name: httpcy.78.cnbaike12893.html
```

Figure 6: the outcome in interpreter(1)



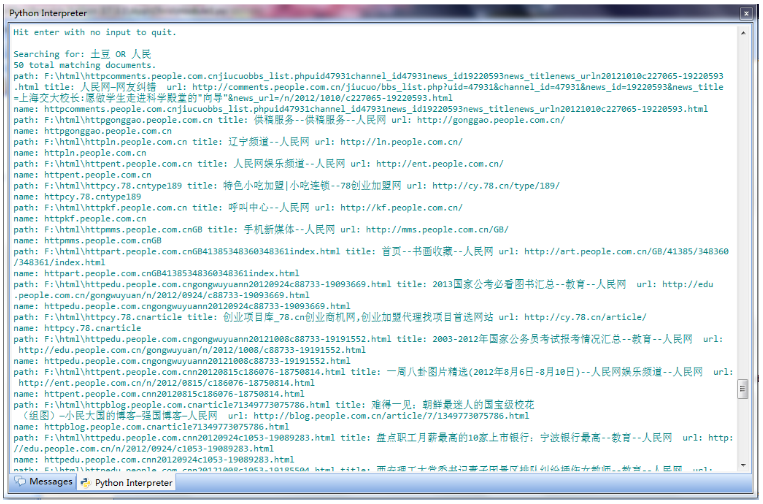
```
Searching for: 游戏 NOT 人民
41 total matching documents.
path: F:\html\httpbbs.pku.edu.cn title: 北大未名BBS url: http://bbs.pku.edu.cn/
name: httpbbs.pku.edu.cn
path: F:\html\httpgame.southcn.comg2012-0914content_54806667.htm title:  王子归来 《波斯王子经典版》登陆安卓 热点新闻 南方网 url: http://game
.southcn.com/g/2012-09/14/content_54806667.htm
name: httpgame.southcn.comg2012-0914content_54806667.htm
path: F:\html\httpftpun.sjtu.edu.cn title: 欢迎使用渔网搜索 url: http://ftpun.sjtu.edu.cn
name: httpftpun.sjtu.edu.cn
path: F:\html\httpcy.78.cntype167 title: 儿童玩具加盟|玩具加盟店--78创业加盟网 url: http://cy.78.cn/type/167/
name: httpcy.78.cntype167
path: F:\html\http9gui.net title: 酒窖酒柜网-专业的酒窖,酒柜,红酒柜资讯网站. url: http://9gui.net/
name: http9gui.net
path: F:\html\httpcy.78.cnarticle344689 title: 全城热驾吧项目加盟、连锁、招商咨询_78.cn创业商机网 url: http://cy.78.cn/article/344689/
name: httpcy.78.cnarticle344689
path: F:\html\httpjyg.gansudaily.com.cnsystem20120905012800845.shtml title: 嘉峪关市城市环境质量预报 -环境质量-每日甘肃-嘉峪关 url: http:/
/jyg.gansudaily.com.cn/system/2012/09/05/012800845.shtml
name: httpjyg.gansudaily.com.cnsystem20120905012800845.shtml
path: F:\html\httpcy.78.cnads.phpactleftadid14urlhttpcy.78.cntype167 title: 儿童玩具加盟|玩具加盟店--78创业加盟网 url: http://cy.78.cn/ads
.php?act=leftad&id=14&url=http://cy.78.cn/type/167/
name: httpcy.78.cnads.phpactleftadid14urlhttpcy.78.cntype167
path: F:\html\httpcy.78.cnbaike12893.html title: 罗罗熊亲子屋_创业项目百科大全 url: http://cy.78.cn/baike/12893.html
name: httpcy.78.cnbaike12893.html
path: F:\html\httphome.focus.cnnewscenter title: 新闻中心-搜狐家居 url: http://home.focus.cn/newscenter/
name: httphome.focus.cnnewscenter
path: F:\html\httphunan.voc.com.cnarticle201210201210190812416969.html title: 天门山翼装飞行世锦赛落幕 南非"鸟人"封王/图 - 三湘万象 -
 湖南在线 - 华声在线 url: http://hunan.voc.com.cn/article/201210/201210190812416969.html
name: httphunan.voc.com.cnarticle201210201210190812416969.html
path: F:\html\httpjs.sjtu.edu.cnlist-167.html title: 党务公开 - 上海交通大学技术学院 url: http://js.sjtu.edu.cn/list-167.html
name: httpjs.sjtu.edu.cnlist-167.html
path: F:\html\httplaw.sjtu.edu.cnArticle0505.aspx title:
    上海交通大学凯原法学院
url: http://law.sjtu.edu.cn/Article0505.aspx
name: httplaw.sjtu.edu.cnArticle0505.aspx
path: F:\html\httplaw.sjtu.edu.cnArticle060106.aspx title:
    上海交通大学凯原法学院
url: http://law.sjtu.edu.cn/Article060106.aspx
name: httplaw.sjtu.edu.cnArticle060106.aspx
path: F:\html\httpedu.dahe.cn201210-19101670129.html title: 大学生组队参加麻将世锦赛 看重文化传承一面-大河教育 url: http://edu.dahe.cn/2012
/10-19/101670129.html
```

Figure 7: the outcome in interpreter(2)



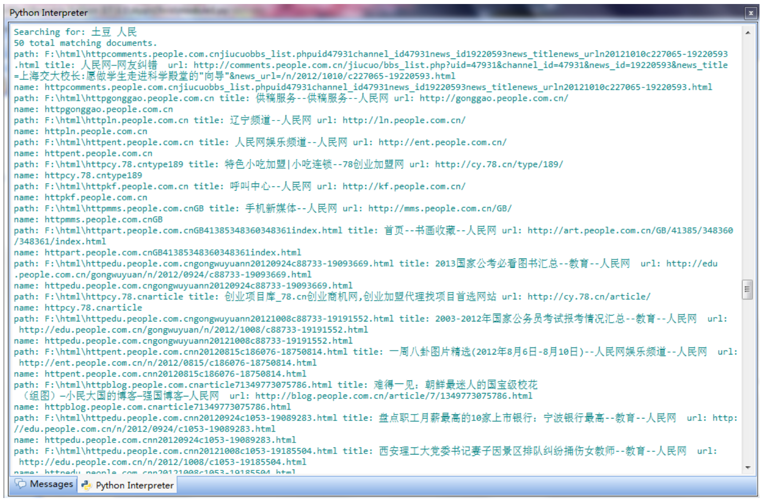
```
Searching for: 土豆 AND 人民
3 total matching documents.
path: F:\html\httpcomments.people.com.cnjiucuobbs_list.phpuid47931channel_id47931news_id19220593news_titlenews_urln20121010c227065-19220593
.html title: 人民网—网友纠错  url: http://comments.people.com.cn/jiucuo/bbs_list.php?uid=47931&channel_id=47931&news_id=19220593&news_title
=上海交大校长:愿做学生走进科学殿堂的"向导"&news_url=/n/2012/1010/c227065-19220593.html
name: httpcomments.people.com.cnjiucuobbs_list.phpuid47931channel_id47931news_id19220593news_titlenews_urln20121010c227065-19220593.html
path: F:\html\httpln.people.com.cn title: 辽宁频道--人民网 url: http://ln.people.com.cn/
name: httpln.people.com.cn
path: F:\html\httpent.people.com.cn title: 人民网娱乐频道--人民网 url: http://ent.people.com.cn/
name: httpent.people.com.cn

Hit enter with no input to quit.
```

Figure 8: the outcome in interpreter(3)



```
Hit enter with no input to quit.

