需求报告——搜索引擎与文本挖掘

日期：2016.10.27

目标：在理解搜索引擎原理及整体流程的基础上，通过亲自动手搭建一个完整、可运行的小型全文检索实验系统，进一步加深对搜索引擎系统底层实现的理解和掌握，熟悉搜索方面的一些经典算法和思想。

可选工具

索引、检索工具：**Lucene**、Indri、Firtex 等

网页爬行器（Crawler）：**Weblech**、Nutch Crawler、Wget、Larbin 等

Web 服务：**Tomcat**、Apache 等

分词组件：对于中文语料来说，不能采用 Lucene 默认的分词器（如 StandardAnalyzer 等），必须引入新的、适合中文的分词器。可以自己实现一个简单的基于最大匹配的分词程序，也可以上网下载 Lucene 的扩展分词组件包 CJKAnalyzer（二字串方式分词），当然也可以采用第三方分词组件，如中科院 ICTCLAS 分词组件等。

功能

要求（必须）：

1、 利用开源的网页爬行器或者自己开发的网页爬行器，爬取一定数目的网页。

2、 对网页进行必要的去噪，预处理工作。

3、 采用 Lucene 等全文检索工具包对数据建立倒排索引，并能提供检索服务

4、 返回给用户的应该是一个进过相关性排序的结果列表

5、 不能采用 Lucene 默认的分词器分词，必须完成一个中文分词接口。

6、 必要的预处理工作。

   中文：分词、停用词过滤。

   英文：大小写转化（Case insensitive）、词干化（Stemming）、停用词过滤。

7、 结果高亮显示

注意事项：

1、 中文分词可以借助现有的组件，如中科院 ICTCLAS 分词组建、哈工大 IRLAS 词法分析系统等。这时需要做的是，根据 Lucene 的 Analyzer 接口规范将上述分词组建进行封装，使之可被索引器、搜索器等调用。

2、 英文的词干化 （Stemming）可以调用网上现有的算法实现，比过 Porter Stemming Algorithm 算法等。

3、 用户界面可以仿照 Google、百度等，不要太花哨，简洁、大方即可。

扩展功能(可选)：

1、 校内 FTP 在线检索

2、 查询扩展（相关反馈）

3、 改进检索算法、模型

4、 其他可选的预处理工作，如人名、地名、组织名识别等。

实验环境设置：

为了能够对系统有一个定量的衡量，现提出一些限制条件：

1、 实现过程采用的工具、语言等没有限制，但尽量选用一些开源、目前比较成熟、稳定的工具

2、**至多两个人一组**（最后提交报告中**必须**说明每个人的工作）

最终评价指标：

1、 系统的完成情况

2、 系统的检索效果

3、 系统的相应时间

4、 用户界面的友好性

5、 报告的撰写情况

书面报告内容：

1、 运行环境

机器配置、选用工具等

2、 分工、每个人完成情况

所做的预处理工作及每一步采用的工具等

3、 结合自己的实际情况，写 500 字左右的总结

联系方式：

Email: dut_ir@163.com (**只发送源码和报告**发到此邮箱)（**不需要发送爬取的网页和索引**）

链接： http://ir.dlut.edu.cn（资料下载）

作业提交：

时间：2016 年 12 月 25 号之前

地点：创新园大厦 A0923 房间