

Python 新闻爬虫大作业

姓名：嵇天颖

班级：计 64

学号：2016010308

一、功能实现

(一) 基础功能

(1) 网络爬虫

1. 从清华新闻网上爬取了超过 5000 条新闻；
2. 实现了新闻数据的预处理与索引，抽取了新闻的内容，实现标题和正文分词处理，并为新闻建立了索引；

(2) Web 设计

1. 实现了首页新闻展示功能；
2. 新闻展示首页可以显示全部的新闻数量，并使用表格等方式显示新闻概要信息（标题、时间和正文摘要），并设计为分页显示；
3. 点击新闻展示首页中的新闻概要，可以跳转到新闻详情页，详情页中显示新闻全文；
4. 新闻主页都实现了一个查询表单，包括一个文本框和一个查询按钮；
5. 新闻查询也实现了支持时间查询的查询表单，包括一个文本框和四个时间选择项；
6. 点击查询按钮后跳转到查询结果页面；

7. 使用 css/javascript 等前端语言或前端框架对页面布局进行美化;
8. 搜索结果列表页中包含关键字的标题和正文部分予以高亮显示;

(二) 加分项

1. **高级搜索**: 在保留基础单关键字查询的基础上, 支持下列几种高级搜索功能

(I)关键字与时间范围的协同查询: 支持查询给定时间段中带有关键字的所有新闻;

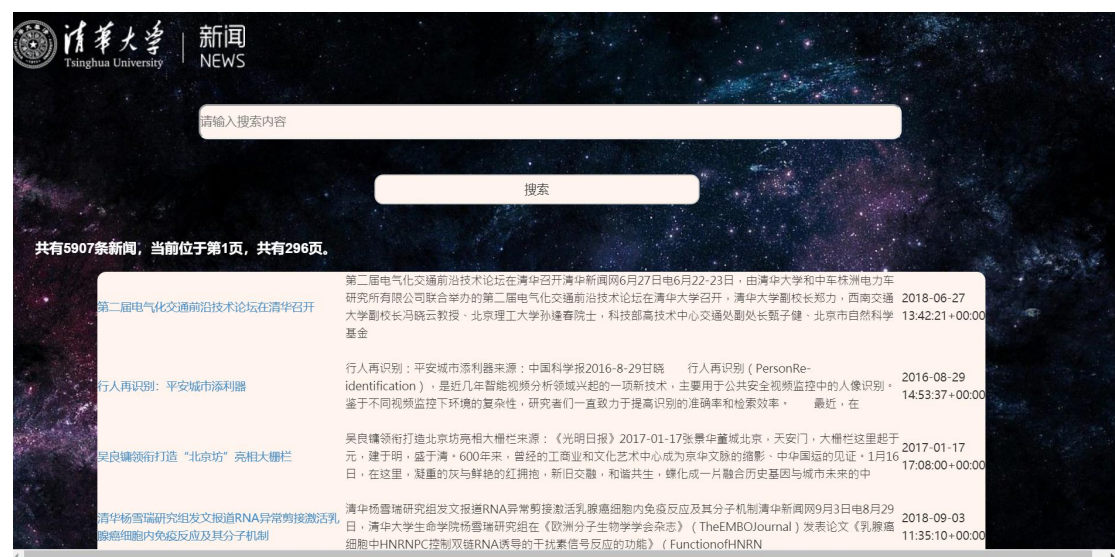
(II)多关键字查询: 支持输入空格分割的多个关键字或一句话, 查询与输入的内容接近的若干条结果;

(III)每次显示出查询耗时;