Searching for: 土豆 OR 人民
50 total matching documents.
path: F:\html\httpcomments.people.com.cnjiucuobbs_list.phpuid47931channel_id47931news_id19220593news_titlenews_urln20121010c227065-19220593
.html title: 人民网—网友纠错  url: http://comments.people.com.cn/jiucuo/bbs_list.php?uid=47931&channel_id=47931&news_id=19220593&news_title
=上海交大校长:愿做学生走进科学殿堂的"向导"&news_url=/n/2012/1010/c227065-19220593.html
name: httpcomments.people.com.cnjiucuobbs_list.phpuid47931channel_id47931news_id19220593news_titlenews_urln20121010c227065-19220593.html
path: F:\html\httpgonggao.people.com.cn title: 供稿服务--供稿服务--人民网 url: http://gonggao.people.com.cn/
name: httpgonggao.people.com.cn
path: F:\html\httpln.people.com.cn title: 辽宁频道--人民网 url: http://ln.people.com.cn/
name: httpln.people.com.cn
path: F:\html\httpent.people.com.cn title: 人民网娱乐频道--人民网 url: http://ent.people.com.cn/
name: httpent.people.com.cn
path: F:\html\httpcy.78.cntype189 title: 特色小吃加盟|小吃连锁--78创业加盟网 url: http://cy.78.cn/type/189/
name: httpcy.78.cntype189
path: F:\html\httpkf.people.com.cn title: 呼叫中心--人民网 url: http://kf.people.com.cn/
name: httpkf.people.com.cn
path: F:\html\httpmms.people.com.cnGB title: 手机新媒体--人民网 url: http://mms.people.com.cn/GB/
name: httpmms.people.com.cnGB
path: F:\html\httpart.people.com.cnGB41385348360348361index.html title: 首页--书画收藏--人民网 url: http://art.people.com.cn/GB/41385/348360
/348361/index.html
name: httpart.people.com.cnGB41385348360348361index.html
path: F:\html\httpedu.people.com.cngongwuyuann20120924c88733-19093669.html title: 2013国家公考必看图书汇总--教育--人民网  url: http://edu
.people.com.cn/gongwuyuan/n/2012/0924/c88733-19093669.html
name: httpedu.people.com.cngongwuyuann20120924c88733-19093669.html
path: F:\html\httpcy.78.cnarticle title: 创业项目库_78.cn创业商机网,创业加盟代理找项目首选网站 url: http://cy.78.cn/article/
name: httpcy.78.cnarticle
path: F:\html\httpedu.people.com.cngongwuyuann20121008c88733-19191552.html title: 2003-2012年国家公务员考试报考情况汇总--教育--人民网  url:
 http://edu.people.com.cn/gongwuyuan/n/2012/1008/c88733-19191552.html
name: httpedu.people.com.cngongwuyuann20121008c88733-19191552.html
path: F:\html\httpent.people.com.cnn20120815c186076-18750814.html title: 一周八卦图片精选(2012年8月6日-8月10日)--人民网娱乐频道--人民网  url:
 http://ent.people.com.cn/n/2012/0815/c186076-18750814.html
name: httpent.people.com.cnn20120815c186076-18750814.html
path: F:\html\httpblog.people.com.cnarticle71349773075786.html title: 难得一见：朝鲜最迷人的国宝级校花
 （组图）—小民大国的博客—强国博客—人民网  url: http://blog.people.com.cn/article/7/1349773075786.html
name: httpblog.people.com.cnarticle71349773075786.html
path: F:\html\httpedu.people.com.cnn20120924c1053-19089283.html title: 盘点职工月薪最高的10家上市银行：宁波银行最高--教育--人民网  url: http:
//edu.people.com.cn/n/2012/0924/c1053-19089283.html
name: httpedu.people.com.cnn20120924c1053-19089283.html
path: F:\html\httpedu.people.com.cnn20121008c1053-19185504.html title: 西安理工大党委书记妻子因景区排队纠纷捅伤女教师--教育--人民网  url:
```

Figure 9: the outcome in interpreter(4)



```
Searching for: 土豆 人民
50 total matching documents.
path: F:\html\httpcomments.people.com.cnjiucuobbs_list.phpuid47931channel_id47931news_id19220593news_titlenews_urln20121010c227065-19220593
.html title: 人民网—网友纠错  url: http://comments.people.com.cn/jiucuo/bbs_list.php?uid=47931&channel_id=47931&news_id=19220593&news_title
=上海交大校长:愿做学生走进科学殿堂的"向导"&news_url=/n/2012/1010/c227065-19220593.html
name: httpcomments.people.com.cnjiucuobbs_list.phpuid47931channel_id47931news_id19220593news_titlenews_urln20121010c227065-19220593.html
path: F:\html\httpgonggao.people.com.cn title: 供稿服务--供稿服务--人民网 url: http://gonggao.people.com.cn/
name: httpgonggao.people.com.cn
path: F:\html\httpln.people.com.cn title: 辽宁频道--人民网 url: http://ln.people.com.cn/
name: httpln.people.com.cn
path: F:\html\httpent.people.com.cn title: 人民网娱乐频道--人民网 url: http://ent.people.com.cn/
name: httpent.people.com.cn
path: F:\html\httpcy.78.cntype189 title: 特色小吃加盟|小吃连锁--78创业加盟网 url: http://cy.78.cn/type/189/
name: httpcy.78.cntype189
path: F:\html\httpkf.people.com.cn title: 呼叫中心--人民网 url: http://kf.people.com.cn/
name: httpkf.people.com.cn
path: F:\html\httpmms.people.com.cnGB title: 手机新媒体--人民网 url: http://mms.people.com.cn/GB/
name: httpmms.people.com.cnGB
path: F:\html\httpart.people.com.cnGB41385348360348361index.html title: 首页--书画收藏--人民网 url: http://art.people.com.cn/GB/41385/348360
/348361/index.html
name: httpart.people.com.cnGB41385348360348361index.html
path: F:\html\httpedu.people.com.cngongwuyuann20120924c88733-19093669.html title: 2013国家公考必看图书汇总--教育--人民网  url: http://edu
.people.com.cn/gongwuyuan/n/2012/0924/c88733-19093669.html
name: httpedu.people.com.cngongwuyuann20120924c88733-19093669.html
path: F:\html\httpcy.78.cnarticle title: 创业项目库_78.cn创业商机网,创业加盟代理找项目首选网站 url: http://cy.78.cn/article/
name: httpcy.78.cnarticle
path: F:\html\httpedu.people.com.cngongwuyuann20121008c88733-19191552.html title: 2003-2012年国家公务员考试报考情况汇总--教育--人民网  url:
 http://edu.people.com.cn/gongwuyuan/n/2012/1008/c88733-19191552.html
name: httpedu.people.com.cngongwuyuann20121008c88733-19191552.html
path: F:\html\httpent.people.com.cnn20120815c186076-18750814.html title: 一周八卦图片精选(2012年8月6日-8月10日)--人民网娱乐频道--人民网  url:
 http://ent.people.com.cn/n/2012/0815/c186076-18750814.html
name: httpent.people.com.cnn20120815c186076-18750814.html
path: F:\html\httpblog.people.com.cnarticle71349773075786.html title: 难得一见：朝鲜最迷人的国宝级校花
 （组图）—小民大国的博客—强国博客—人民网  url: http://blog.people.com.cn/article/7/1349773075786.html
name: httpblog.people.com.cnarticle71349773075786.html
path: F:\html\httpedu.people.com.cnn20120924c1053-19089283.html title: 盘点职工月薪最高的10家上市银行：宁波银行最高--教育--人民网  url: http:
//edu.people.com.cn/n/2012/0924/c1053-19089283.html
name: httpedu.people.com.cnn20120924c1053-19089283.html
path: F:\html\httpedu.people.com.cnn20121008c1053-19185504.html title: 西安理工大党委书记妻子因景区排队纠纷捅伤女教师--教育--人民网  url:
 http://edu.people.com.cn/n/2012/1008/c1053-19185504.html
name: httpedu.people.com.cnn20121008c1053-19185504.html
```

Figure 10: the outcome in interpreter(5)

to the ”Q.ppt”.

2.2 The Second part of the Experiment

It easy to find that the crawler part of the latter half of lab3 is nothing different from the first part, so I will only give out the index and search part of each experiment.

2.2.1 Make Index with the Information of Site and Search It

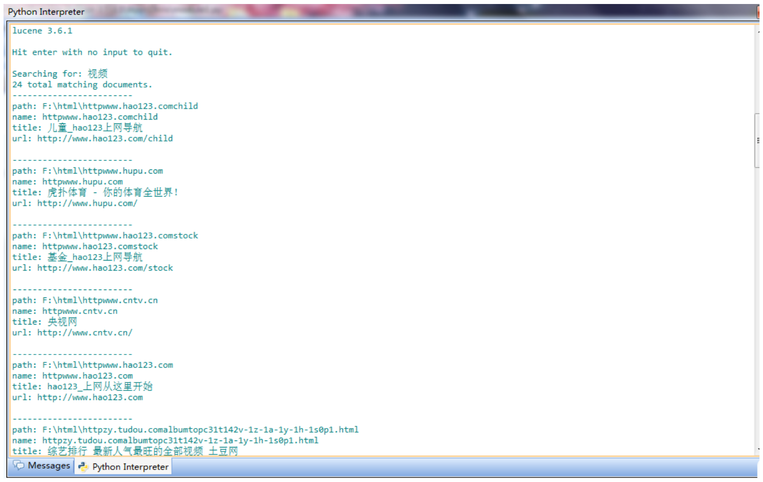
There is little different between this program and the indexing program of the first part, except that we need to find the domain name of the website and add it to the final index. And I happened to find an easy way to get the domain name using urllib, and I will present it in the following codes. By the way, I use the word 'site' instead of 'domain name' in my program.

