**清华大学新闻检索系统设计文档**

**席照炜 计55 2015011354**

**一、简介**

清华大学新闻检索系统是一款基于Django网络框架开发的搜索引擎，它采用了MVC设计模式，模型方面运用了MySQL数据库作为后台支持，视图方面采用Django自带的Python视图机制，前端设计则是html + CSS的设计模式。它利用Python爬虫爬取了清华大学新闻网(news.tsinghua.edu.cn)目录下的所有子网页，并读取新闻的标题、内容、时间等信息，利用jieba分词工具建立MySQL数据库，实现后端数据的获取。前段则是搜索框加检索链接的标准界面，简单实用。

**二、需求分析**

**1、爬取清华新闻网中的新闻信息**

分析：利用Python进行爬虫，如何做到不重不漏，且要模拟浏览器行为。

**2、处理爬取页面信息**

分析：对html去标签化处理，在此之前需要利用标签来将信息分类，并抓取关键信息。

**3、进行分词**

分析：利用jieba分词工具

**4、建立倒排列表**

分析：利用MySQL数据库，将内容存在文件，检索对应信息存在数据库。

**5、实现网页**

分析：学习html和css的有关知识，还有js和jQuery等高阶手段。

**6、拓展功能**

分析:美化界面、分页、高亮、筛选等，在此不赘述。

**三、功能实现**

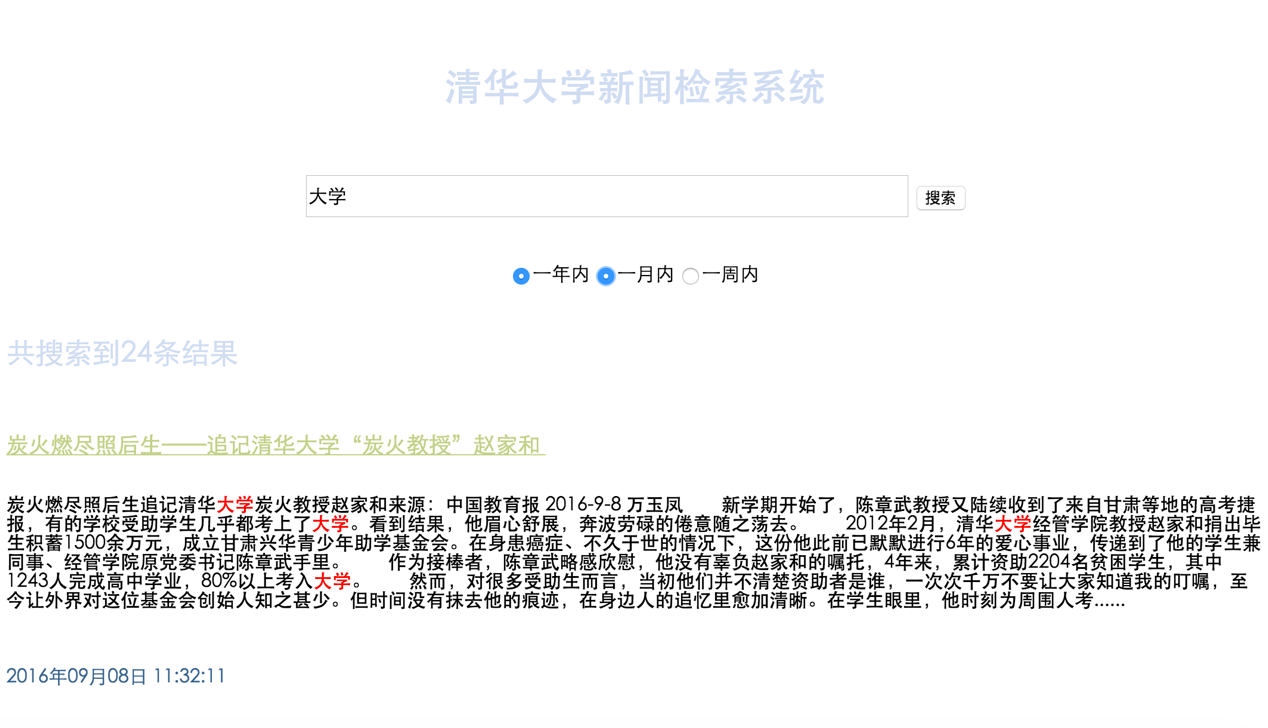
由于上传文件中不含爬取的html文件，记录标题、内容、时间的文件，也没有用于搜索的数据库的相关信息，所以只好以图片的方式展示本搜索引擎的功能。

