

李墨珩 2017050025

程序设计训练第三周大作业说明文档

September 15, 2018

程序设计训练第三周大作业

新闻搜索引擎设计

综合运用了 Python3 Jieba BeautifulSoup4 正则表达式 Django HTML CSS Javascript Mysql Bootstrap4 搭建了新闻搜索引擎。

新闻搜索引擎设计

程序设计夏季学期第三周大作业

我们最近收集了很多新闻，您想看些什么？

请输入关键字 多个关键字间用空格隔开

(可选) 搜索中包含的最早新闻时间 输入格式为 20150308

(可选) 搜索中包含的最晚新闻时间 输入格式为 20181018

开始搜索

或者你想看看我们都收集了什么？

近几天发生了这些事情：

**企业舆情：中国环企50强仅7家
过百亿 亟须“输血”**

20180914
据中国人民大学绿色金融团队测算，2015年-2020年，国内的绿色金融需求为15万-30万亿元，空间巨大。其中环保、新能源、环境基础设施、环境修复、工业污染治理、能源与资源节约等五大领域，绿色金融需求将达到14.6万亿元。

**每日监测(13日)：取消“清考”大学
教育不再“放水”**

20180913
近日，教育部印发《关于推进新时代全国高等学校本科教育工作会议精神落实的通知》，对加强本科教育首次“加严”，要求严格本科教育教学过程管理，淘汰“水课”，加大过程考核成绩在课程总成绩中的比重，严把毕业出口关，坚决取消“清考”制度。

**幼儿园问卷调查家庭住房情况，
不妥在哪**

20180914
家庭住房情况是重要的个人隐私，幼儿园问卷调查包含这样的内容，涉嫌侵犯家长隐私，引发争议在所难免。对此，有家长称“这是描述我的小区啊，明明是描述我的经济状况”，反映了家长的忧虑和呼声。

**雷军启动小米首次组织结构变
革，把一线阵地交给80后**

20180913
这是小米上市之后的首次重大调整，也是小米成立以来最大的组织架构变革。

一、数据收集

分为三部分：获取URL、获取全文以及关键字处理。

第一部分：

从新华网的不同板块出发，首先采集新闻链接 url、新闻标题 title 和新闻摘要 summary，存入 Mysql 数据库。以自建 id 为关键字（以采集次序自增），在向数据库中存入数据时对 url 查重，确保 url 唯一。

并对在采集过程中可能出现的错误进行了预处理，避免运行时报错，降低效率。

```
7 db_config={
8     'host': '127.0.0.1',
9     'port': 3306,
10    'user': 'root',
11    'password': 'passMY.18',
12    'db': 'pytest',
13    'charset': 'utf8'
14 }
15 connection = pymysql.connect(**db_config)
16 oldurls=['http://www.xinhuanet.com/world/index.htm', 'http://www.xinhuanet.com/overseas/index.htm',
17 'http://www.xinhuanet.com/local/index.htm', 'http://www.xinhuanet.com/house/index.htm',
18 'http://www.xinhuanet.com/politics/rs.htm', 'http://www.xinhuanet.com/tech/index.htm',
19 'http://www.xinhuanet.com/politics/leaders/index.htm', 'http://www.xinhuanet.com/politics/xhll.htm',
20 'http://www.xinhuanet.com/food/index.htm', 'http://www.xinhuanet.com/auto/index.htm',
21 'http://www.xinhuanet.com/info/index.htm', 'http://www.xinhuanet.com/health/',
22 'http://www.xinhuanet.com/travel/', 'http://www.news.cn/money/index.htm', 'http://www.xinhuanet.com/
23 'http://www.news.cn/comments/index.htm', 'http://www.news.cn/fashion/index.htm',
24 'http://www.xinhuanet.com/legal/ffu.htm', 'http://www.news.cn/fortune/caiyan.htm', 'http://www.news.cn/
25 'http://www.news.cn/abroad/index.htm', 'http://www.news.cn/city/index.htm', 'http://www.news.cn/zl
26 'http://www.news.cn/info/spsy/index.htm', 'http://www.news.cn/xhsd/index.htm'],
27 urls=['http://www.news.cn/politics/xgc.htm', 'http://www.news.cn/finance/', 'http://chanye.news.cn/
28 headers = {'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Ge
29 try:
30     with connection.cursor() as cursor:
31         sql = 'insert into newslinks(title, url, summary) values(%s, %s, %s)'
32         for url in urls:
33             page = request.Request(url, headers=headers)
34             page_info = request.urlopen(page).read().decode('utf-8')
35             soup = BeautifulSoup(page_info, 'html.parser')
36             allis = soup.find_all(name='li', attrs={'class': 'clearfix'})
37             for li in allis:
38                 if li.find(name='a') is not None:
39                     title = li.find(name='a').string
40                     if li.find(name='a').attrs['href'] is None:
41                         url='None'
42                     else:
43                         url = li.find(name='a').attrs['href']
44                     if li.find(name='p', attrs={'class': 'summary'}) is None:
45                         summary = 'None'
46                     else:
47                         summary = li.find(name='p', attrs={'class': 'summary'}).string
48                     cursor.execute(sql, (title, url, summary))
49                     connection.commit()
50                     print('finished a website!')
51 finally:
52     connection.close()
53 print('findlinks.py all finished!')
```