```python
import sys, os, lucene, threading, time,chardet,urllib2
from datetime import datetime
from BeautifulSoup import BeautifulSoup
from ctypes import *
import urllib

"""
This class is loosely based on the Lucene (java implementation) demo class
org.apache.lucene.demo.IndexFiles. It will take a directory as an argument
and will index all of the files in that directory and downward recursively.
It will index on the file path, the file name and the file contents. The
resulting Lucene index will be placed in the current directory and called
'index'.
"""

class Ticker(object):

    def __init__(self):
        self.tick = True

    def run(self):
        while self.tick:
            sys.stdout.write('.')
            sys.stdout.flush()
            time.sleep(1.0)

class IndexFiles(object):
    """Usage: python IndexFiles <doc_directory>"""

    def __init__(self, root, storeDir, analyzer):

        if not os.path.exists(storeDir):
            os.mkdir(storeDir)
        store = lucene.SimpleFSDirectory(lucene.File(storeDir))
        writer = lucene.IndexWriter(store, analyzer, True,
                                    lucene.IndexWriter.MaxFieldLength.LIMITED)
        writer.setMaxFieldLength(1048576)
        self.indexDocs(root, writer)
        ticker = Ticker()
        print 'optimizing index',
        threading.Thread(target=ticker.run).start()
        writer.optimize()
        writer.close()
        ticker.tick = False
        print 'done'

    def indexDocs(self, root, writer):
        for root, dirnames, filenames in os.walk(root):
            for filename in filenames:
                if filename.endswith('.txt'):
                    continue
                print "adding", filename
                try:
                    path = os.path.join(root, filename)
                    file = open(path)
                    buf = file.read()
                    contents=buf
                    result = chardet.detect(buf)['encoding']
                    if result=='GB2312':
                        contents = buf.decode('gbk').encode('utf8')
                    file.close()
                    soup=BeautifulSoup(contents)
                    url=mydict[filename]
                    proto, rest = urllib.splittype(url)
                    site, rest = urllib.splithost(rest)
                    title=str(soup.head.title.string).decode('utf8')
                    contents=''.join(soup.findAll(text=True))
                    doc = lucene.Document()
                    doc.add(lucene.Field("name", filename,
                                         lucene.Field.Store.YES,
                                         lucene.Field.Index.NOT_ANALYZED))
                    doc.add(lucene.Field("path", path,
                                         lucene.Field.Store.YES,
                                         lucene.Field.Index.NOT_ANALYZED))
                    doc.add(lucene.Field("url", url,
                                         lucene.Field.Store.YES,
                                         lucene.Field.Index.NOT_ANALYZED))
                    doc.add(lucene.Field("title", title,
                                         lucene.Field.Store.YES,
                                         lucene.Field.Index.NOT_ANALYZED))
                    doc.add(lucene.Field("site", site,
                                         lucene.Field.Store.YES,
                                         lucene.Field.Index.ANALYZED))
                    if len(contents) > 0:
                        dll=cdll.LoadLibrary("F:\\ICTCLAS50_Windows_32_C\ICTCLAS
                        dll.ICTCLAS_Init(c_char_p("F:\\ICTCLAS50_Windows_32_C"))
                        strlen = len(c_char_p(contents).value)
                        t =c_buffer(strlen*6)
                        bSuccess = dll.ICTCLAS_ParagraphProcess
                        (c_char_p(contents),c_int(strlen),t,c_int(0),0)
                        contents=t.value.decode('gbk').encode('utf8')
                        ##list=t.value.split()
                        ##print ' '.join(list)
                        dll.ICTCLAS_Exit()
                        doc.add(lucene.Field("contents", contents,
                                             lucene.Field.Store.NO,
                                             lucene.Field.Index.ANALYZED))
                    else:
                        print "warning:␣no␣content␣in␣%s" % filename
                    writer.addDocument(doc)
                except Exception, e:
                    print "Failed␣in␣indexDocs:", e

if __name__ == '__main__':
##    if len(sys.argv) < 2:
##        print IndexFiles.__doc__
##        sys.exit(1)
    lucene.initVM()
    print 'lucene', lucene.VERSION
    start = datetime.now()
    dic= open('F:\\html\index.txt')
    d = dic.readlines()
    dic.close()
    mydict = {}
    for word in d:
        key = word.split(';')[0]
        value = word.split(';')[1]
        mydict[key] = value
    try:
##        IndexFiles(sys.argv[1], "index", lucene.WhitespaceAnalyzer(lucene.Vers
        IndexFiles('F:\\html', "F:\\index", lucene.WhitespaceAnalyzer(lucene.Ver
        end = datetime.now()
        print end - start
    except Exception, e:
        print "Failed:␣", e
```

As for the search part, we only need to add the keyword 'site' in the program to make it possible to search the contents in a certain site. In this part, we have to use BooleanQuery to accomplish the query consisted of several kinds of keywords, which can be learnt in the ppt file given.

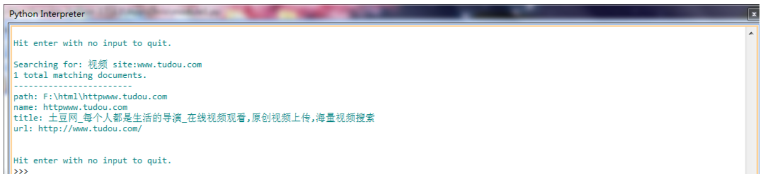
```
from lucene import \
    QueryParser, IndexSearcher, WhitespaceAnalyzer, SimpleFSDirectory, File, \
    VERSION, initVM, Version, BooleanQuery, BooleanClause


"""
This script is loosely based on the Lucene (java implementation) demo class
org.apache.lucene.demo.SearchFiles. It will prompt for a search query, then it
will search the Lucene index in the current directory called 'index' for the
search query entered against the 'contents' field. It will then display the
'path' and 'name' fields for each of the hits it finds in the index. Note that
search.close() is currently commented out because it causes a stack overflow in
some cases.
"""

def parseCommand(command):
    '''
    input: C title:T author:A language:L
    output: {'contents':C, 'title':T, 'author':A, 'language':L}

