期末大作业说明（大数据系统及综合实验 2023）

一、实验概述；

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

二、实验要求：

1、选取不少于 8 个科大网站；
2、爬取内容的最低要求：若网站包含“下载中心”此类文件专栏，则需包含；例如：

3、尽可能多的用到本课程中的技术；

三、实验形式：

1. 组队情况说明

1-3 人组队，组队信息于 12 月 1 日 23：59：59 前提交至问卷：
https://docs.qq.com/sheet/DSEJKQmJtYmNQdFFU?tab=BB08J2
想单独一人做的同学也请填写一下，逾期未填的同学默单独一人；

2. 实验环境说明

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. 实验验收说明

验收需求：课程设计汇报+课程实验报告提交（附源码）
课程设计汇报时间：分两次课汇报（12.22 和 12.29），汇报形式材料可自行决定，实验报告.pdf 或额外制作 PPT
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：wengbingjie@mail.ustc.edu.cn。实验报告命名格式：学号_姓名_exp.zip（多人组队只需要写其中一人的学号姓名即可）如：PB19000666_张三_exp.zip

4. 实验报告说明

需包含：

1、组员名单和具体分工

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

4、核心代码块（可截图放上去）

5、该组所有同学【都要写】各自部分的总结与心得（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
中科大本科生招生网 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）