

期末大作业说明（大数据系统综合实验2024）

一、实验概述

从科大的网站爬取文件数据，存在分布式数据库（HBase）中，做一个搜索引擎，实现校内文件搜索的目的。

二、实验要求

1. 尽可能多的爬取科大的网站，要求包含各学院官网以及各管理部门官网，文档最后列举了一部分网站。
2. 爬取内容最低要求：若网站包含“下载中心”“此类文件专栏，则需包含，例如中科大财务处：



3. 搜索引擎需实现的基本结果是：根据搜索内容召回文档（独立的文件）或文本信息，召回结果应按照相关性和契合度排序。
4. 请尽可能多的使用本课程中学到的技术。

三、实验形式

1.组队情况

本作业可以由1-3人组队完成，按照之前问卷中各同学提交的组队名单完成，未及时提交的同学默认单人完成。如有特殊情况请与助教联系。

2. 实验说明

项目的实验环境在本地进行搭建。(由于开源软件的不同版本搭配，可能会存在各种各样的 bug，这里提供了一个[实验环境搭建](#)进行参考，也可以参考平时实验的实验平台)

3.验收说明

1. 验收需求： 课程设计汇报+课程实验报告提交（附源码）
2. 课程设计汇报时间： 分两次课汇报（暂定12.19 和 12.26，后续可能根据进度调整），汇报形式和材料可自行决定， 可以用实验报告.pdf 或额外制作 PPT或视频等。
3. 实验报告提交截止时间： 汇报结束一周内， 将实验报告提交至助教邮箱：
lvhang1001@mail.ustc.edu.cn。
4. 实验报告命名格式： 学号_姓名_exp.zip（多人组队只需要写队长的学号姓名即可） 如： PB21000001_张三_exp.zip。如有其他特殊情况请在邮件正文中说明。

4.实验报告内容

实验报告形式自由，但至少需要包含的信息有

- 1、小组成员名单和具体分工
- 2、技术路线（介绍一下该项目用到的主要技术并做简要介绍，尤其是与本课程相关的技术）
- 3、实现功能介绍和相应的效果展示
- 4、核心代码块（可截图放上去）
- 5、该组所有同学各自的总结与心得（如踩坑、错误总结、实验收获等）

参考网站：

大数据学院 <http://sds.ustc.edu.cn/main.htm>
中科大本科招生网 <https://zsb.ustc.edu.cn/main.htm>
中科大就业信息网 <http://www.job.ustc.edu.cn/index.htm>
中科大教务处 <https://www.teach.ustc.edu.cn/>
中科大财务处 <https://finance.ustc.edu.cn/main.htm>
学工一体化 <https://xgyth.ustc.edu.cn/usp/home/main.aspx>
中科大研究生院 <http://gradschool.ustc.edu.cn/>
中科大保卫与校园管理处 <https://bwc.ustc.edu.cn/5655/list.htm>
中科大出版社 <http://press.ustc.edu.cn/xzzq/main.htm>
中科大信息科学实验中心 <http://ispc.ustc.edu.cn/6299/list.htm>
中科大科技成果转移转化办公室 <http://zhb.ustc.edu.cn/18534/list1.htm>
青春科大 <http://young.ustc.edu.cn/15056/list.htm>
中科大网络信息中心 <http://ustcnet.ustc.edu.cn/main.htm>
中科大资产与后勤保障处 <https://zhc.ustc.edu.cn/main.htm>
中科大计算机科学与技术学院 <http://cs.ustc.edu.cn/main.htm>
中科大网络空间安全学院 <http://cybersec.ustc.edu.cn/main.htm>
中科大数学科学学院 <https://math.ustc.edu.cn/main.htm>
中科大信息科学技术学院 <https://sist.ustc.edu.cn/main.htm>

中科大苏州高等研究院 <https://sz.ustc.edu.cn/index.html>

中科大软件学院 <https://sse.ustc.edu.cn/main.htm>

中科大先进技术研究院 <https://iat.ustc.edu.cn/iat/index.html>

..... (可行选择)