|  |
| --- |
|  |
| **大作业设计文档** |
| 小学期第三周：导师查询系统 |
|  |
| **计15 2011011300 周若凡** |
| **2012/9/9** |

|  |
| --- |
| [前期由python编写的网页爬虫爬取资料，并进行分词和建立倒排列表、服务器端由Django实现的B/S架构查询系统。] |

1. **功能介绍**

（开启方式：开启服务器后输入网址<http://127.0.0.1:8000/search>）

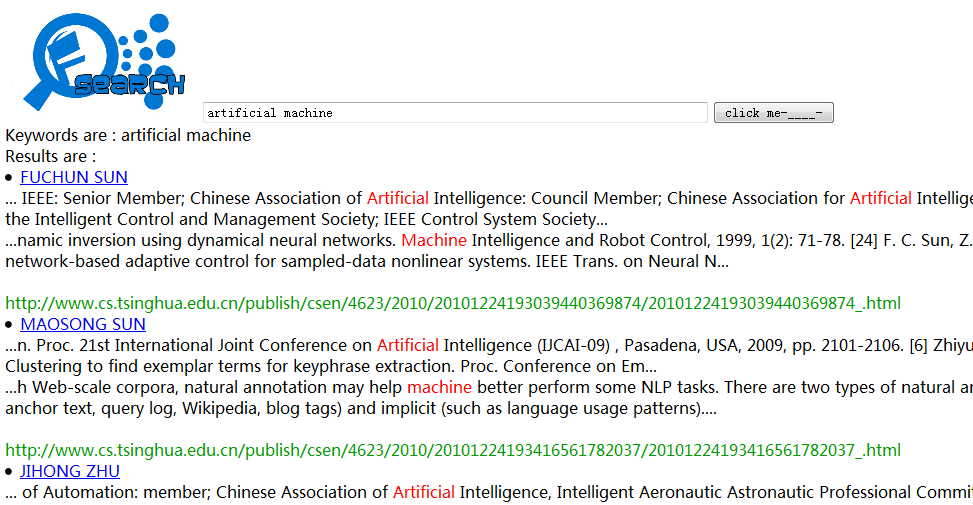
**1．输入导师姓名**



输入导师姓名（如上图），点击按钮后网页会自动连接到该导师主页。

\*大小写忽略。但字符必须全部匹配。

1. **输入关键词**



输入关键词，则对爬取的网页进行全文搜索（多个关键词则以空格为分割，且最终结果取交集）。返回如上图列表，会返回符合条件导师姓名（点击后可进入其主页）、关键词附近信息以及导师URL。

符合条件的导师以爬取网页的顺序，即ID序号从小到大给出。

\*大小写忽略

**2.爬取、分词**

共爬取269个网页（webs目录下即爬取的网页），其中包括清华网页上爬取的导师95名，CMU网页上爬取的导师174名。所有爬取下来的网页都进行过处理，去除无用html标签以及网页中大部分与导师无关的内容。

采取暴力分词法，提取所有网页，以空格，标点等为分界的所有单词并进行倒排列表的建立。

**3.文件功能阐述**

1./manage.py、/mysite/\_\_init\_\_.py、/mysite/settings.py、/mysite/wsgy.py以及/mysite/urls.py都是Django生成的工程文件。

2./mysite/view.py为后台工作程序，完成了接收信息到处理到返回的过程。

3./mysite/z\_htmlParser.py为基本的网页爬虫，利用递归（最深3层）爬取网页url信息。

4./mysite/z\_selectHTML.py为选取与导师相关网页的爬虫2。其中/mysite/webs/中所有的网页都由该程序爬下。

5./mysite/z\_rewrite.py为删除爬下资料无用信息的程序。

6./mysite/z\_wordsearch.py则用来分词，取关键字。

7./mysite/z\_sort.py用来建倒表。

8./mysite/z\_NameList.txt为导师姓名列表；/mysite/z\_UrlList.txt为爬取网页的url；/mysite/z\_Wordlist.txt为分词出来的关键词列表；/mysite/z\_sort.txt为倒排列表

9./logo.jpg是为网页设计的有标记。