    Sample:
    input:'contenance title:henri language:french author:william shakespeare'
    output:{'author': ' william shakespeare',
                    'language': ' french',
                    'contents': ' contenance',
                    'title': ' henri'}
    '''
    allowed_opt = ['site']
    command_dict = {}
    opt = 'contents'
    for i in command.split('␣'):
        if ':' in i:
            opt, value = i.split(':')[:2]
            opt = opt.lower()
            if opt in allowed_opt and value != '':
                command_dict[opt] = command_dict.get(opt, '') + '␣' + value
        else:
            command_dict[opt] = command_dict.get(opt, '') + ' ' + i
    return command_dict


def run(searcher, analyzer):
    while True:
        print
        print "Hit enter with no input to quit."
        command = raw_input("Query:")
        if command == '':
            return

        print
        print "Searching for:", command

        command_dict = parseCommand(command)
        querys = BooleanQuery()
        for k,v in command_dict.iteritems():
            query = QueryParser(Version.LUCENE_CURRENT, k,
                                analyzer).parse(v)
            querys.add(query, BooleanClause.Occur.MUST)
        scoreDocs = searcher.search(querys, 50).scoreDocs
        print "%s total matching documents." % len(scoreDocs)

        for scoreDoc in scoreDocs:
            doc = searcher.doc(scoreDoc.doc)
##            explanation = searcher.explain(query, scoreDoc.doc)
            print "------------------------------"
            print 'path:', doc.get("path")
            print 'name:', doc.get("name")
            print 'title:', doc.get('title')
            print 'url:', doc.get('url')
##            print explanation


if __name__ == '__main__':
    STORE_DIR = "F:\\index"
    initVM()
    print 'lucene', VERSION
    directory = SimpleFSDirectory(File(STORE_DIR))
    searcher = IndexSearcher(directory, True)
    analyzer = WhitespaceAnalyzer(Version.LUCENE_CURRENT)
    run(searcher, analyzer)
    searcher.close()
```

In this part, to make it fast, I only get 30 pages as an example, you can crawl more pages by modifying the variable in the crawler program. And here are some screenshots of the effect of my program.



```
lucene 3.6.1

Hit enter with no input to quit.

Searching for: 视频
24 total matching documents.
------------------------
path: F:\html\httpwww.hao123.comchild
name: httpwww.hao123.comchild
title: 儿童_hao123上网导航
url: http://www.hao123.com/child

------------------------
path: F:\html\httpwww.hupu.com
name: httpwww.hupu.com
title: 虎扑体育 - 你的体育全世界!
url: http://www.hupu.com/

------------------------
path: F:\html\httpwww.hao123.comstock
name: httpwww.hao123.comstock
title: 基金_hao123上网导航
url: http://www.hao123.com/stock

------------------------
path: F:\html\httpwww.cntv.cn
name: httpwww.cntv.cn
title: 央视网
url: http://www.cntv.cn/

------------------------
path: F:\html\httpwww.hao123.com
name: httpwww.hao123.com
title: hao123_上网从这里开始
url: http://www.hao123.com

------------------------
path: F:\html\httpzy.tudou.comalbumtopc31t142v-1z-1a-1y-1h-1s0p1.html
name: httpzy.tudou.comalbumtopc31t142v-1z-1a-1y-1h-1s0p1.html
title: 综艺排行 最新人气最旺的全部视频 土豆网
```

Figure 11: the outcome of the site part(1)



```
Hit enter with no input to quit.

Searching for: 视频 site:www.tudou.com
1 total matching documents.
------------------------
path: F:\html\httpwww.tudou.com
name: httpwww.tudou.com
title: 土豆网_每个人都是生活的导演_在线视频观看,原创视频上传,海量视频搜索
url: http://www.tudou.com/


Hit enter with no input to quit.
>>>
```

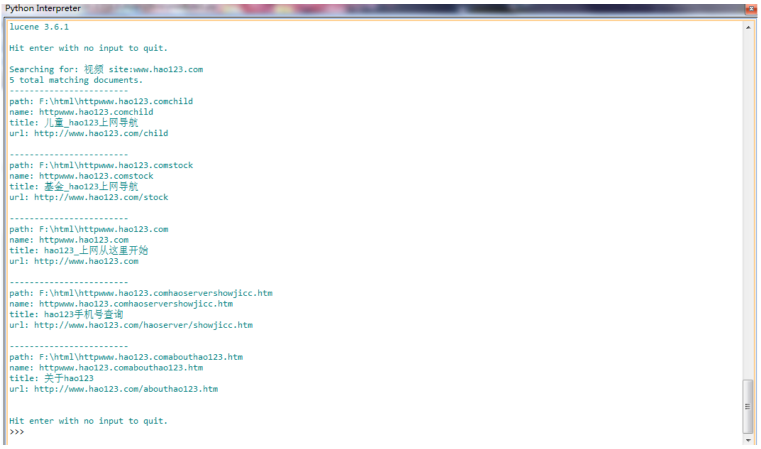
Figure 12: the outcome of the site part(2)

2.2.2 Index and Search for the Pictures

Well, after finish this part, I have to say it is not as easy as it seems to be at first. It is not because we have to get the url of the picture, the url of the website it's on or the title of the web page, but is the difficulties to get information, or contents of the pictures. Since the structure of the website is quite complex, I really took some time to get enough information I need to search certain pictures.

I choose "http://www.ommoo.com/" to be the page I'm going to index, which is a website offering pictures of the desktop of your computer. So after analyzing the structure of the website for a really hard time, I get the following codes at last, which can make a quite exact index of these pictures.

By the way, if you want to use the index program on other websites, you will



```
lucene 3.6.1

Hit enter with no input to quit.

Searching for: 视频 site:www.hao123.com
5 total matching documents.
------------------------
path: F:\html\httpwww.hao123.comchild
name: httpwww.hao123.comchild
title: 儿童_hao123上网导航
url: http://www.hao123.com/child

------------------------
path: F:\html\httpwww.hao123.comstock
name: httpwww.hao123.comstock
title: 基金_hao123上网导航
url: http://www.hao123.com/stock

------------------------
path: F:\html\httpwww.hao123.com
name: httpwww.hao123.com
title: hao123_上网从这里开始
url: http://www.hao123.com

------------------------
path: F:\html\httpwww.hao123.comhaoservershowjicc.htm
name: httpwww.hao123.comhaoservershowjicc.htm
title: hao123手机号查询
url: http://www.hao123.com/haoserver/showjicc.htm

------------------------
path: F:\html\httpwww.hao123.comabouthao123.htm
name: httpwww.hao123.comabouthao123.htm
title: 关于hao123
url: http://www.hao123.com/abouthao123.htm


Hit enter with no input to quit.
>>>
```

Figure 13: the outcome of the site part(3)

have to re-analyse the structure of that site and modify some of the variables in the program so that it can fit the target website.

```python
import sys, os, lucene, threading, time,chardet,urllib2,re
from datetime import datetime
from BeautifulSoup import BeautifulSoup
from ctypes import *
import urllib
import Queue
import urlparse

"""
This class is loosely based on the Lucene (java implementation) demo class
org.apache.lucene.demo.IndexFiles.  It will take a directory as an argument
and will index all of the files in that directory and downward recursively.
It will index on the file path, the file name and the file contents. The
resulting Lucene index will be placed in the current directory and called
'index'.
"""

class Ticker(object):

    def __init__(self):
        self.tick = True

    def run(self):
        while self.tick:
            sys.stdout.write('.')
            sys.stdout.flush()
            time.sleep(1.0)

class IndexFiles(object):
    """Usage: python IndexFiles <doc_directory>"""

    def __init__(self, root, storeDir, analyzer):

        if not os.path.exists(storeDir):
            os.mkdir(storeDir)
        store = lucene.SimpleFSDirectory(lucene.File(storeDir))
        writer = lucene.IndexWriter(store, analyzer, True,
                                     lucene.IndexWriter.MaxFieldLength.LIMITED)
        writer.setMaxFieldLength(1048576)
        self.indexDocs(root, writer)
        ticker = Ticker()
        print 'optimizing index',
        threading.Thread(target=ticker.run).start()
        writer.optimize()
        writer.close()
        ticker.tick = False
        print 'done'

