清华信息网新闻检索系统—设计文档

计 52 周京汉 2015011245

1 开发背景

应用 python 爬虫功能获得新闻信息及用 Django 框架搭建的一个新闻检索系统。

2 需求分析

2.1 获得网页

应用 python 爬虫功能爬取清华大学清华新闻网中的全部新闻信息并且对爬取得到的信息进行处理。为每一个新闻页面赋予一个 ID，抽取新闻中的主题关键内容并对这些关键内容进行 jieba 分词并建立倒排列表。

2.2 检索网页

网页需包含输入文本框和搜索按钮。用户可以在新闻检索的网络界面中输入关键词检索出包含相应关键词的新闻，并且可以检索出的新闻数量上限大于 500 条。并且检索网页所需时间需小于 1 秒。搜寻的时候支持多个关键词的查询，并且用户输入的关键词也要进行 jieba 分词。

2.3 附加需求

运用 CSS 框架美化页面。对搜索结果进行分页，并在搜索结果部分对关键词进行红字标识。

3 软件概述

3.1 软件介绍

本软件是跨平台的新闻信息检索系统，名叫 Search_TsinghuaNews，实现通过关键词查询所有清华的新闻，提供该新闻的正文，并且提供新闻原网站链接的功能。

3.2 运行环境

支持 Windows，Linux 和 OSX 平台（需要安装 Django）。

3.3 文件组成

文件包含 3 个部分。第一部分是 Tools 文件夹中的全部工具，这些工具是用来爬取清华新闻网中的全部新闻，并且对其进行解析，提取正文，和进行 jieba 分词的工具。第二部分是一 manage.py 控制的的 Django 框架，包含 searchnews, static, collection 和 templates 四个文件夹，分别包含了 django 框架的 python 程序（包括 url 等），和程序所用的 jQuery 文件，和处理网页信息的的 python 程序和网页的 HTML 模板。第三部分就是新闻检索所用的数据库。数据库共有两个，第一个包括了全部新闻的所有关键词及其所对应的网页的 id，第二个数据库包含了所用网页的 id，及其所对应的网页的新闻标题，内容及网址。

3.4 文件运行

本新闻检索系统要通过控制台进行运行。先在控制台中到达本软件所对应的文件夹下，然后输入：python manage.py runserver, 本检索软件便在 127.0.0.1:8000 网页开始运行了，此时只需在浏览器中打开此网页即可。

4 整体架构

本检索软件的主题有 django 框架构成，主要包含三个部分。

4.1 管理部分

第一个部分程序主要在 searchnews 文件夹中，主要管理网页运行时对于进入其目录下的请求和其所对应的需要调用的 python 程序。这一部分的主要代码都是由系统自动添加而成的，只需要在 url 文件中添加网页的目录下包含的部分和所需的调用的函数。

4.2 解析部分

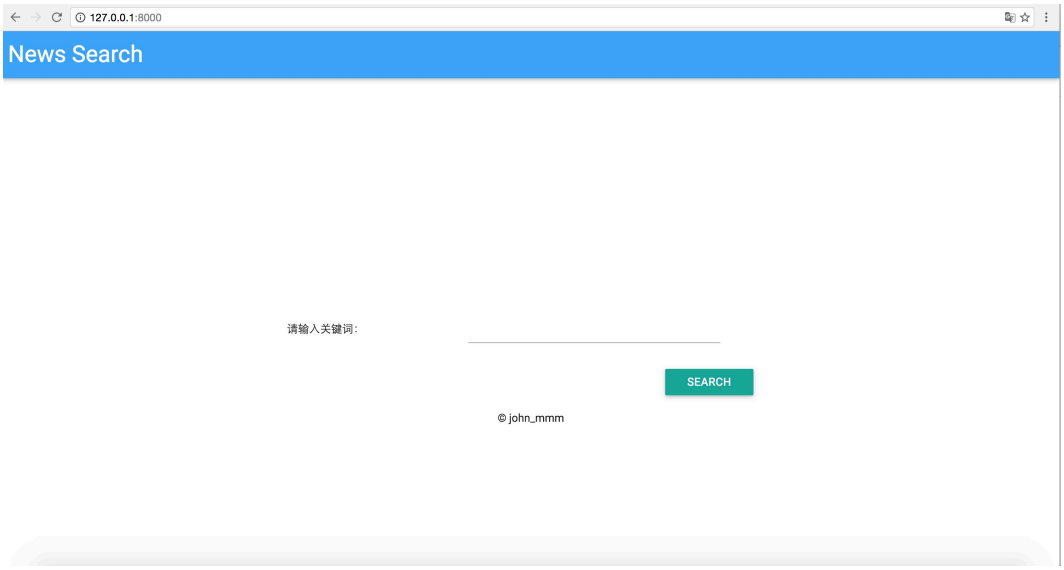
第二个部分主要是在网页发出请求的时候调用的程序，它对网页中传入的 request 进行解析获得信息并读取数据。

