信息检索大作业结果文档

作业名称： 小型搜索引擎——Seeker

小组成员：

姓名	学号	班号
刘家强	2015303181	14011505
李泽豪	2014302644	10011407
严珂	2014301549	10011405
何祺玮	2017371014	10011401
梁瑀	2014302645	10011407

引言： 本次信息检索课程大作业，我们小组选择自己搭建一个小型的搜索引擎——Seeker，语料库是爬取的西工大新闻网站(http://news.nwpu.edu.cn)中的一百多个网页，所用语言是python，遵守了标准的信息检索过程。项目成果可以访问 (http:// 114.115.206.83:80)

过程详述：

整个项目大体可以分为两个部分：

- 文档集收集
- 文档检索

(详情请查看源代码)

1. 文档集收集

在本项目中，文档集收集采用的是爬虫技术，主要文件放在/Spider 文件夹目录下。大致过程为：以 http://news.nwpu.edu.cn 为起始网页-->获取网页内容，并解析网页中的链接→将解析出的链接放入 url 队列中→将访问过的链接放入已访问列表中，从队列中获取新的 url 链接，采用多线程爬取。

针对每个访问的网页，解析出标题、时间、网页网址、主体内容，将每个网页抽象为一个 URL 类，并将以上信息写入/Main/Info/*.txt 文件中。此外，每个 URL 类还将记录页面中链出的链接，最终根据所有网页的互相链接信息，构成链接矩阵，迭代运用 pagerank 算法，得出每个网页的 rank,从而可以为每个网页的重要性提供参考值。

在此过程中，还构建了倒排表。由于爬虫可以获取每个网页的内容，所以所有的网页内容构成了语料库。在对原始文档的预处理中，运用的是 python 的结巴分词工具包。在对文档的词干化，规范化之前，还做了停用词过滤，停用词集合在/Spider/stopword.txt 中。构建好的倒排表存储在小型数据库 sqlite 中，保存在/Main/my_engine_data_base.db 中。

2. 文档检索

本搜索引擎的网页前端入口采用的是 python 的 web.py 模块。文档检索过程大致为：前端页面获取用户输入搜索关键字→对搜索关键字切词(词干化，规范化，停用词过滤)→计算搜索关键字和每个文档的 tf-idf 评分→按序返回评分高的网页(由于文档集小的原因，如果没有发现相关文档，则随机返回一些网页)→页面前端结果显示

最终效果图片展示(可自行访问 http://114.115.206.83:80 体会)：

Seeker 西工大 Seeker !

Seeker为您找到了约15条结果，用时0.0834251485971秒

视窗-西北工业大学新闻网

闻网 首 页 工大要闻 校园动态 媒体工大 通知公告 摄影报道 西工大报 网络视频系统 党委宣传部 宣传教育 高教视点 科技前沿 网观天下 学术信息 电子读报系统 校园风光 热点专题 “两学一做”学习教育 聚

http://news.nwpu.edu.cn/index.htm

视窗-西北工业大学新闻网

闻网 首 页 工大要闻 校园动态 媒体工大 通知公告 摄影报道 西工大报 网络视频系统 党委宣传部 宣传教育 高教视点 科技前沿 网观天下 学术信息 电子读报系统 校园风光 1 2 3 4 5 u_6pnw

http://news.nwpu.edu.cn/index.htm

校园动态-视窗-西北工业大学新闻网

闻网 首 页 工大要闻 校园动态 媒体工大 通知公告 摄影报道 西工大报 网络视频系统 党委宣传部 宣传教育 高教视点 科技前沿 网观天下 学术信息 电子读报系统 校园风光 校园动态 当前位置: 首 页 >>

http://news.nwpu.edu.cn/news/xyxw.htm

工大要闻-视窗-西北工业大学新闻网

闻网 首 页 工大要闻 校园动态 媒体工大 通知公告 摄影报道 西工大报 网络视频系统 党委宣传部 宣传教育 高教视点 科技前沿 网观天下 学术信息 电子读报系统 校园风光 工大要闻 当前位置: 首 页 >>

http://news.nwpu.edu.cn/news/gdyw.htm

宣传教育-视窗-西北工业大学新闻网

闻网 首 页 工大要闻 校园动态 媒体工大 通知公告 摄影报道 西工大报 网络视频系统 党委宣传部 宣传教育 高教视点 科技前沿 网观天下 学术信息 电子读报系统 校园风光 宣传教育 当前位置: 首 页 >>

http://news.nwpu.edu.cn/wenhua/xcjy.htm

专题列表-视窗-西北工业大学新闻网

闻网 首 页 工大要闻 校园动态 媒体工大 通知公告 摄影报道 西工大报 网络视频系统 党委宣传部 宣传教育 高教视点 科技前沿 网观天下 学术信息 电子读报系统 校园风光 专题列表 当前位置: 首 页 >>

http://news.nwpu.edu.cn/zdzy/rdzt.htm

西工大举行2017届本科毕业生代表座谈会..

西工大举行2017届本科毕业生代表座谈会-视窗-西北工业大学新闻网 首 页 工大要闻 校园动态 媒体工大 通知公告 摄影报道 西工大报 网络视频系统 党委宣传部

http://news.nwpu.edu.cn/info/1002/49878.htm

校长汪劲松与学生面对面“论政”-视窗-西..