    def indexDocs(self, root, writer):
        for root, dirnames, filenames in os.walk(root):
            for filename in filenames:
                if filename.endswith('.txt'):
                    continue
                print "adding", filename
                try:
                    path = os.path.join(root, filename)
                    file = open(path)
                    buf = file.read()
                    contents=buf
                    result = chardet.detect(buf)['encoding']
                    if result=='GB2312':
                        contents = buf.decode('gbk').encode('utf8')
                    file.close()
                    soup=BeautifulSoup(contents)
                    url=mydict[filename]
                    proto, rest = urllib.splittype(url)
                    site, rest = urllib.splithost(rest)
                    title=str(soup.head.title.string.strip()).decode('utf8')
                    flag2=0
                    for i in soup.findAll('img'):
                        contents=""
                        flag1=0
                        flag3=0
                        try:
                            contents=contents+' '+i['alt']
                        except:
                            pass
                        tempurl=i['src']
                        imgurl=urlparse.urljoin(url,tempurl)
                        temp=i.parent.parent
                        try:
                            photoid=temp.find('a')['data-photo-id']
                            flag1=1
                        except:
                            pass
                        try:
                            picid=temp.parent.find('article')['id']
                            flag3=1
                        except:
                            pass
                        try:
                            for t in temp.findAll('b'):
                                try:
                                    contents=contents+' '+t.string.strip()
                                except:
                                    pass
                        except:
                            pass
                        try:
                            for k in temp.findAll('p'):
                                try:
                                    contents=contents+' '+k.string.strip()
                                except:
                                    pass
                        except:
                            pass
                        try:
                            for j in temp.findAll('span',{'class':'title'}):
                                try:
                                    contents=contents+' '+j.string.strip()
                                except:
                                    pass
                        except:
                            pass
                        if flag1==1:
                            timetowait=0
                            try:
                                for p in temp.parent.findAll('div',{'class':'car
                                    if timetowait<flag2:
                                        timetowait+=1
                                        continue
                                    contents=contents+' '+p.string.strip()
                                    flag2+=1
                                    break
                            except:
                                pass
                        if flag3==1:
                            try:
                                for q in temp.parent.findAll('div',{'class':'pos
                                    r=q.find('h1')
                                    contents=contents+' '+str(r.string).decode('
                                    break
                            except:
                                pass
                        contents=contents.strip()
                        doc = lucene.Document()
                        doc.add(lucene.Field("imgurl", imgurl,
                                             lucene.Field.Store.YES,
                                             lucene.Field.Index.NOT_ANALYZED))
                        doc.add(lucene.Field("url", url,
                                             lucene.Field.Store.YES,
                                             lucene.Field.Index.NOT_ANALYZED))
                        doc.add(lucene.Field("title", title,
                                             lucene.Field.Store.YES,
                                             lucene.Field.Index.NOT_ANALYZED))
                        if len(contents) > 0:
                            dll=cdll.LoadLibrary("F:\\ICTCLAS50_Windows_32_C\ICT
                            dll.ICTCLAS_Init(c_char_p("F:\\ICTCLAS50_Windows_32_
                            strlen = len(c_char_p(contents).value)
                            t =c_buffer(strlen*6)
                            bSuccess = dll.ICTCLAS_ParagraphProcess(c_char_p(con
                            contents=t.value.decode('gbk').encode('utf8')
                            ##list=t.value.split()
                            ##print ' '.join(list)
                            dll.ICTCLAS_Exit()
                            doc.add(lucene.Field("contents", contents,
                                                 lucene.Field.Store.NO,
                                                 lucene.Field.Index.ANALYZED))
                        else:
                            print "warning: no content in part of %s" % filename
                        writer.addDocument(doc)
                except Exception, e:
                    print "Failed␣in␣indexDocs:", e

if __name__ == '__main__':
##      if len(sys.argv) < 2:
##          print IndexFiles.__doc__
##          sys.exit(1)
    lucene.initVM()
    print 'lucene', lucene.VERSION
    start = datetime.now()
    dic= open('F:\\html\index.txt')
    d = dic.readlines()
    dic.close()
    mydict = {}
    for word in d:
        key = word.split(';')[0]
        value = word.split(';')[1]
        mydict[key] = value
    try:
##          IndexFiles(sys.argv[1], "index", lucene.WhitespaceAnalyzer(lucene.Vers
        IndexFiles('F:\\html', "F:\\imgindex", lucene.WhitespaceAnalyzer(lucene.
        end = datetime.now()
        print end - start
    except Exception, e:
        print "Failed:␣", e
```

And after finishing the index part, it's pretty easy to accomplish the rest part. The code of the searching part is as follows, as it's easy, I won't explain it explicitly.

```
from lucene import \
    QueryParser, IndexSearcher, WhitespaceAnalyzer, SimpleFSDirectory, File, \
    VERSION, initVM, Version


"""
This script is loosely based on the Lucene (java implementation) demo class
org.apache.lucene.demo.SearchFiles. It will prompt for a search query, then it
will search the Lucene index in the current directory called 'index' for the
search query entered against the 'contents' field. It will then display the
'path' and 'name' fields for each of the hits it finds in the index. Note that
search.close() is currently commented out because it causes a stack overflow in
some cases.
"""
def run(searcher, analyzer):
    while True:
        print
        print "Hit␣enter␣with␣no␣input␣to␣quit."
        command = raw_input("Query:")
        if command == '':
            return
        print
        print "Searching␣for:", command
        query = QueryParser(Version.LUCENE_CURRENT, "contents",
                            analyzer).parse(command)
        scoreDocs = searcher.search(query, 50).scoreDocs
        print "%s␣total␣matching␣documents." % len(scoreDocs)

        for scoreDoc in scoreDocs:
            doc = searcher.doc(scoreDoc.doc)
            print 'title:', doc.get("title"), 'url:',doc.get("url"), 'imgurl:',



if __name__ == '__main__':
    STORE_DIR = "F:\\imgindex"
    initVM()
    print 'lucene', VERSION
    directory = SimpleFSDirectory(File(STORE_DIR))
    searcher = IndexSearcher(directory, True)
    analyzer = WhitespaceAnalyzer(Version.LUCENE_CURRENT)
    run(searcher, analyzer)
    searcher.close()
```

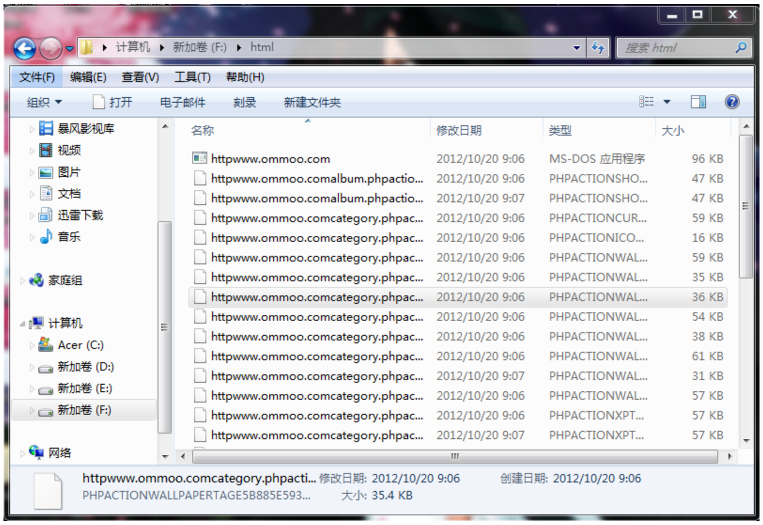
And here are also some screenshots of the files I crawled in the target file ’F:\imgindex’, and the pictures of the searching outcomes.

3 The Problems I Met in the Experiment and My Solution

Well, there are so many problems I have met in the experiment.
In the first part, I find that sometimes I can’t get the content of the pages for the node problem, so I just use ’chardet’ in python to find what the code method is and then transform them all to utf8. And then I find a problem in establish the ”F:\html” file, and then I find these codes and add them to my program:

```
if not os.path.exists("F:\\html"):
    os.mkdir("F:\\html")
```

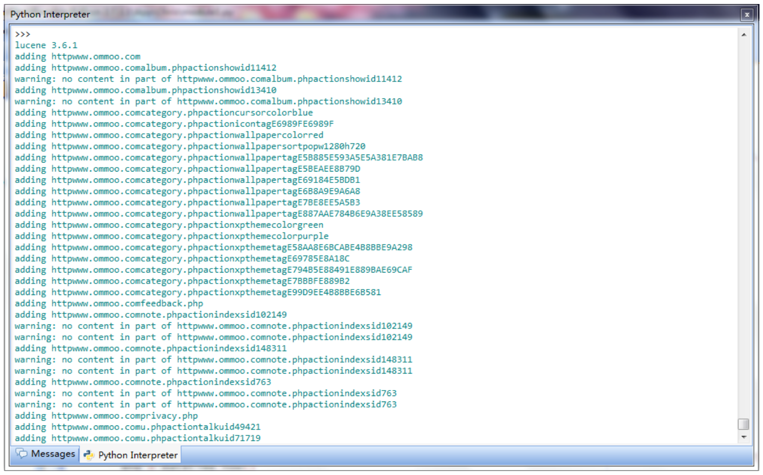
These codes mean that if ”F:\html” doesn’t exist, then it will be set up. And then comes the problem of saving the url and filenames of each page. To make it easy, I save each pair of them in the same line us ”;” to separate them from



名称	修改日期	类型	大小
httpwww.ommoo.com	2012/10/20 9:06	MS-DOS 应用程序	96 KB
httpwww.ommoo.comalbum.phpactio...	2012/10/20 9:06	PHPACTIONSHO...	47 KB
httpwww.ommoo.comalbum.phpactio...	2012/10/20 9:07	PHPACTIONSHO...	47 KB
httpwww.ommoo.comcategory.phpac...	2012/10/20 9:06	PHPACTIONCUR...	59 KB
httpwww.ommoo.comcategory.phpac...	2012/10/20 9:06	PHPACTIONICO...	16 KB
httpwww.ommoo.comcategory.phpac...	2012/10/20 9:06	PHPACTIONWAL...	59 KB
httpwww.ommoo.comcategory.phpac...	2012/10/20 9:06	PHPACTIONWAL...	35 KB
httpwww.ommoo.comcategory.phpac...	2012/10/20 9:06	PHPACTIONWAL...	36 KB
httpwww.ommoo.comcategory.phpac...	2012/10/20 9:06	PHPACTIONWAL...	54 KB
httpwww.ommoo.comcategory.phpac...	2012/10/20 9:06	PHPACTIONWAL...	38 KB
httpwww.ommoo.comcategory.phpac...	2012/10/20 9:06	PHPACTIONWAL...	61 KB
httpwww.ommoo.comcategory.phpac...	2012/10/20 9:07	PHPACTIONWAL...	31 KB
httpwww.ommoo.comcategory.phpac...	2012/10/20 9:06	PHPACTIONWAL...	57 KB
httpwww.ommoo.comcategory.phpac...	2012/10/20 9:06	PHPACTIONXPT...	57 KB
httpwww.ommoo.comcategory.phpac...	2012/10/20 9:07	PHPACTIONXPT...	57 KB

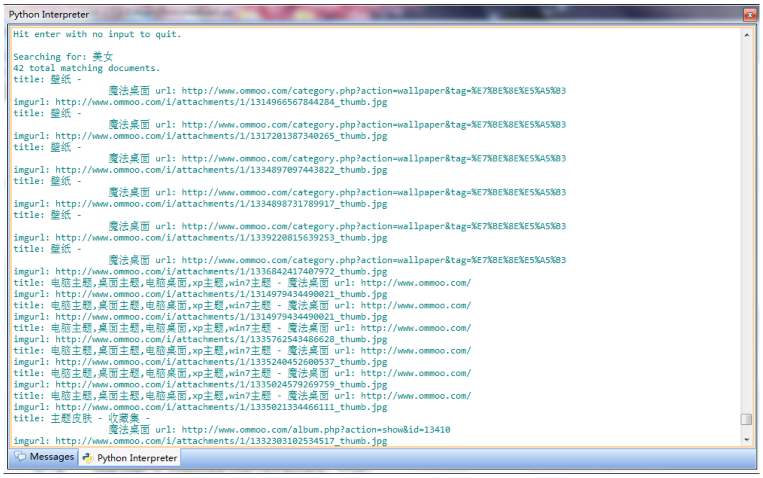
httpwww.ommoo.comcategory.phpacti... 修改日期: 2012/10/20 9:06 创建日期: 2012/10/20 9:06
PHPACTIONWALLPAPERTAGE5B885E593... 大小: 35.4 KB

Figure 14: the files we get in 'F:\html'



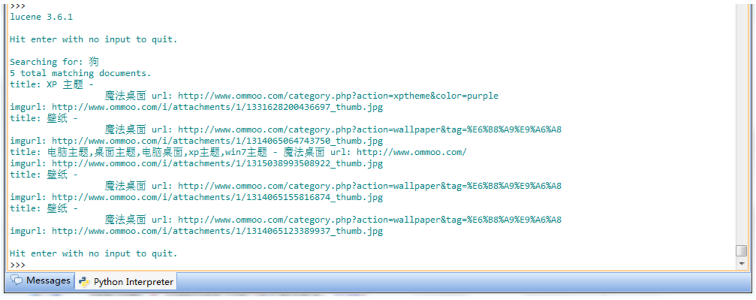
```
>>>
lucene 3.6.1
adding httpwww.ommoo.com
adding httpwww.ommoo.comalbum.phpactionshowid11412
warning: no content in part of httpwww.ommoo.comalbum.phpactionshowid11412
adding httpwww.ommoo.comalbum.phpactionshowid13410
warning: no content in part of httpwww.ommoo.comalbum.phpactionshowid13410
adding httpwww.ommoo.comcategory.phpactioncursorcolorblue
adding httpwww.ommoo.comcategory.phpactionicontagE6989FE6989F
adding httpwww.ommoo.comcategory.phpactionwallpapercolorred
adding httpwww.ommoo.comcategory.phpactionwallpapersortpopw1280h720
adding httpwww.ommoo.comcategory.phpactionwallpapertagE5B885E593A5E5A381E7BAB8
adding httpwww.ommoo.comcategory.phpactionwallpapertagE5BEAEE8B79D
adding httpwww.ommoo.comcategory.phpactionwallpapertagE69184E5BDB1
adding httpwww.ommoo.comcategory.phpactionwallpapertagE6B8A9E9A6A8
adding httpwww.ommoo.comcategory.phpactionwallpapertagE7BE8EE5A5B3
adding httpwww.ommoo.comcategory.phpactionwallpapertagE887AAE784B6E9A38EE58589
adding httpwww.ommoo.comcategory.phpactionxpthemecolorgreen
adding httpwww.ommoo.comcategory.phpactionxpthemecolorpurple
adding httpwww.ommoo.comcategory.phpactionxpthemetagE58AA8E6BCABE4B8BBE9A298
adding httpwww.ommoo.comcategory.phpactionxpthemetagE69785E8A18C
adding httpwww.ommoo.comcategory.phpactionxpthemetagE794B5E88491E889BAE69CAF
adding httpwww.ommoo.comcategory.phpactionxpthemetagE7BBBFE889B2
adding httpwww.ommoo.comcategory.phpactionxpthemetagE99D9EE4B8BBE6B581
adding httpwww.ommoo.comfeedback.php
adding httpwww.ommoo.comnote.phpactionindexsid102149
warning: no content in part of httpwww.ommoo.comnote.phpactionindexsid102149
warning: no content in part of httpwww.ommoo.comnote.phpactionindexsid102149
adding httpwww.ommoo.comnote.phpactionindexsid148311
warning: no content in part of httpwww.ommoo.comnote.phpactionindexsid148311
warning: no content in part of httpwww.ommoo.comnote.phpactionindexsid148311
adding httpwww.ommoo.comnote.phpactionindexsid763
warning: no content in part of httpwww.ommoo.comnote.phpactionindexsid763
warning: no content in part of httpwww.ommoo.comnote.phpactionindexsid763
adding httpwww.ommoo.comprivacy.php
adding httpwww.ommoo.comu.phpactiontalkuid49421
adding httpwww.ommoo.comu.phpactiontalkuid71719
```

Figure 15: the outcomes of the indexing part



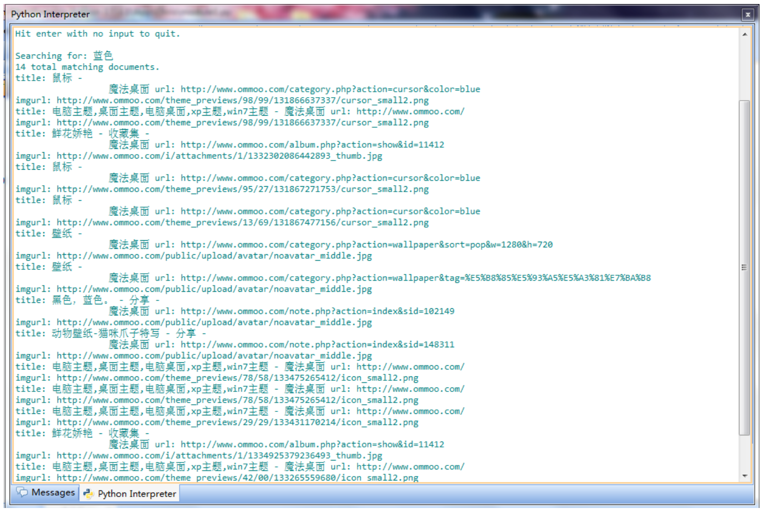
```
Hit enter with no input to quit.

