**贵州民族大学人文科技学院**

**本科毕业论文（设计）任务书**

**学院：**大数据与信息工程学院 **年级：**2017级 **专业班级：** 计算机科学与技术2班

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **学生姓名** | | 李苇 | **指导教师** | 李卿 | **职称** | 讲师 |
| **论文（设计）题目** | | 基于python与spark的大规模数据采集平台设计 | | | | |
| **毕业 论文(设计)工作 内容** | 1. **研究内容**   **1、在这个信息时代，未知的信息对每个人都有着吸引力，比如需要付费的百度文章，VIP电影，微信小程序等。很多时候简单的搜索引擎已经满足不了人们对信息的采集，此设计的目的是为了方便人们对所需信息数据的获取，确保获取信息的真实性和实用性。**  **2、本平台可用于各个学校，图书馆，企业等。它可解决人们找某个范围的资料或是信息时需要在纷繁复杂的浏览器或是APP上查找来减少人们的工作量。**   1. **主要要求**   **1、spark框架以读取文件数据，将数据进行拆分，统计，转换，最后将转换的结果采集到控制台上，它可以快速的处理数据集。将爬虫技术和spark框架结合起来。python对数据进行采集，采集的数据放在MongoDB数据库中，用spark对数据进行计算，在用python将计算结果显示出来。**  **2、实现人们对信息的查询功能。**  **3、本系统适用于广大的人民群众，让人们每时每刻都能查询自己想要的信息。**  **4、性能要求:此系统具有稳定性，可控性，安全性，可维护性，响应时间快，支持多个用户同时查询。**  项目的推进目标：   1. 第一个月达成需求分析，完成开题 2. 1.5个月达成框架设计，中期报告 3. 2个月达成最小实验系统搭建 4. 3个月完成系统搭建并测试系统稳定性 5. 3.5个月完成论文初稿，提交初稿 6. 初稿润色修改，完成答辩PPT，准备答辩 | | | | | |
| **指导教师： （签名） 年 月 日**  **系 主 任： （签名） 年 月 日**  **学院院长： （签名） 年 月 日** | | | | | | |
| **备注** | 该表由学生与指导老师共同讨论后，确定论文题目，并在“工作内容”栏填写“完成论文撰写所需完成的主要任务”。 | | | | | |