第二部分：

从数据库中收集的 url 出发，利用 BeautifulSoup4 和 正则表达式 逐条寻找每篇新闻的发布时间 date 和新闻全文 text，并对在采集过程中可能出现的错误进行了预处理，避免运行时报错，降低效率。

```

27     cursor.execute(sqlr)
28     results=cursor.fetchall()
29     for row in results:
30         url = row[3]
31         page = request.Request(url, headers=headers)
32         try:
33             page_info = request.urlopen(page).read().decode('utf-8')
34         except urllib.error.HTTPError as e:
35             print('HTTPError',row[0])
36             continue
37         except UnicodeDecodeError as e:
38             print('UnicodeDecodeError')
39             continue
40
41         soup = BeautifulSoup(page_info, 'html.parser')
42
43         text=''
44         div = soup.find(name='div',attrs={'id':'p-detail'})
45         if div is not None:
46             #print(div)
47             for p in div.find_all(name='p'):
48                 if p.string is not None:
49                     text+=p.string
50                     text+='\n'
51             #print(text)
52
53
54         date=0
55         span=soup.find(name='span',attrs={'class':'h-time'})
56         if span is not None and span.string is not None:
57             datat=span.string
58             date=int(datat[1:5])*10000+int(datat[6:8])*100+int(datat[9:11])
59         else:
60             date=99999999
61         #print(date)
62
63         sqlu = 'UPDATE newslinks SET date = %s WHERE (id = %s)'
64         cursor.execute(sqlu, (date,row[0]))
65         connection.commit()
66
67         sqlu = 'UPDATE newslinks SET text = %s WHERE (id = %s)'
68         cursor.execute(sqlu, (text,row[0]))
69         connection.commit()
70
71         print('finished one row!'+str(row[0]))
72         time.sleep(0.7)
73