**1、基本的搜索功能**

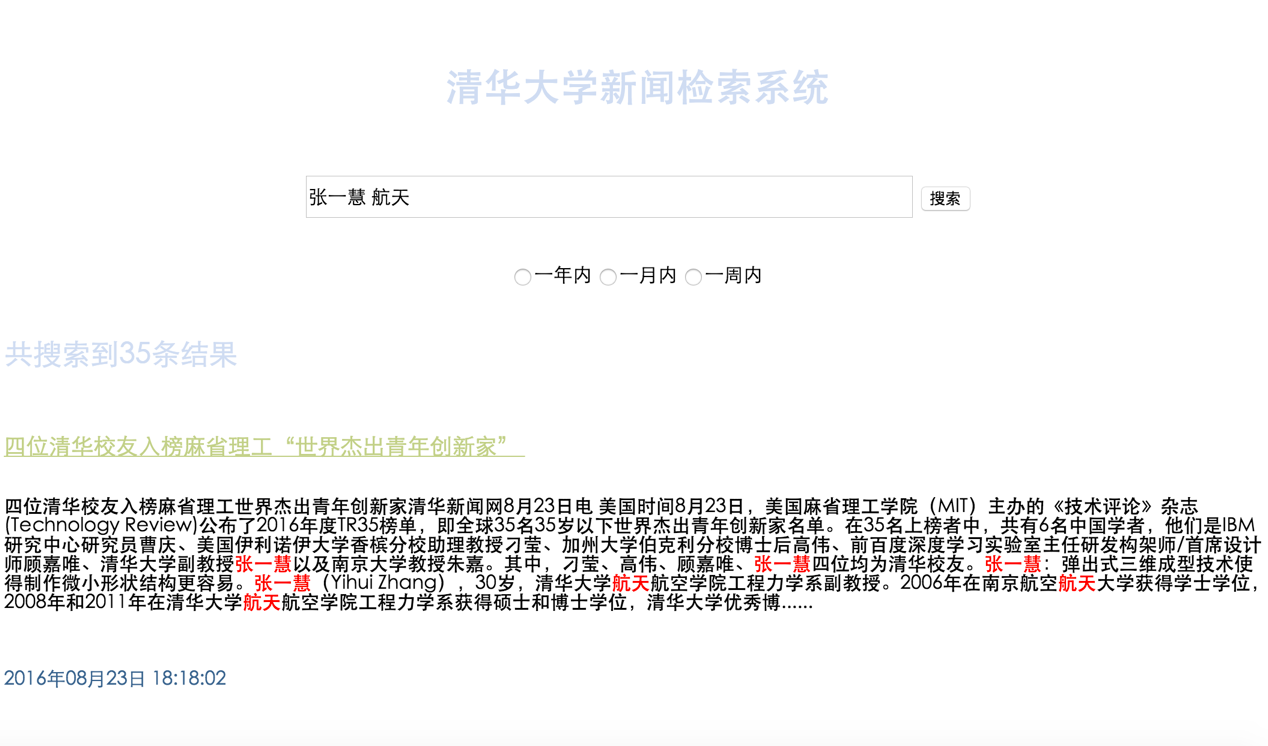
能根据输入的关键字返回相关链接



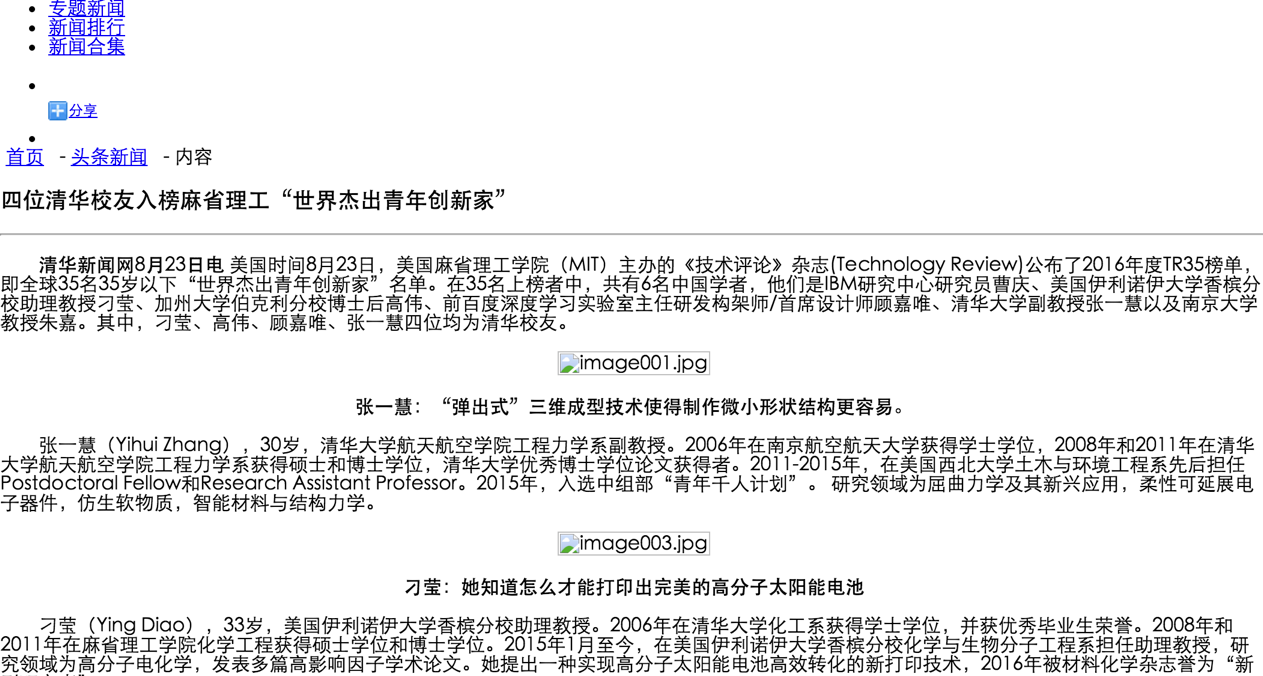
**2.关键字高亮（已展示**）

**3.根据时间筛选显示结果**  


**4.多关键字查询**



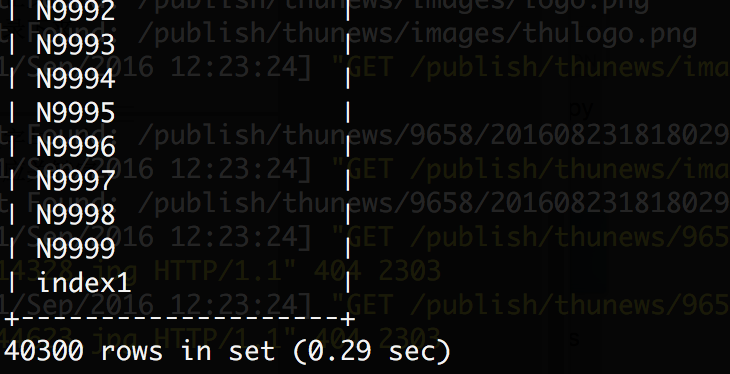
**5.超链接形式打开网址**



**6、后台搜索实现(基于MySQL数据库)**

表格清单（一张索引表和许多网页ID映射表）



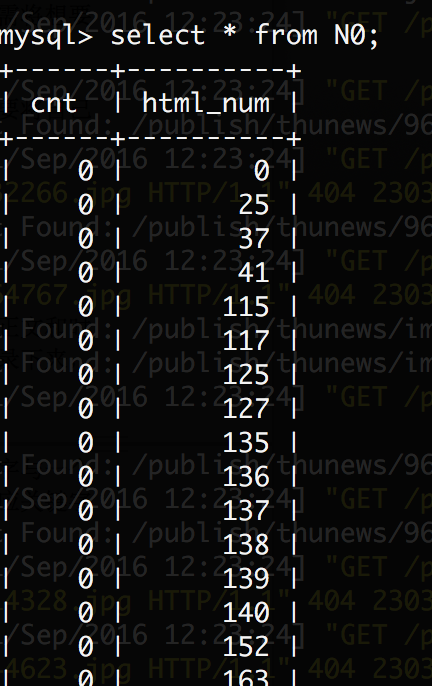


索引表概览



