python爬虫实验报告

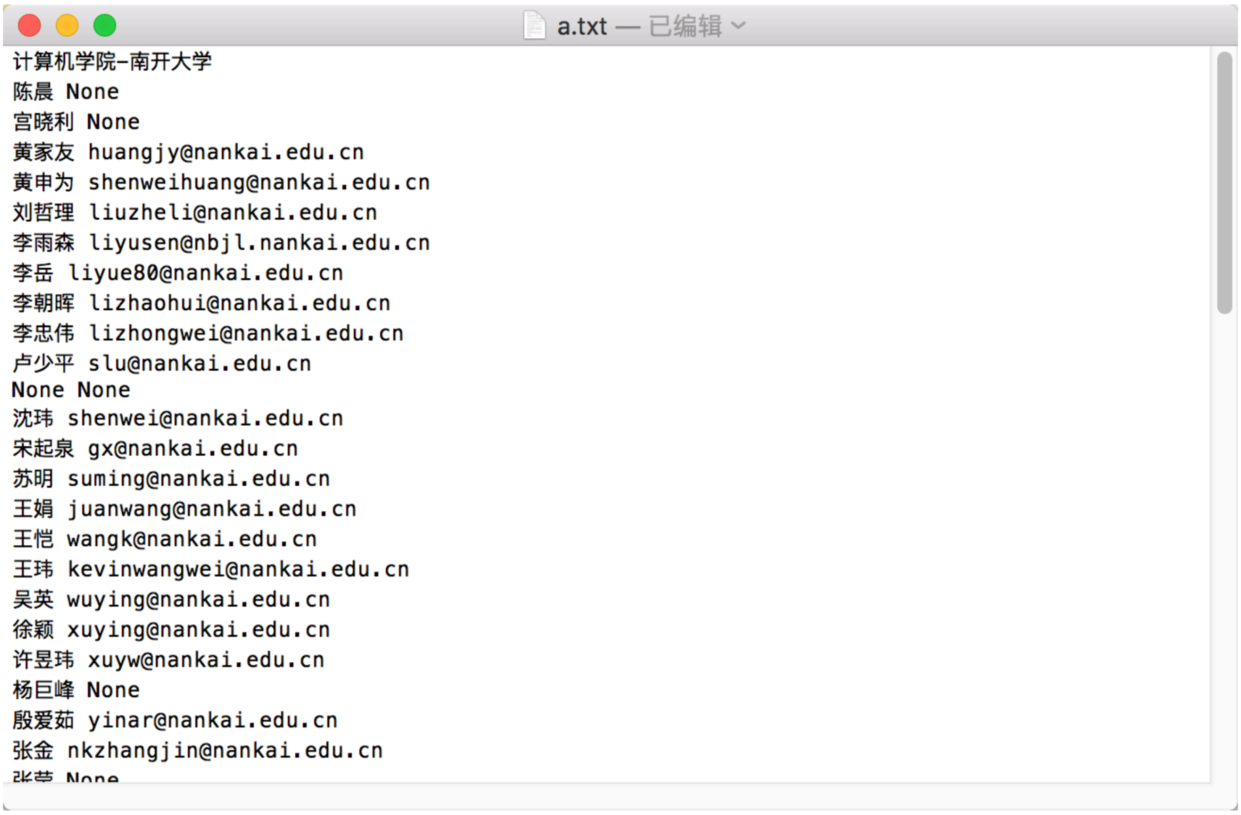
1611736 钟腾

本次试验要求爬南开大学所有老师的邮箱并发送邮件，一开始找到起始网址'http://www.nankai.edu.cn/213/list.htm'，里面有各个学院的网站，各学院网站有老师信息网页，从起始网站开始搜索nankai后缀的网站，然后开始匹配邮箱的正则表达式，如果有就保存在新建文件中。再从文件中读出所有老师名字和邮箱给他们发送邮件。

