**2024-2025小学期大数据系统开发实践项目要求**

**项目目标**

通过大数据处理案例进行大数据技术的理解和运用

1, 大数据需求分析

2, 大数据技术框架的理解与应用

3, 大数据的结果分析与展示

**要求1： 不超过6人组成一个小组，不限班级，在相应的技术方案文档中写清楚小组成员及分工。**

**要求2：代码实现的部分需要在分布式环境下进行，编写实验报告（可以与技术文档放在一起），实验报告中给出实现结果的截图。**

**项目内容**

**一、项目1**

1. 技术方案-编写搜索引擎实现的技术方案
   1. 功能描述：描述所实现的搜索引擎的功能结构、各分项功能。
   2. 技术选型：（要求选择使用Hadoop生态圈技术)，并说明选型依据，阐述使用Hadoop进行开发搜索引擎的优势（**技术方案要求的重点内容**）与不足，可以根据技术方案做进一步说明补充。
   3. 功能实现：结合上面的功能描述与技术选型，描述主要的功能（**主要与Hadoop有关的部分**）的实现方案，例如，倒排索引的实现等。
   4. 工作计划：开发阶段划分，每个阶段的工作安排。
   5. 组织结构：描述开发小组成员（可以虚拟）及分工。
   6. 软件质量保证、非功能性保证等，可以简写。
2. 实验报告-实现文档的倒排索引
3. 运用MapReduce算法计算,构建一个倒排索引, 将倒排索引存储在HBase中
4. 数据集，压缩文件sentences.txt.zip,大小500MB，解压文件1.43GB，下载地址：i北理课程群
5. 下载数据之后，按照文件大小或者句子数量（例如10000个句子）构成一个文件，形成一个文件集合。可以编程实现文件分割或者使已有的文件分割工具软件。
6. 建议的软件环境

* VMware WorkStation 16 Pro
* hadoop-3.3.0
* hbase-2.1.0
* jdk1.8.0\_241
* zookeeper-3.7.1
* CentOS Linux release 7.7.1908 (Core)

**三 作业包括文档（pdf）和代码**

1. 实验报告的内容：（建议）

* 实验要求
* 数据准备
* 环境的安装与配置
* 算法及实现
* 运行结果与分析
* 总结（心得、体会）

**四 提交时间： 9月29日（周日）23：59分截止。**

**五 提交方式：邮件至guoguisuo@bit.edu.cn**