```

第三部分：

对数据库中每条新闻的摘要和全文做分词处理（搜索引擎模式），整理并统计词频，并将重要的关键字写入数据库。整理关键字时，为确保搜索效率，做出了如下规定：

- 1:关键字长度大于1；
- 2:关键字词频大于3；
- 3:以词频为主要权重，倒序排列，最多取10个关键字写入数据库。

对得到的关键字建立倒排列表，便于后续索引。对在采集过程中可能出现的错误进行了预处理，避免运行时报错，降低效率。

```
connection = pymysql.connect(**db_config)
cursor = connection.cursor()
sqlr = "select * from newslinks where id>2000 "
cursor.execute(sqlr)
results=cursor.fetchall()

for row in results:
    try:
        ans=''
        wordcnt=0
        wordlist=[]
        worddict={}
        text=''
        if row[4] is not None:
            text+= row[4]
        text+= row[5]
        if text!='':
            wordlist=jieba.cut_for_search(text)
            for item in wordlist:
                if len(str(item))>1:
                    if item not in worddict:
                        worddict[item]=1
                    else:
                        worddict[item]+=1
            orderList=list(worddict.values())
            orderList.sort(reverse=True)
            for i in range(len(orderList)):
                for key in worddict:
                    if worddict[key]==orderList[i] and worddict[key]>3 and wordcnt<11:
                        ans+=(key+str(worddict[key])+',')
                        wordcnt+=1
                        worddict[key]=0

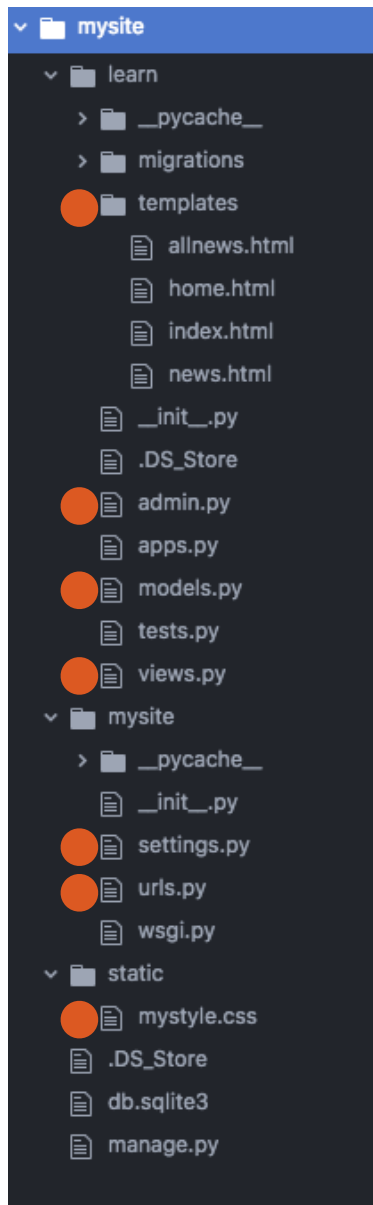
            sqlu = 'UPDATE newslinks SET keywords = %s WHERE (id = %s)'
            cursor.execute(sqlu, (ans,row[0]))
            connection.commit()
            print('finished one!'+str(row[0]))
        else:
            print('text is empty!'+str(row[0])+ '!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!')
            continue
    except Exception as e:
        raise e
        print('something is wrong with'+str(row[0])+ '!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!')
        continue

print('all finished!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!')
```

二、网站搭建

首先建立 project 命名为 mysite ，在建立 app 命名为 learn 。

文件目录如图：



主要操作的是红色标注的几个文件：

templates 只要存储我使用的网页模版，四个模版由上至下别是：整合所有新闻并展示的页面、搜索首页、搜索结果显示页面、新闻详情页面。

admin.py 是我后台整理数据的操作文件。

models.py 是使用 MVC 架构时与数据库接轨的一个操作文件。

views.py 是使用 MVC 架构时处理model传来的数据，并将数据发送到html模版中的操作文件。

settings.py 是管理项目设置的文件。

urls.py 是处理不同url与view.py不同函数对接关系的操作文件。

mystyle.css 是CSS文件，为了设计并美化html模版而用。

下面分页面进行说明：

搜索首页：

新闻搜索引擎设计

程序设计夏季学期第三周大作业

我们最近收集了很多新闻，您想看些什么？

或者你想看看我们都收集了什么？

近些天发生了这些事情：

企业舆情：中国环企50强仅7家过百亿 亟须“输血”
20180914

据中国人民大学绿色金融团队测算，2016年-2020年，国内绿色金融需求为15万-30万亿元，空间巨大，其中可持续发展、环境基础设施、环境修复、工业污染治理、能源与能源节约等五大领域，绿色融资需求将达到14.6万亿元。