2. **新闻推荐**: 在每条新闻详情页的最下方, 增加“新闻推荐”部分, 推荐 1-3 篇其他新闻, 点击后可跳转到被推荐新闻的详情页

二、界面与功能展示

- (一) **新闻首页**: 新闻首页支持关键词查询, 并展示了所有新闻, 同时具备分页功能



点击'previous'可以跳转到上一页，点击'next'可以跳转到下一页，中间白色字体部分显示当前所在页面：

清华大学教授李稻葵获批国家社会科学基金重大项目	清华大学教授李稻葵获批国家社会科学基金重大项目 清华大学新闻网11月23日电11月，清华大学经济管理学院弗里曼讲席教授、金融系李稻葵申请的“国家社会科学基金重大项目我国历史上的GDP及其结构研究（980-1840年）”获得批准。 此重大项目是国家社会科学基金2015年度第二批投标项目，并将于年底启动。该	2015-11-23 10:36:43+00:00
清华清华幼儿园举行70周年园庆系列活动	清华清华幼儿园举行70周年园庆系列活动清华大学新闻网6月12日电6月9日，清华清华幼儿园迎来了70周年华诞，清华大学和市区教委各级领导和幼教专家齐聚一堂，共同庆祝这美好的节日。清华大学党委常务副书记、副校长姜胜耀，副校长吉俊民，党委原书记贺美英，校务委员会副主任、党委原副书记韩景阳，国务院妇儿工委办协调	2018-06-12 15:47:55+00:00
2017清华大学招生凸显三个关键词：“大”“新”“更”	2017清华大学招生凸显三个关键词：大新更来源：新华网2017-05-01在清华大学106周年校庆前夕，招办主任刘震邀请人文与社会类首席教授、社科学院党委书记刘海雄和外交系主任、世界文学与文化研究院（简称世文院）院长、世界文学与文化实验班（简称世文班）首席教授颜海平共同接受媒体采访，阐释2017年	2017-05-06 20:08:56+00:00
朝阳区与清华签约开展教育医疗合作	朝阳区与清华签约开展教育医疗合作来源：《北京青年报》2017-10-11刘旭昨天，朝阳区与清华大学签署《朝阳区人民政府清华大学全面合作协议》，双方将在区域建设、教育事业、医疗卫生、人才培养等方面开展系列合作。签约仪式上，清华大学附属垂杨柳医院正式授牌。据悉，未来，清华大学附属垂杨柳医院将充分运用	2017-10-11 17:53:45+00:00
清华电子系博士生获2017国际传输技术大会最佳学生论文奖	清华电子系博士生获2017国际传输技术大会最佳学生论文奖清华大学新闻网10月10日电9月24至27日，由国际电气与电子工程师学会（IEEE）主办的第86届国际传输技术大会（2017IEEE86thVehicularTechnologyConference，IEEEVTC2017-Fall）在加拿大多伦多	2017-10-10 15:38:35+00:00
【组图】清华美院建院60周年校友服装设计作品发布会	清华美院建院60周年系列活动清华美院（中央工艺美术学院）建院60周年校友服装设计作品发布会清华大学新闻网11月7日电11月1日，清华大学美术学院（中央工艺美术学院）建院60周年校友服装设计作品发布会在清华大学艺术博物馆首层大厅举行。本次活动由清华大学美术学院主办，清华大学美术学院染织服装艺术设计系承办。本次	2016-11-07 16:13:52+00:00
清华大学“星火计划”第十期学员赴英国产业调研	清华大学星火计划第十期学员赴英国产业调研清华大学新闻网9月1日电（通讯员：修新羽）8月1日至8月12日，科技创新，星火燎原清华大学学生创新人才培养计划第十期（以下简称星火十期）共39名师生赴英国进行了以探访日不落为主题的产业调研活动，这也是继去年走进日本后，星火计划第二次集体进行的海外产业调研。同学们从	2016-09-01 16:40:43+00:00

首页支持关键词和多关键词查询：点击'搜索'按钮既可以显示在全体新闻中的全部搜索结果：

	清华大学 Tsinghua University	新闻 NEWS
<input type="text" value="清华"/>		
<input type="text" value="清华"/>		
<input type="button" value="搜索"/>		
共有5907条新闻，当前位于第2页，共有296页。		
在线教育引领“学习的革命”	在线教育引领学习的革命来源：《大河报》2017-04-07赵志疆4月6日上午，河南省人民政府与清华大学签署《省校战略合作协议》。当天下午，河南日报报业集团与清华大学的清华教育投资有限公司签署《清华大学在线教育战略合作协议》，依托学堂在线，建设河南省在线教育课程（MOOC）平台，将清华大学、清华附中。	2017-04-07 13:57:04+00:00
清华举办2016交叉信息科学国际研讨会	清华举办2016交叉信息科学国际研讨会来源：中国科学报2016-12-22王之康 12月18日，2016交叉信息科学国际研讨会在清华大学开幕。会议围绕理论计算机科学、量子信息、信息安全、大数据与人工智能、计算机与生命科学和金融科技与产业等六个方向，总结了领域前沿最新科研成果，展望了未来发展前景。	2016-12-22 15:37:28+00:00
2017两岸清华科技论坛在苏州举行	2017两岸清华科技论坛在苏州举行两岸清华学者与国内外企业代表共话创新、智能和环保清华大学新闻网10月16日电（记者李晴倩图科研院）10月15日，两岸清华科技论坛在苏州金鸡湖国际会议中心举行，苏州市委常委、副市长吴庆文，清华大学副校长薛其坤院士，台湾新竹清华大学副校长陈信文教授等出席并致辞。论坛由北京清	2017-10-16 11:34:40+00:00
清华大学校长邱勇寄语新生：向美而行，照亮自己也温暖世界	清华大学校长邱勇寄语新生：向美而行，照亮自己也温暖世界来源：中国青年网2017-08-24吴楚刘语潇同学们，世界是多彩的，希望你们与美相伴；人生是漫长的，希望你们向美而行。8月24日上午9点，清华大学2017年本科新生开学典礼隆重召开，校长邱勇发表讲话，3700余名中外新生正式开启清华圆梦之旅。	2017-08-25 15:53:33+00:00