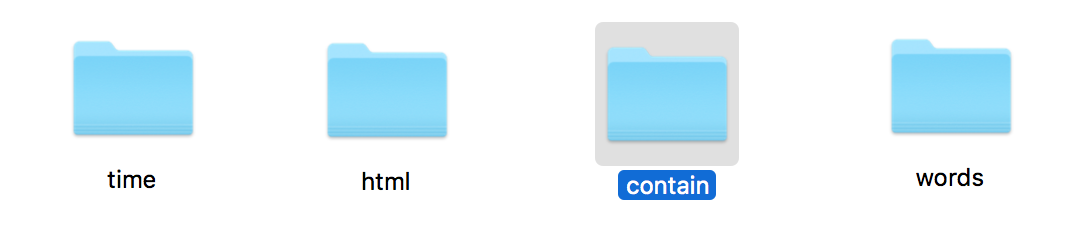


网页ID映射表概览

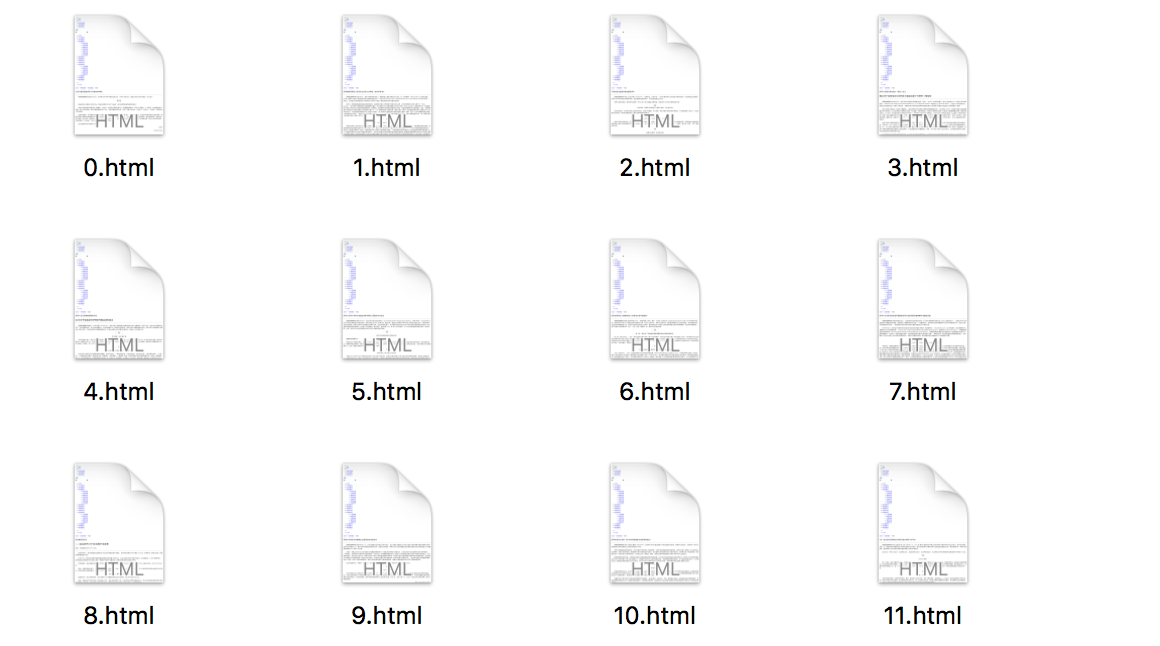


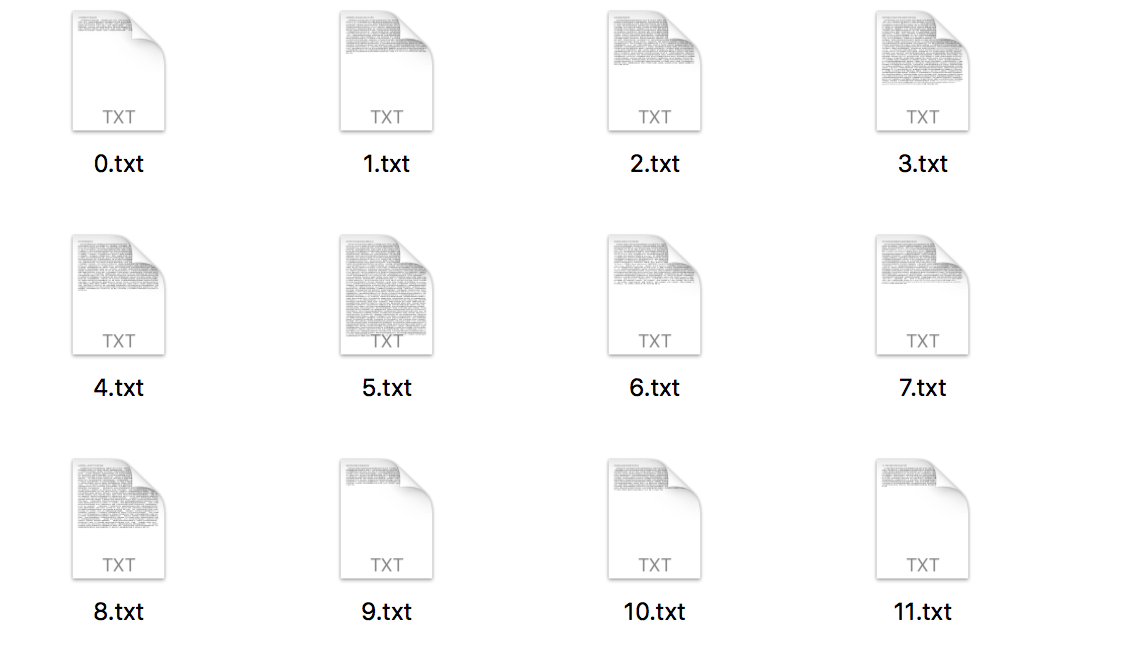
**7、爬取内容展示**

包含四个部分：html(爬取的网页)，contain(文章的标题和内容)，time(文章的时间)，words(文章内所含词语)



内部展示





**四、设计架构**

主要采用Django网络框架，基于MVC模式设计开发。

***在SearchNews/Search文件夹中***

**1.views.py**

采用了单一视图对应多个html的渲染模式，运用GET方法获得所需参数。