对于查询需求，我先将其运用 jieba 分词进行解析，获得更为详细的关键词然后在调取数据库获得所有关键词对应的网页 id。然后应用 set 函数将 id 取交集，得出最后要返回的全部网页的 id。由于空间有限。我最终只是输出了全部符合要求的网页中的前 540 个。并将全部的网页的 id，题目和关键词作为参数返回给 query.html。

对于查询一个网页更加详细的需求，程序会先读取数据库，网页返回的网页的 id 的详细信息，包括题目，新闻内容和新闻原网页的链接，然后将这三项内容作为返回值返回给 news.html。

4.3 网页模板

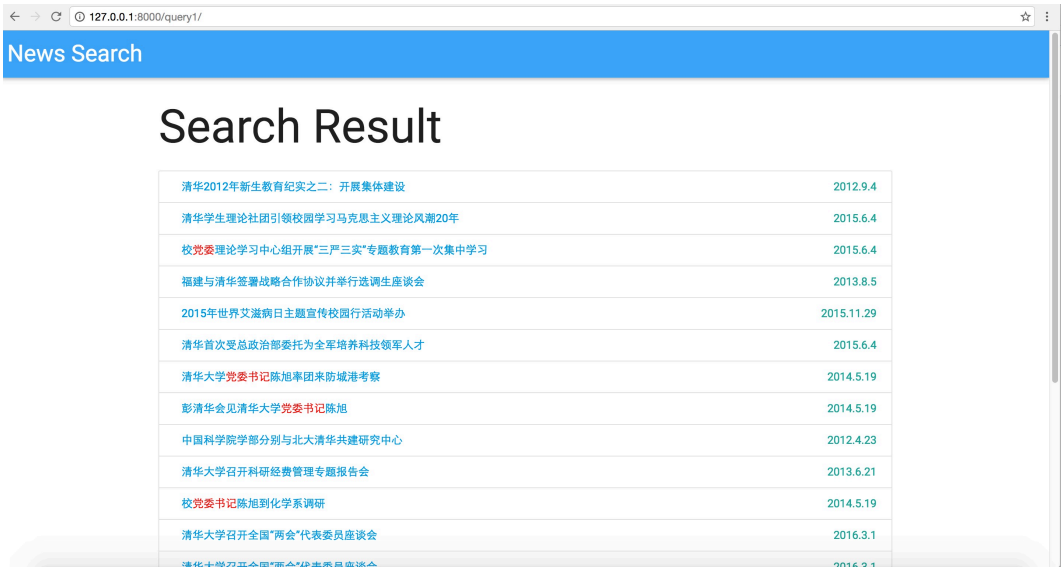
第三个部分主要包含网页建立所需的 HTML 模板和 jQuery 所需的 js 文件。网页模板为几个网页查询页面的页面组成的基本模板，总共有四个程序。其中 navbar.html 程序为所有网页的标题设置，应用 CSS 模板，可以在任意网页点击该网页并返回主页。index.html 为主页的模板布置，应用“container”包含一个用于输入的横线和一个点击搜索的按钮；query.html 为显示搜索结果的页面，包含了一个表格和分页的页码，每页可显示 20 个搜索结果；news.html 为显示新闻信息的页面，显示了一些文字信息，包含新闻标题，新闻正文内容和原网页的链接。后三个模板中均包含了第一个模板作为整个网页的大标题，点击这个大标题可以返回网页的主页。



5 其他设计

5.1 分页功能

本检索系统实现了分页功能，每页显示 20 条搜索结果，最多可以有 27 页，通过点击页面下方的按钮来切换页面。分页功能的实现应用了 pagination 类，将全部的搜索结果等分，每页 20 个结果显示出来。



5.2 关键词标红功能

在显示结果的 query.html 和显示新闻的 news.html 中，我应用 JavaScript 中的 jQuery 寻找所有的包含的关键词，并在其两边加上标签，令其文字颜色变为红色，一次来实现关键词标红。