Searching for: 美女
42 total matching documents.
title: 壁纸 -
                魔法桌面 url: http://www.ommoo.com/category.php?action=wallpaper&tag=%E7%BE%8E%E5%A5%B3
imgurl: http://www.ommoo.com/i/attachments/1/1314966567844284_thumb.jpg
title: 壁纸 -
                魔法桌面 url: http://www.ommoo.com/category.php?action=wallpaper&tag=%E7%BE%8E%E5%A5%B3
imgurl: http://www.ommoo.com/i/attachments/1/1317201387340265_thumb.jpg
title: 壁纸 -
                魔法桌面 url: http://www.ommoo.com/category.php?action=wallpaper&tag=%E7%BE%8E%E5%A5%B3
imgurl: http://www.ommoo.com/i/attachments/1/1334897097443822_thumb.jpg
title: 壁纸 -
                魔法桌面 url: http://www.ommoo.com/category.php?action=wallpaper&tag=%E7%BE%8E%E5%A5%B3
imgurl: http://www.ommoo.com/i/attachments/1/1334898731789917_thumb.jpg
title: 壁纸 -
                魔法桌面 url: http://www.ommoo.com/category.php?action=wallpaper&tag=%E7%BE%8E%E5%A5%B3
imgurl: http://www.ommoo.com/i/attachments/1/1339220815639253_thumb.jpg
title: 壁纸 -
                魔法桌面 url: http://www.ommoo.com/category.php?action=wallpaper&tag=%E7%BE%8E%E5%A5%B3
imgurl: http://www.ommoo.com/i/attachments/1/1336842417407972_thumb.jpg
title: 电脑主题,桌面主题,电脑桌面,xp主题,win7主题 - 魔法桌面 url: http://www.ommoo.com/
imgurl: http://www.ommoo.com/i/attachments/1/1314979434490021_thumb.jpg
title: 电脑主题,桌面主题,电脑桌面,xp主题,win7主题 - 魔法桌面 url: http://www.ommoo.com/
imgurl: http://www.ommoo.com/i/attachments/1/1314979434490021_thumb.jpg
title: 电脑主题,桌面主题,电脑桌面,xp主题,win7主题 - 魔法桌面 url: http://www.ommoo.com/
imgurl: http://www.ommoo.com/i/attachments/1/1335762543486628_thumb.jpg
title: 电脑主题,桌面主题,电脑桌面,xp主题,win7主题 - 魔法桌面 url: http://www.ommoo.com/
imgurl: http://www.ommoo.com/i/attachments/1/1335240452600537_thumb.jpg
title: 电脑主题,桌面主题,电脑桌面,xp主题,win7主题 - 魔法桌面 url: http://www.ommoo.com/
imgurl: http://www.ommoo.com/i/attachments/1/1335024579269759_thumb.jpg
title: 电脑主题,桌面主题,电脑桌面,xp主题,win7主题 - 魔法桌面 url: http://www.ommoo.com/
imgurl: http://www.ommoo.com/i/attachments/1/1335021334466111_thumb.jpg
title: 主题皮肤 - 收藏集 -
                魔法桌面 url: http://www.ommoo.com/album.php?action=show&id=13410
imgurl: http://www.ommoo.com/i/attachments/1/1332303102534517_thumb.jpg
```

Figure 16: the outcomes of the searching part(1)



```
>>>
lucene 3.6.1

Hit enter with no input to quit.

Searching for: 狗
5 total matching documents.
title: XP 主题 -
                魔法桌面 url: http://www.ommoo.com/category.php?action=xptheme&color=purple
imgurl: http://www.ommoo.com/i/attachments/1/1331628200436697_thumb.jpg
title: 壁纸 -
                魔法桌面 url: http://www.ommoo.com/category.php?action=wallpaper&tag=%E6%B8%A9%E9%A6%A8
imgurl: http://www.ommoo.com/i/attachments/1/1314065064743750_thumb.jpg
title: 电脑主题,桌面主题,电脑桌面,xp主题,win7主题 - 魔法桌面 url: http://www.ommoo.com/
imgurl: http://www.ommoo.com/i/attachments/1/1315038993508922_thumb.jpg
title: 壁纸 -
                魔法桌面 url: http://www.ommoo.com/category.php?action=wallpaper&tag=%E6%B8%A9%E9%A6%A8
imgurl: http://www.ommoo.com/i/attachments/1/1314065155816874_thumb.jpg
title: 壁纸 -
                魔法桌面 url: http://www.ommoo.com/category.php?action=wallpaper&tag=%E6%B8%A9%E9%A6%A8
imgurl: http://www.ommoo.com/i/attachments/1/1314065123389937_thumb.jpg

Hit enter with no input to quit.
>>>
```

Figure 17: the outcomes of the searching part(2)



```
Hit enter with no input to quit.

