**深 圳 大 学 实 验 报 告**

**课程名称：Java程序设计**

**实验项目名称： 选实验6 输入输出应用**

**学院： 计算机与软件学院**

**专业： 计算机科学与技术**

**指导教师：潘微科**

**报告人： 吴艇 学号： 2020281061 班级： 计科01**

**实验时间：2021年12月17日（周五）~2021年12月22日（周三）**

**实验报告提交时间： 2021年12月23日星期四**

**教务部制**

|  |
| --- |
| **实验目的与要求：**  **实验目的：** 掌握I/O程序设计，能够读写本地文件等。  **实验要求：**  (1). 数据解析和统计   * <https://snap.stanford.edu/data/web-Amazon.html> 网站上有很多Amazon的数据集供研究人员下载使用。 * 本次实验使用Watches.txt.gz数据集，请下载后解压。 * 格式说明请看网页上的“Data Format”部分。 * 在报告中附上程序截图、运行结果（如每个输出文件前10行的截图等）和简要文字说明。   （1.1）使用Java语言读取解压后的文件（Watches.txt），并得到以下文件（40分）：   * **review.txt**：每行3列，以分号作为分隔符，第1列是userID，第2列是productID，第3列是summary；表示(user, product, summary)三元组。该文件中不同行之间的顺序，按照userID从小到大排列，当userID相同时按照productID从小到大排列。注：如果没有按照userID或productID从小到大排序扣5分。   （1.2）使用Java语言根据review.txt进行计算，并得到以下文件（40分）：   * **userNeighborhood.txt**：每行11列，以分号作为分隔符，第1列是userID，第2-10列是与该用户最相似的10个用户的userID，按相似度值从大到小排列，其中相似度是通过review.txt中的前两列计算得到的Jaccard index值。该文件中不同行之间的顺序，按照第1列的userID从小到大排列。注：如果没有使用多线程实现扣5分，如果没有按相似度从大到小排列扣5分。   报告写作。要求：主要思路有明确的说明，重点代码有详细的注释，行文逻辑清晰可读性强，报告整体写作较为专业。（20分）  **说明：**  （1）本次实验课作业满分为100分，占总成绩的比例（待定）。  （2）本次实验课作业截至时间2021年12月22日（周三）21:59。  （3）报告正文：请在指定位置填写，本次实验**需要单独提交源程序文件**（源程序单独打包在Blackboard中上传，不要包含外部导入的包）。  （4）个人信息：WORD文件名中的“姓名”、“学号”，请改为你的姓名和学号；实验报告的首页，请准确填写“学院”、“专业”、“报告人”、“学号”、“班级”、“实验报告提交时间”等信息。  （5）提交方式：截至时间前，请在Blackboard平台中提交。  （6）发现抄袭（包括复制&粘贴整句话、整张图），**抄袭者和被抄袭者的成绩记零分（含抄袭往届同学的作业）。**  （7）延迟提交，不得分；如有特殊情况，请于截至日期之后的48小时内发邮件到panweike@szu.edu.cn，并在邮件中注明课程名称、作业名称、姓名、学号等信息，以及特殊情况的说明，我收到后会及时回复。  （8）期末考试阶段补交无效。 |

|  |
| --- |
| （1.1）使用Java语言读取解压后的文件（Watches.txt），并得到以下文件（40分）：   * **review.txt**：每行3列，以分号作为分隔符，第1列是userID，第2列是productID，第3列是summary；表示(user, product, summary)三元组。该文件中不同行之间的顺序按照userID从小到大排列，当userID相同时按照productID从小到大排列。注：如果没有按照userID或productID从小到大排序扣5分。   ①这里编写了三个方法来实现归并排序，以UserID为序。  ②字符串从前到后进行比较，字符串中字符按照ASKII码表比较。  ③读写文件，用对象数组来存储文件中的多组数据指标。  Review类：  将UserID单独拿出来进行排序，将排序结果用sortNum数组来反映。  以productId来排序。由于userID相同的比较少，故使用冒泡排序来进行比较排序。  ④输出结果，关闭文件。  文件写入结果：  此页截图中userID都为unknown，可以更直观地反映出根据productID排序。  （1.2）使用Java语言根据review.txt进行计算，并得到以下文件（40分）：   * **userNeighborhood.txt**：每行11列，以分号作为分隔符，第1列是userID，第2-10列是与该用户最相似的10个用户的userID，按相似度值从大到小排列，其中相似度是通过review.txt中的前两列计算得到的Jaccard index值。该文件中不同行之间的顺序，按照第1列的userID从小到大排列。注：如果没有使用多线程实现扣5分，如果没有按相似度从大到小排列扣5分。   ①需要进行计算比较的两种数据单独编写一个类  ②与第一题相同，使用归并排序  ③计算两个字符串相似度的方法  ④计算相似度前的准备工作  将userID和productID合并在一起，userID相同的话，将多个productID进行拼接  ⑤计算相似度，并且进行排序，将结果写入文件中  计算得到相似度  将相似度排序  以顺序数组sortNum为序写入文件  ⑥结果  ++++++++++++++++++++++++++++++++++++++++++++++++++++++  **其他（例如感想、建议等等）。** |

深圳大学学生实验报告用纸

|  |
| --- |
| 指导教师批阅意见：  成绩评定：  指导教师签字：  2021年 月 日 |
| 备注： |

注：1、报告内的项目或内容设置，可根据实际情况加以调整和补充。

2、教师批改学生实验报告时间应在学生提交实验报告时间后10日内。