(二) 新闻查询页

(1) 新闻查询页支持关键词查询、多关键词查询、时间范围查询，点击‘搜索’按钮既可以显示在要求范围内中的全部搜索结果；

单关键词查询效果——



多关键词检查结合时间范围限定的查询效果——



搜索整句效果——



(2) 搜索栏目下方会展示出本次查询所消耗时间, 以及结果总数目

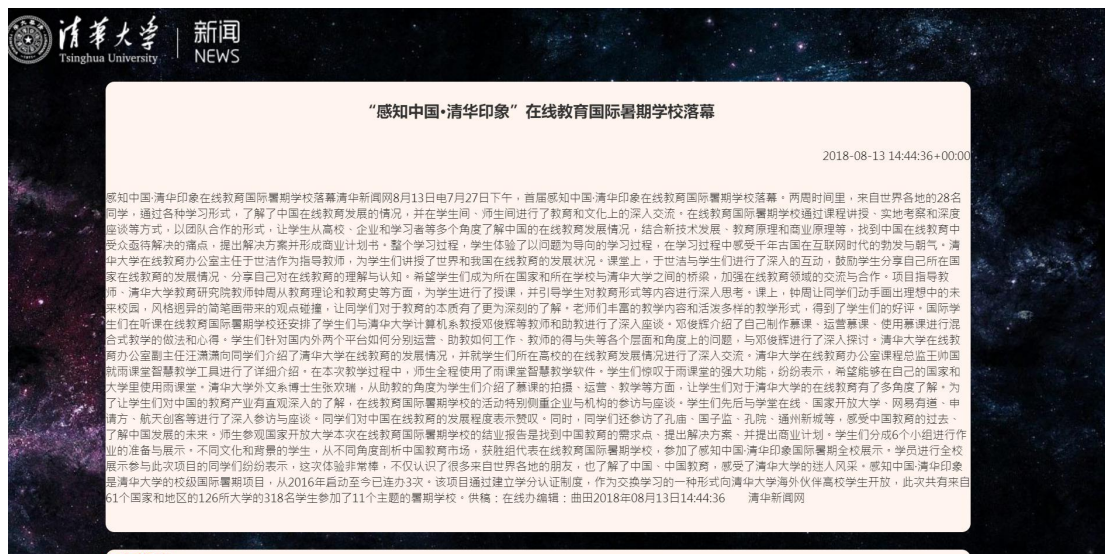


(3) 检查结果显示具备分页功能

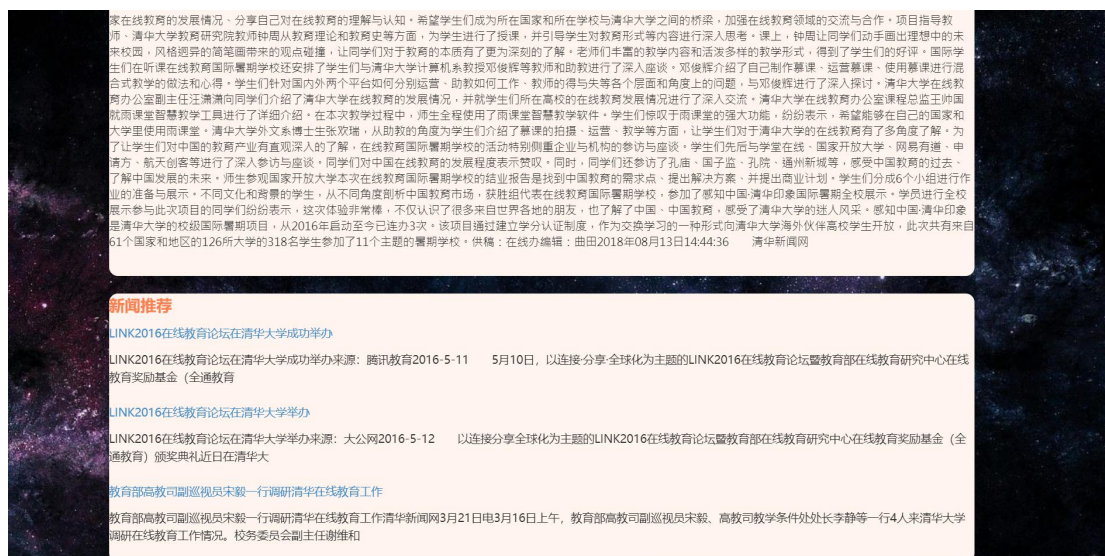


(三) 新闻详情页

(1) 新闻详情页展示出新闻的标题、正文和时间



(2) 详情页有新闻推荐栏目，根据文本相似度判别，点击即可以跳转相应的详情页



三、查询算法与推荐算法

(一) 查询算法

对搜索关键词和语句进行断开处理，存入 wordlist 后，设置时间限制的 filter，与数据库进

行比对查询：

```

for each_word in wordlist:
    curWord = Word.objects.get(pk = each_word)
    if articlelist:
        articlelist = articlelist&set(curWord.article.filter(article_time__gte=datetime.date.today()-datetime.timedelta(days=7)))
    else:
        articlelist = set(curWord.article.filter(article_time__gte=datetime.date.today()-datetime.timedelta(days=7)))
    articlelist = list(articlelist)
    time_start=time.time();
    print(time.strftime('%Y-%m-%d %H:%M:%S'))
    articlelist = sorted(articlelist, key=lambda x:x.article_time, reverse=True)
    dicts = [{ 'id':each.article_id, 'url':each.article_url, 'title':redtitle(each.article_title, wordlist), 'text':redtext(each.article_text, wordlist) } for each in articlelist]

```

(二) 推荐算法

利用包 gensim: 对所有文章进行分词操作，去掉无用词后，生成词典，根据先前建立好的文章索引，获得相关文章的索引，存入矩阵之间，便于后面的查询

```

#对所有文章进行分词操作

texts = [jieba.lcut_for_search(document) for document in documents]

print("分词操作完成")

#去掉无用词
stoplist = list(' , . ( ) 的 地 和 你 它 我 他 是 。 , + - * / : ; [ ] = ! ~ ? < > '.split(" "))
texts = [word for word in texts if word not in stoplist]

print("无用词清除结束")

dictionary = gensim.corpora.Dictionary(texts)