源代码如下：

```python
import re
import urllib.request
import urllib
from collections import deque
import smtplib
from email.mime.text import MIMEText
from email.header import Header

# 提取文章标题的正则表达式
REX_TITLE=r'<title>(.*?)</title>'
# 提取所需链接的正则表达式
CC_REX_URL=r'/introduce/[\w]*'
# 种子url，从这个url开始爬取
BASE_URL='http://www.nankai.edu.cn/213/list.htm'

init_queue = deque()
init_visited = set()
init_queue.append(BASE_URL)


def ccFindTeacher(CURRENT_URL): #计控学院
    #   爬虫用到的两个数据结构，队列和集合
    queue = deque()
    visited = set()
    queue.append(CURRENT_URL)
    count = 0
    email_list = []
    name_list = []

    while queue:
        url = queue.popleft()  # 队首元素出队
        if count == 0 :
            list_CURRENT_URL = CURRENT_URL.split('/')
            list_CURRENT_URL = list_CURRENT_URL[0:4]
            print(list_CURRENT_URL)
            CURRENT_URL = ('/').join(list_CURRENT_URL)
            print(CURRENT_URL)
        visited |= {url}  # 标记为已访问

        print('已经抓取: ' + str(count) + '   正在抓取 <---  ' + url)
        count += 1
        urlop = urllib.request.urlopen(url)

        # 避免程序异常中止
        try:
            data = urlop.read().decode('utf-8')
        except:
            continue
        if count != 1:
            email=re.search(r'[\w-]+(\.[\w-]+)*@[\w-]+(\.[\w-]+)+',data)
            if email:
                email_list.append(email.group(0))
            else:
                email_list.append("None")
            name=re.search(r'<span class="attribute">[\s]*姓    名[\s]*:[\s]*</span>[\s]*<span>[\s]*[\w]*[\s]*</span>',data)
            if name:
                list_name = name.group(0).split(' ')
                print(list_name[67])
                name_list.append(list_name[67])
            else:
                name_list.append("None")

        # 正则表达式提取页面中所有链接，并判断是否已经访问过，然后加入待爬队列
        linkre = re.compile(CC_REX_URL)
        for sub_link in linkre.findall(data):
            sub_url=CURRENT_URL+sub_link
            if sub_url in visited:
                pass
            else:
                # 设置已访问
                visited |= {sub_url}
                # 加入队列
                queue.append(sub_url)
                print('加入队列 --->  ' + sub_url)

    for i in range(0,len(email_list)):
        # if email_list[i] != 'None':
        #     sender = '754159742@qq.com'
        #     # mail_pass = ''
        #     receivers = [email_list[i]]
        #     message = MIMEText('老师好，Python 作业邮件发送，请勿理会。','plain','utf-8')
        #     message['From'] = Header('作业','utf-8')
        #     message['To'] = Header('测试','utf-8')
        #     subject = 'Python 作业邮件'
        #     message['Subject'] = Header(subject,'utf-8')
        #     try:
        #         smtpObj = smtplib.SMTP('localhost')
        #         smtpObj.sendmail(sender,receivers,message.as_string())
        #         print('邮件发送成功')
        #     except smtplib.SMTPException:
        #         print("ERROR: 无法发送邮件")
        # with open('/Users/zhongteng/Desktop/a.txt','a') as f:
                # f.write(name_list[i]+" "+email_list[i]+'\n')
            print(name_list[i]+""+email_list[i]+'\n')

while init_queue:
    current_url = init_queue.popleft()
    init_visited |= {current_url}

    print('   正在抓取 <---  ' + current_url)
    current_urlop = urllib.request.urlopen(current_url)

        # 避免程序异常中止
    try:
        first_data = current_urlop.read().decode('utf-8')
    except:
        continue
    college = re.findall(r'http://[\w]*.nankai.edu.cn',first_data)

    if 'http://cc.nankai.edu.cn' in college:
        page_count = 1
        init_page = 'http://cc.nankai.edu.cn/teachers/search/'
        page = init_page+str(page_count)
        urlop = urllib.request.urlopen(page)
        data = urlop.read().decode('utf-8')
        linkre = re.search(REX_TITLE,data)
        if(linkre):
            print('获取文章标题：'+linkre.group(1))
            with open('/Users/zhongteng/Desktop/a.txt','a') as f:
                f.write(linkre.group(1)+'\n')
        try:
            while data:
                page_count += 1
                page = init_page+str(page_count)
                urlop = urllib.request.urlopen(page)
                data = urlop.read().decode('utf-8')
                ccFindTeacher(page)
        except:
            pass
```

爬出结果示例如下：



```
计算机学院-南开大学
陈晨 None
宫晓利 None
黄家友 huangjy@nankai.edu.cn
黄申为 shenweihuang@nankai.edu.cn
刘哲理 liuzheli@nankai.edu.cn
李雨森 liyusen@nbjl.nankai.edu.cn
李岳 liyue80@nankai.edu.cn
李朝晖 lizhaohui@nankai.edu.cn
李忠伟 lizhongwei@nankai.edu.cn
卢少平 slu@nankai.edu.cn
None None
沈玮 shenwei@nankai.edu.cn
宋起泉 gx@nankai.edu.cn
苏明 suming@nankai.edu.cn
王娟 juanwang@nankai.edu.cn
王恺 wangk@nankai.edu.cn
王玮 kevinwangwei@nankai.edu.cn
吴英 wuying@nankai.edu.cn
徐颖 xuying@nankai.edu.cn
许昱玮 xuyw@nankai.edu.cn
杨巨峰 None
殷爱茹 yinar@nankai.edu.cn
张金 nkzhangjin@nankai.edu.cn
张莹 None
```

然后利用smtp协议发送邮件即可。