**XXX大学XXX学院课程实验报告**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 姓 名 |  | 学 号 |  | 学 院 |  |
| 年 级 |  | 专 业 |  | 邮 箱 |  |
| 课 头 号 |  | 课程名称 |  | 指导教师 |  |
| 实验类型 | □独立实验课 ☑含实验的理论课 □计划外自选开放实验  □学生自主式开放实验 □大学生科技竞赛 | | | | |
| 实验名称 | ItgInsight数据清洗与主题词提取 | | | | |
| 实验日期 |  | 实验学时 | 2-4学时 | 成 绩 |  |
| **一、实验概况**  **（一）实验目的与内容**   1. 了解数据清洗在科技文本挖掘中的重要意义。 2. 掌握同名异指、异名同指的含义；ItgInsight进行机构名清洗的基本原理，人名清洗的基本原理，人名机构名清洗的相同与不同点；ItgInsight快速模式清洗的原理，指纹的作用、修改与使用；ItgInsight进行项目、出版物、技术类别、关键词清洗的原理。 3. 掌握ItgInsight进行主题词提取的原理；词长度、词频率进行术语提取的阈值设置；术语度指标、Tfidf指标进行候选术语筛选的过程；了解语法规则词典设置。 4. 掌握ItgInsight各种词表的使用，依据词典进行分组，与DDA词表的相同与不同，将DDA词表转化为ItgInsight词表的方法。 5. 掌握ItgInsight数据关联Goto Related操作方式。分组统计、组合分析，输出excel和txt格式报表。 6. 掌握Dataset进行数据查询、过滤的方式，组合查询、过滤的条件，名称批量替换，大小写转换，文档保存为Excel，列表保存为Excel，Dataset保存为Excel等多种保存为Excel的方式。   **（二）主要仪器设备（含必要的元器件、工具）：**  1、操作系统  Windows 10及以上版本操作系统  2、工具软件  ItgInsight、Notepad++、Microsoft Office系列  3、数据集（以下数据集选一即可）   * 选定自己感兴趣的领域的SCIE/SSCI论文、知网论文或者其他数据集合； * 2012-2021年期间被SCIE/SSCI收录的图书情报领域的研究论文（Article）集合；   **（三）实验内容**  1、运用ItgInsight进行数据清洗，读取数据全部字段，进行著录项目分组，比如机构分组，自动分组，手工分组，撤销分组，保存分组，停止分组，依据词表进行分组，生成词表。  2、基于ItgInsight生成的\*.dataset文件，进行时间分段，精确到年的操作。  3、基于ItgInsight生成的\*.dataset文件，进行分组统计、组合分析，输出excel和txt格式报表、可视化。  4、应用词表功能进行数据的二次读取，观察使用词表前后，数据读取和分析结果的差异。  \*拓展题：  1、编制过滤器，读取软件文件“example\data\自然基金数据”的项目数据。  2、编制过滤器，读取软件文件example\data子目录下名为“example-data-patsnap”的专利数据，从题目和摘要中提取主题词。  3、编制词典，进行二次数据读取，观察使用词典后的\*.dataset与使用前的区别。  4、了解字典的使用方法：正向筛选、反向筛选、正则表达式筛选、大小写设置。  5、了解读取数据时只提取排序第一的作者、排序第一的机构。  6、使用哪一个过滤器能够读取SCI数据的通信作者。 | | | | | |
| **二、实验操作部分（可续页，附上文字说明和重要内容截图）**  1、  2、  3、  4、  5、  6、 | | | | | |
| **三、实验效果分析**  **（一）实验完成情况**  **（二）实验室完成（未完成）的总结** | | | | | |
| **四、实验反思与反馈**  **（一）本次实验的心得体会**  **（二）本次实验的提升建议** | | | | | |
| **教师评语** |  | | | | |