每日监测(13日)：取消“清考”大学教育不再“放水”
20180913

近日，教育部印发《关于借鉴新时代全国高等学校本科教育工作会议精神落实的通知》，对加强本科教育两次“加码”，要求严格本科教育教学过程管理，淘汰“水课”，加大过程考核成绩在课程总成绩中的比重，严把毕业出口关，坚决取消“清考”制度。

幼儿园问卷调查家庭住房情况，不妥在哪
20180914

家庭住房情况是重要的个人隐私，幼儿园问卷调查包含这样的内容，涉嫌侵犯家长隐私，引发争议在所难免。对此，有家长称“这要是描述我的小区啊，明明是描述我的经济状况”，反映了家长的忧虑和呼声。

雷军启动小米首次组织结构变革，把一线阵地交给80后
20180913

这是小米上市之后的首次重大调整，也是小米成立以来最大的组织架构变革。

华硕发布百款ROG系列新品
20180913

发布了包括PC游戏、主机游戏和手机游戏在内的三大游戏高端解决方案。

盘点人工智能的行业应用
20180913

人工智能作为新一轮产业变革的核心驱动力，必须与各行各业融合才能发挥作用，形成真正有效的行业智能。以此来助力传统行业转型升级，加快人工智能应用落地。

5G开启超高清视频时代
20180913

超高清视频是继数字化、高清化之后的新一轮重大技术革新，未来5年将是我国超高清视频产业发展的战略机遇期。

构建了一个POST表单，里面包含三个搜索栏，一个提交按钮。

关键字搜索必需项，支持多关键词搜索，两个时间搜索为可选项。点击提交后，会自动判断有无时间限定。

多个关键词输入后，我们会将关键词分开，对每个关键词设立权值，单独搜索，最后整合搜索结果，若有时间限定，我们再过滤掉不合法的时间，按权值大小展示结果，得到最优的答案。

页面中“或者你想看看我们都收集了什么？”可以链接到“所有新闻页面”的第一页。

“近些天发生了这些事情：”会收集最近日期的热点新闻，并展示出来。

搜索结果展示页面：

我们对 "北京" 的检索结果如下：

本次检索耗时为0.003731250762939453秒，为您找到221条相关新闻。

我们对得到的新闻进行了分页展示，共有 23 页，当前您所浏览的是第 1 页。[点击此处带您回到搜索首页。](#)

新房质量验收 呼唤独立验房机制

20180912

9月10日，记者从**北京市**住建委获悉，近日，**北京市**住建委发布了《关于对新建住宅交付使用前实施房屋质量查验的通知（征求意见稿）》。

北京楼市量价平稳 多家互联网中介违规被点名

20180911

互联网平台必须做到房源信息、经纪人信息真实并实时更新。对此，**北京市**住建委要求各网站对检查发现的问题立即整改，严把房源发布准入关，下架不合格房源信息。

二手房成交量持续下滑 楼市还会有金九银十吗？

20180906

金九银十，楼市传统意义上的旺季，而今年的九月份，**北京**楼市是否会延续往年的惯例，走出一波新行情？对此，无论是新房开发商，还是二手房中介，都显得有些信心不足。

北京3000多套共有产权房可申购了！位置好价格低

20180906

据**北京市**住建委官网消息，位于石景山区、长安街以北的玉景阳光共有产权房6日开始网申，均价3.5万元/平米，购房人拥有50%的产权。链家数据显示，玉景阳光周边的二手商品房价格在6万-7万元/平米。

得到关键字之后，在 views.py 中执行相应函数，在数据库中进行检索，得到数据后，将数据传递给 index.html 模版，进行展示。

检索耗时：计算了函数开始检索到检索完毕的耗时；

分页显示：（后端分页模式）在传递新闻条目时，并不是全部传递到模版。每页只显示10条结果，每次回传递当前页页码和所对应的十条结果，当点击下一页的时候，采用GET方法获取想要显示的页码，在对应好数据，传递给模版。