index视图方法判断目前处于初始状态、检索状态还是找不到状态，调用render函数分别渲染home.html，index.html，notfound.html（这些html文件放在templates文件夹中）。而对于检索所需的参数，采用GET方法来获取。

**2.models.py**

由于事先建好了数据库，并不需要在models.py中建数据库。只需要在views.py中用MySQLdb.connect()函数来连接数据库即可。

***在SearchNews/SearchNews文件夹中***

**1.urls.py**

这是Django自己设计的用于网址url与视图方法之间的映射表，只需将想要的url地址与上述的视图方法写入映射表。

**2.settings.py**

这是将自己建立的app应用程序文件夹加入工程的地方，同样只需要将自己建的Search这个app加入表中。

在hw文件夹中

**1.hw.py**

实现网页爬虫，在爬虫过程中重点关注html文件，并注意去重。

同时还调用save\_in\_file()函数进行网页信息抓取。

**2.file.py**

实现save\_in\_file()函数，通过html标签来抓取网页中新闻的标题、正文和时间，并利用jieba分词工具进行分词，把一篇文章中包含的词语信息也记录下来，将这些信息以文件的形式保存在本机服务器中。

**3.sql.py**

读取由file.py创建并写入的文件，先建一张索引表，把每一个关键字与一个ID对应起来，然后每一个ID i建一张名为N i的表，记录包含这个ID对应关键字的html文件的ID。由此完成检索的映射。