print("词典生成完毕")

dictionary.save(fileaddr+"worddict.dict")

print("词典保存完毕 位置:%sworddict.dict" % fileaddr)

corpus = [dictionary.doc2bow(text) for text in texts]
corpora.MmCorpus.serialize(fileaddr+'mmcorpus.mm',corpus)
print("corpus保存完毕 位置:%smmcorpus.mm" %fileaddr)

```

四、目录结构

以下为项目目录的文件结构（只列出必要的文件与目录）

- project: 项目文件夹
 - ArticleFile1.pickle: 以页面 ID 为索引的分词数据 (1)
 - ArticleFile1.txt: ArticleFile1.pickle 的文本形式
 - ArticleFile2.pickle: 以页面 ID 为索引的分词数据 (2)

- ArticleFile2.txt: ArticleFile1.pickle 的文本形式
- IDFile.pickle: 以页面 ID 为索引的标题、正文、时间、路径
- IDFile.txt: IDFile.pickle 的文本形式
- crawler.py: 爬取新闻网页并对其进行分析, 将分析结果保存至 IDFile
- wordJieba.py: 从 IDFile 读取分析结果, 对正文与标题分词, 分词结果保存至

WordFile

- articleJieba.py: 从 IDFile 读取分析结果, 对正文与标题分词, 分词结果保存至

ArticleFie

- similarity.py: 获取文章文本相似度数据
- WordFile.pickle: 以词组为索引的分词数据
- WordFile.txt: WordFile.pickle 的文本形式
- similarity4.txt: similarity4.pickle 的文本形式
- similarity4.pickle: 文章的相似度数据
- document: 文档目录
- * document: 设计文档
- mysite: 检索系统目录
- * db.sqlite3: 存储页面信息与分词数据的数据库
- * inputarticle.py: 将页面数据导入数据库

- * inputconnect.py: 将分词结果导入数据库
- * inputword.py: 将词组导入数据库
- * inputsim.py: 将文本相似度导入数据库
- * manage.py: Django 生成
- * app: 检索系统目录
 - admin.py: Django 生成
 - models.py: 定义了两个 Model: Article、Word
 - urls.py: Django 生成
 - views.py: 搜索功能函数以及其他函数
 - static: 存储 CSS、js 文件和图片
 - templates: 存储网页模板
- * mysite
 - settings.py: Django 生成
 - urls.py: Django 生成

五、开发版本

Python :3.6.5 Django:2.1.1

```
C:\Users\DELL\Desktop\News\mysite>python --version
Python 3.6.5
C:\Users\DELL\Desktop\News\mysite>python -m django --version
2.1.1
```