关键字高亮：采用前端 Javascript 方法，对带有特定id的板块进行关键字正则匹配，对其添加额外的标签。对应代码如下：

```
<script>
function SearchHighlight(idVal, keyword) {
    var pucl = document.getElementById(idVal);
    if ("" == keyword) return;
    var temp = pucl.innerHTML;
    var htmlReg = new RegExp("<.*?>", "i");
    var arrA = new Array();
    for (var i = 0; true; i++) {
        var m = htmlReg.exec(temp);
        if (m) {
            arrA[i] = m;
        } else {
            break;
        }
        temp = temp.replace(m, "{[" + i + "]}");
    }
    words = unescape(keyword.replace(/+/g, ' ')).split(/\s+/);
    for (w = 0; w < words.length; w++) {
        var r = new RegExp("(" + words[w].replace(/[\(){}.\+*?^$|\\[\]]/g, "\\$&") + ")", "ig");
        temp = temp.replace(r, "<b style='color:Red;'>$1</b>");
    }
    for (var i = 0; i < arrA.length; i++) {
        temp = temp.replace("{[" + i + "]", arrA[i]);
    }
    pucl.innerHTML = temp;
}
SearchHighlight("news", "{originalinput}");
</script>
```

所有新闻页面：

这是我们收集到的所有新闻：

我们对得到的 6947 条新闻进行了分页展示，共有 695 页，当前您所浏览的是第 1 页。[点击此处带您回到搜索首页。](#)

新华国际时评：引领中俄关系保持高水平发展

20180911

在两国元首的战略引领下，中俄全面战略协作伙伴关系步入高水平、大发展的新时代。习近平主席和普京总统都曾用“典范”形容当前的中俄关系。

财经观察：龙狮共舞共促繁荣

20180911

在中非合作升温的大背景下，“中国龙”与“非洲狮”共同起舞，踩在经济全球化的时代鼓点上，不仅会给 2 6 亿中非人民带来深远福祉，也将增进中非以及世界经济的发展繁荣。

施压巴解组织和国际刑事法院 美欲“一石三鸟”恐难如愿

20180911

特朗普政府意在迫使巴方接受美国调停的巴以和平方案，阻挠国际刑事法院对美国及其盟国开展战争罪行调查，并在美中期选举前巩固对外强硬捍卫国家利益的形象以争取选民支持，不过这些目的恐难实现。

国际观察：当“北欧福利主义”遭遇极右浪潮

20180910

作为“北欧福利主义”代表的“瑞典模式”遭遇极右浪潮，未来新政府组阁和社会福利制度的改革走向扑朔迷离，也给难民问题带来的“欧洲困境”增添新案例。

外媒：美国反对调查美涉战争罪行

和搜索结果展示页面异曲同工，不在重复叙述。

新闻详情页面：

新闻详情页

[点击此处带您回到搜索首页。](#)

施压巴解组织和国际刑事法院 美欲"一石三鸟"恐难如愿

20180911

[查看此新闻出处](#)

新华社记者朱东阳 刘晨

美国政府10日宣布关闭巴勒斯坦解放组织（巴解组织）驻华盛顿办事处，并威胁制裁国际刑事法院的法官和检察官。

分析人士认为，特朗普政府此举意在迫使巴方接受美国调停的巴以和平方案，阻挠国际刑事法院对美国及其盟国开展战争罪行调查，并在美中期选举前巩固对外强硬捍卫国家利益的形象以争取选民支持，不过这些目的很可能难以实现。

强化对巴施压

近来美国与巴勒斯坦矛盾日益明显。据报道，美国政府目前正在酝酿一项所谓“世纪协议”设想，以推动巴以双方实现最终和平。但这一协议尚未出台就已遭到巴方明确反对。为此，美国已多次对巴施压。就在8月31日，美国刚刚宣布不再向联合国近东巴勒斯坦难民救济和工程处提供资金。

美国智库布鲁金斯学会高级研究员达雷尔·韦斯特指出，美国在巴以矛盾中“拉偏架”的迹象非常明显，巴方不再视美国为中东和平谈判的中立调停方，认为对话将完全偏向以色列，故而拒绝和以方谈判。美国此次关闭巴解组织驻华盛顿办事处就是要对巴方拒绝参加和谈进行“报复”，试图借此迫使巴方重回谈判。

