附件一：

苏州大学本科生毕业设计（论文）任务书

学院：计算机科学与技术

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 设计（论文）题目：面向高校学者信息大数据平台的分布爬虫设计与实现 | | | | | |
| 指导教师 | 王进 | 职称 | 教授 | 类别 | 毕业设计 |
| 学生 | 魏素忠 | 学号 | 1527406028 | 设计（论文）类型 | 应用型 |
| 专业 | 软件工程 | 班级 | 15软件工程 | 是否隶属科研项目 | 否 |
| 1. 设计（论文）的主要任务及目标   主要任务：   1. 了解分布式爬虫的工作原理。 2. 针对爬取页面设计合理的爬取逻辑。 3. 了解Hadoop大数据平台的基本架构。 4. 掌握MapReduce编程思想。 5. 对数据进行可视化展示。   目标：通过分布式爬虫获取高校学者的基本信息，运用大数据平台对学者信息进行  清洗和统计，最后以web形式将数据可视化。 | | | | | |
| 1. 设计（论文）的主要内容 2. 解析爬取页面的逻辑结构。 3. 分析分布式爬虫的难点以及解决方法。 4. 分布式爬虫的部署。 5. Hadoop集群环境搭建。 6. HDFS基本操作和MapReduce编程。 7. 整个数据可视化平台的搭建，包括数据库的设计、前后端的通信。 | | | | | |
| 1. 设计（论文）的基本要求 2. 分布式爬虫的设计以及实现。 3. 利用MapReduce进行编程。 4. 掌握web开发的整套流程。 5. 完成3000字的文献综述、不少于3000个单词的外文翻译。 6. 撰写毕业论文时，做到条理清晰，思维严谨，格式遵守科研写作规范，字数达到规定要求。 7. 提交毕业论文，展示毕业成果。 | | | | | |
| 4、主要参考文献  [1]崔庆才.Python3网络爬虫开发实战.人民邮电出版社，2018.  [2]余明辉，张良均.Hadoop大数据开发基础.人民邮电出版社,2017.  [3]林大贵.Python+Spark2.0+Hadoop机器学习与大数据实战.清华大学出版社，2017.  [4]安俊秀，王鹏，靳宇倡.Hadoop大数据处理技术基础与实践.人民邮电出版社，2015. | | | | | |
| 5、进度安排   |  |  |  | | --- | --- | --- | |  | 设计（论文）各阶段任务 | 起 止 日 期 | | 1 | 调查、资料收集，阅读学习相关知识。 | 2018.02.10―2018.02.25 | | 2 | 撰写任务书、文献综述、外文翻译。 | 2018.02.26―2018.03.31 | | 3 | 进行课题的主要设计与编码工作。 | 2018.04.01―2018.04.26 | | 4 | 完善课题的设计、进行论文编写。 | 2018.04.27―2018.05.10 | | 5 | 论文定稿、提交打印。 | 2018.05.11―2018.05.15 | | 6 | 准备答辩。 | 2018.05.16―2018.05.21 | | | | | | |