Searching for: 蓝色
14 total matching documents.
title: 鼠标 -
            魔法桌面 url: http://www.ommoo.com/category.php?action=cursor&color=blue
imgurl: http://www.ommoo.com/theme_previews/98/99/131866637337/cursor_small2.png
title: 电脑主题,桌面主题,电脑桌面,xp主题,win7主题 - 魔法桌面 url: http://www.ommoo.com/
imgurl: http://www.ommoo.com/theme_previews/98/99/131866637337/cursor_small2.png
title: 鲜花娇艳 - 收藏集 -
            魔法桌面 url: http://www.ommoo.com/album.php?action=show&id=11412
imgurl: http://www.ommoo.com/i/attachments/1/1332302086442893_thumb.jpg
title: 鼠标 -
            魔法桌面 url: http://www.ommoo.com/category.php?action=cursor&color=blue
imgurl: http://www.ommoo.com/theme_previews/95/27/131867271753/cursor_small2.png
title: 鼠标 -
            魔法桌面 url: http://www.ommoo.com/category.php?action=cursor&color=blue
imgurl: http://www.ommoo.com/theme_previews/13/69/131867477156/cursor_small2.png
title: 壁纸 -
            魔法桌面 url: http://www.ommoo.com/category.php?action=wallpaper&sort=pop&w=1280&h=720
imgurl: http://www.ommoo.com/public/upload/avatar/noavatar_middle.jpg
title: 壁纸 -
            魔法桌面 url: http://www.ommoo.com/category.php?action=wallpaper&tag=%E5%B8%85%E5%93%A5%E5%A3%81%E7%BA%B8
imgurl: http://www.ommoo.com/public/upload/avatar/noavatar_middle.jpg
title: 黑色，蓝色。 - 分享 -
            魔法桌面 url: http://www.ommoo.com/note.php?action=index&sid=102149
imgurl: http://www.ommoo.com/public/upload/avatar/noavatar_middle.jpg
title: 动物壁纸-猫咪爪子特写 - 分享 -
            魔法桌面 url: http://www.ommoo.com/note.php?action=index&sid=148311
imgurl: http://www.ommoo.com/public/upload/avatar/noavatar_middle.jpg
title: 电脑主题,桌面主题,电脑桌面,xp主题,win7主题 - 魔法桌面 url: http://www.ommoo.com/
imgurl: http://www.ommoo.com/theme_previews/78/58/133475265412/icon_small2.png
title: 电脑主题,桌面主题,电脑桌面,xp主题,win7主题 - 魔法桌面 url: http://www.ommoo.com/
imgurl: http://www.ommoo.com/theme_previews/78/58/133475265412/icon_small2.png
title: 电脑主题,桌面主题,电脑桌面,xp主题,win7主题 - 魔法桌面 url: http://www.ommoo.com/
imgurl: http://www.ommoo.com/theme_previews/29/29/133431170214/icon_small2.png
title: 鲜花娇艳 - 收藏集 -
            魔法桌面 url: http://www.ommoo.com/album.php?action=show&id=11412
imgurl: http://www.ommoo.com/i/attachments/1/1334925379236493_thumb.jpg
title: 电脑主题,桌面主题,电脑桌面,xp主题,win7主题 - 魔法桌面 url: http://www.ommoo.com/
imgurl: http://www.ommoo.com/theme_previews/42/00/133265559680/icon_small2.png
```

Figure 18: the outcomes of the searching part(3)

each other. And in this experiment, I choose ”www.hao123.com” as the seed web site since I think it can link to plenty of web pages.

As for the second part, I think it is the most complex part in my experiment. I this part, I have to manage to save the name, path, url, title and content of each document in F:\html, except index.txt. And since the method of getting path ,name and the content of the document has been given in the demo, so I only need to get the url and title. And in order to get the url quickly and correctly, I will use the ’index.txt’ I set up in step 1. And in this part, I use the following codes to change it into a dictionary.

```
dic= open('F:\\html\index.txt')
d = dic.readlines()
dic.close()
mydict = {}
for word in d:
key = word.split(';')[0]
value = word.split(';')[1]
mydict[key] = value
```

And in this way we make a dictionary ’mydict’, making it easy to find the corresponding url of each page. Then I use BeautifulSoup to find the title using this code:

```
title=str(soup.head.title.string).decode('utf8')
```

It seems that this part is finished, but actually, it is far from saying so now. If we read the third part carefully, we will find that we have to separate the

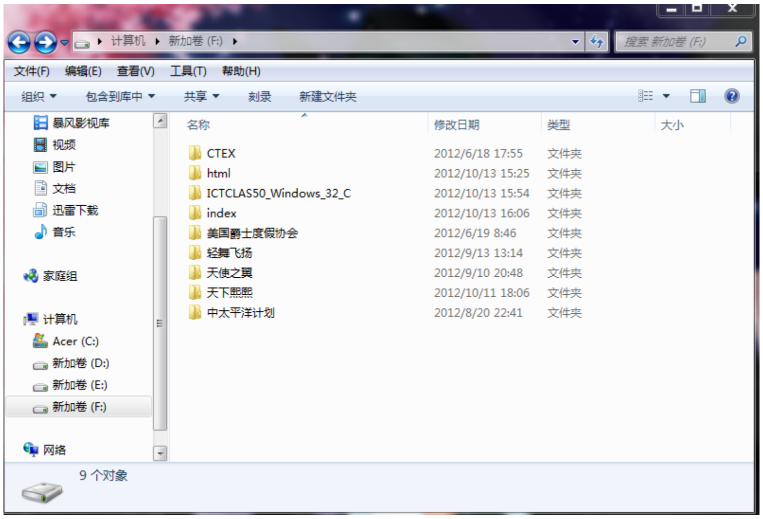
Chinese words using some dictionaries ad then pass them to the analyzer in lucene. Without this ,we will find the outcome of our search will be a mass. So I decide to use ICTCLAS50 to separate the Chinese words with these codes:

```
from ctypes import *
dll=cdll.LoadLibrary("F:\\ICTCLAS50_Windows_32_C\ICTCLAS50.dll")
dll.ICTCLAS_Init(c_char_p("F:\\ICTCLAS50_Windows_32_C"))
strlen = len(c_char_p(contents).value)
t =c_buffer(strlen*6)
bSuccess = dll.ICTCLAS_ParagraphProcess(c_char_p(contents),c_int(strlen),t,c_int
contents=t.value.decode('gbk').encode('utf8')
##list=t.value.split()
##print ' '.join(list)
dll.ICTCLAS_Exit()
```

When I finish the separating work, I find I couldn't find any result in the third part. Then I realized that it is because I didn't change the contents splited by the dictionary to utf8 code, since ICTCLAS50 can support gbk code, which means maybe my contents are just in gbk form. And then when I changed the code, the result is ok. And I need to say, in order to make the content easily to be checked later, I use SimpleAnalyzer to separate in the following part.
In the third part, it is quite easy since the most difficult parts have been solved earlier. And what I need to do is to change the StandardAnalyzer in the demo to SimpleAnalyzer, and make sure the outcome contains path, title, url and name.
Here is a screenshot of where I put those files.
In the second half, I happened to find a method that can get the domain name of the website quickly on the Internet. It is attained by using the urllib library. And the codes are as follows:

```
soup=BeautifulSoup(contents)
url=mydict[filename]
proto, rest = urllib.splittype(url)
site, rest = urllib.splithost(rest)
```

And in this way we can get 'site' as the domain name.
And as for the image search, I decided to use 'www.taobao.com' at first, but later I found there is some problem with getting the url of the pictures. In this condition, some pictures have the url form as 'data-ks-lazyload' instead of the normal 'src' form. I looked it up and found it was because the so-called 'delay loading of the picture'. And as time presses, I hardly have any time to study it deeply to solve the problem, so I have to change the target website. Maybe I will try to solve it later.



名称	修改日期	类型	大小
CTEX	2012/6/18 17:55	文件夹	
html	2012/10/13 15:25	文件夹	
ICTCLAS50_Windows_32_C	2012/10/13 15:54	文件夹	
index	2012/10/13 16:06	文件夹	
美国爵士度假协会	2012/6/19 8:46	文件夹	
轻舞飞扬	2012/9/13 13:14	文件夹	
天使之翼	2012/9/10 20:48	文件夹	
天下熙熙	2012/10/11 18:06	文件夹	
中太平洋计划	2012/8/20 22:41	文件夹	

Figure 19: the files we need(in English)

4 Some of My Thoughts

From lab1 to lab3, it's not hard to find that we are nearer to the search engine step by step. We learn how to get urls in lab1, how to crawl the pages in lab2 and how to search according to the query after making an index in this experiment. So I really think it's fun to study in this field. Though I have some tough time learning things I never heard of and spend tons of time programming. But when I see the achievement from my own hands, I feel real happiness that can't be expressed with words.

And as for the second half, we know that it can be more exactly when we try to search some information on a certain website. And I do think it is of great use because most of the times, we really want to search something on one page, and our experiment enables us to achieve it.

In the experiment, I have ever forgot to save the successful code and lost them, taking me some time to write them again. But luckily, in this process, I think I am better at commanding the principle of the searching method. But I think the most annoying thing is that python is really slow when running. If it can be a little faster, I think it will be a more relaxed thing searching information using the program written of our own.