事实上，美国政府对此也并不讳言。美国国务院发言人诺尔特10日在宣布关闭巴解组织驻华盛顿办事处的声明中称，巴解组织不仅没有采取措施推动巴以开始“直接而有意”的谈判，其领导层还一直谴责美国尚未出台的巴以和平计划，并拒绝与美国政府就相关事宜进行接触。此外，巴方还试图推动国际刑事法院对以色列进行调查。美国政府因此做出上述决定。

阻挠国际调查

就在同一天，国际刑事法院也受到了来自美国的威胁。美国总统国家安全事务助理博尔顿10日发表演讲时称，如果该法院启动对美国在阿富汗等地所涉的战争罪行调查，并因此起诉美国、以色列或者其他盟国，美国将在本国法律许可的范围内对该法院法官和检察官采取反制行动。

分析人士认为，美国政府在“9·11”事件纪念日前夕发布此种威胁，意在减轻对美以在

从目前情况看，其愿望恐怕很难实现。

巴勒斯坦方面10日通过各种渠道发声谴责美国的举动，表示不会在巴勒斯坦建国、耶路撒冷地位和难民回归等原则问题上改变立场。巴解组织驻华盛顿办事处代表胡萨姆·佐姆洛特发表声明说，美国的决定将促使巴方向国际刑事法院施压，以加快对以色列的审判。

关于上述举动的选举效果，美国智库布鲁金斯学会外交政策高级研究员欧汉龙认为，这些举动反映出博尔顿等特朗普政府内部强硬派人士的诉求，但很难说会赢得多少美国选民的心，大多数美国选民对博尔顿的呼声并不关心。

此外，一些美国专家认为上述举动还将对美国自身产生负面影响。布鲁金斯学会高级研究员韦斯特说，政府此举损害了美国的全球形象，再次加深了国际社会对美国“不公平”和“单边主义”的印象，这将导致其他国家进一步质疑美国的全球领导力。（参与记者：赵悦、杨媛媛）

我们还为您准备了一些相关的新闻：

外媒：美国反对调查美涉战争罪行

20180912

外媒称，美国10日对国际刑事法院发起前所未有的严厉抨击，并威胁说如果该法院盯住美国人或者以色列人，就对其法官和检察官进行制裁。

美国“勒令”关闭巴解办事处 巴勒斯坦：这等同于宣战

20180912

美国已向巴解组织发出正式通知，称将关闭其驻华盛顿的办事处。“这等同于向我们国家和地区宣战，尽管他们曾承诺带来和平”，巴民族权力机构发言人说。

美国众议员承诺保住链式移民 吁华人选民积极投票

20180228

据美国《世界日报》报道，包括美国伊利诺伊州国会众议员夏考斯基(Jan Schakowsky)、奎利(Mike Quigley)等十多位民代、政府官员，日前出席华人侨团“新春庆团圆”活动时表示，“一定会致力保存家庭链式移民”，夏考斯基也呼吁合格选民，尽快完成选民登记，才能在期中选举时，达到“以票发...

女美下作第江被禁 中国印地安“美同禁”近乡

利用GET方法得到新闻id之后，在 views.py 中执行相应函数，在数据库中进行检索，得到数据后，将数据传递给 news.html 模版，进行展示。

相关新闻：获取到当前新闻的第一个和第二个关键字，以该关键字再次进行检索，将得到的结果展示出来，只展示最相关的前四个。

具体urls对应关系：

```
urlpatterns = [
    path('', learn_views.showhomepage,name='home'),
    path('admin', admin.site.urls,name='admin'),
    path('news/<int:id>', learn_views.showNewsDetail,name='newsdetail'),
    path('search', learn_views.index,name='index'),
    path('search/<int:wantpagecnt>', learn_views.showwantindex,name='pagechangeto'),
    path('allnews/<int:wantpagecnt>', learn_views.showwantallnews,name='allnewschangeto'),
]
```

三：具体实现

请移步代码，那里有比较详细的注释说明。