**《数据结构》课程设计任务书**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 课程设计名称 | 《数据结构》课程设计 | | | | | | | |
| 适用专业 | 软件2011、2012、2013 | | | 培养层次 | | 本科 | | |
| 学期 | 第1学期 | 周数 | 2 | 学分 |  | | 总学时 | 2周 |
| **一、课程设计目的与要求**  **1.课程设计目的**  在数据结构理论学习和基础实验的基础上，开发规模较大的程序，掌握应用计算机解决实际问题的基本方法。学会综合应用数据结构理论知识解决问题，为其他专业课打下基础。  **2.课程设计要求**  （1）利用C、C++语言成系统的设计，使用多文件完成系统，要求：   * 做好模块设计，每个模块完成相应功能，不得所有功能都放一个函数内。 * 题目要求的数据必须来自教师提供文件数据，不得自己建。 * 减少全局变量使用。通过地址、变量实现函数间数据传递。 * 数据结构要丰富，不要单一结构，引入链表、map容器、树、图结构实现。 * 变量名、函数名、文件名等所有用户标识符命名要规范。代码规范参考资料如下：IMG_256https://www.cnblogs.com/zhrb/p/6270496.html * 建议用C++的文件读写函数，更稳定。 * 借助其他图形工具做结果的可视化展示。如python的图形库（matplotlib等）、c语言的easyx库、Neo4j数据库等。   （2）系统扩展性要好，设计函数尽量考虑可以给多个模块共享使用。程序功能变化，只需要对某个函数修改，而不是全部代码变化。  （3）系统交互性要好，对用户的每步操作都有文字提示，错误操作能有友好提示，功能操作简单易懂。  （4）系统容错性要好。系统异常情况或临界数据均有相应处理，程序不会异常退出。 | | | | | | | | |
| **二、选题说明**  学生自选课程设计题目，设计题目从任务书所列选题表中选取，每教学班每题不得超过6人。  **选题原则：PTA期中小测前30名只能选择第4或者第5题，31-60名限定在2-4题，61-92限定在1-3题。**  学生原则上可以结合个人爱好自选课题，课题要有一定的深度与难度，有一定的算法复杂性，能够巩固数据结构课程所学的知识。学生自选课题需在18周周二前报课程设计指导教师批准方可生效。 | | | | | | | | |
| **三、课程设计内容与要求**  **题目1、景区评论文本的热度调查**  **1622771633(1)**  **图1 景区评论数据格式**  **【功能】**  图1所示为“景区评论.csv”数据。现在要求统计50个景区的热度词。词的热度主要以词出现的频率表示。要求设计数据结构为：基于词汇的哈希链，然后基于哈希链统计每个词出现的频率。实现功能：  1. 以中文词作为关键字，设计每个文件的哈希链，并存入文件景区名“HashWord.txt”中。存入格式如下：（主要用分词那列的数据）  亲子->1->20->50->1245 //表示“亲子”这个词在编号为1,20,50,1245的评论出现。  票价->12->120->150->1245->2345  注意：共有50个景区，故最后需要有50个哈希链文件展示。  输入一个景区，展示对应的哈希链。  2. 输入一个景区，展示所有词的词频。可以用线性表保存。  3. 基于堆排序，统计每个景区词频排名前20的词。展示结果如下：  **1622773132(1)**  4. 基于合并排序，对50个景区的评论进行合并排序，并展示结果。注意，相同的词要合并计算。如“不错”，合并后应该是3000多。  系统菜单请自行设计，要体现上述功能。本题词汇量较大，请务必注意动态申请及链式结构，不能依赖顺序表、数组。  **【注意】**   1. 本题哈希链、排序算法不得用python语言写。 2. 如果有做**可视化**，可以借用python的图形库展示。如词云图、热度图。   **题目2、三国人物关系图谱**  基于三国人物关系数据，构建三国人物关系图。图的顶点为三国人物。边为人物关系（如儿子、义弟、主公）。本题需要实现在图结构上做的各种人物关系查询。人物的属性包括：人名、所属国。  如刘备（刘备，蜀国）  **【功能】**  1.基于邻接表构建三国人物关系图，并能输出展示邻接表，保存在文件”TRKGraph.txt”中。文件格式如下：  刘备-->张飞--->关羽--->赵云  注意：图的构建，应该先整理图中的顶点数据，即要整理所有三国人物数据作为图顶点。  2.统计人物关系数量最多的前10个三国人物及其所属国。  3.输入2个人物，查找2个人物之间所有关系。基于图的深度遍历，展示关系路径。并能说明最短路径是哪条。答辩时候展示：有关系的人物。无关系的人物。  4.**智能问答**。输入荀彧的主公的儿子是谁？能输出关系路径，并给出答案。  **题目3、北京地铁关系网**  基于北京21条地铁路线建一个地铁关系图。图节点信息在station.csv中。节点关系在line1,line2....lineyz中的文件。无向图。2个站的距离值为边的权值。注意：1个站点属于多条线。  **【功能】**   1. 基于邻接表结构，构建地铁关系图，存入文件subway.csv。文件关系如下：   苹果园->古城->  本功能可以结合python的networkx库输出展示，不同地铁线用不同颜色输出，检查与实际路线图是否一致。   1. 输入某条线，展示该线路经过站点。 2. 输入某个站点，展示经过该站点的所有路线，结合图的广度遍历。 3. 输入2个站点，输出经过2个站点的所有路径及应乘坐的地铁线，并输出最短路径。   （所有路径结合图的深度遍历、最短路径结合Dijkstra算法），建议保存在文件，结合python做可视化输出，也可以直接采用python和C++混合编程实现可视化输出。   1. 输入2个站点，输出最佳换乘方案，即不考虑路径长度，考虑站点最少的换乘方法。   （结合图的广度遍历）  【参考】  1.[北京市地铁图 - 百度地图 (baidu.com)](https://map.baidu.com/@13151828,2796584,13z)  **题目4、字典树**  大家熟悉的英文分词（python的jieba分词库），如果只是用空格作为单词的标志，这是不合理。因为单词可能书写不对。所以合理的自然语言处理（NLP）分词，需要一个有效字典维护。现有的英文字典包含单词量高达10万多条，在这样一个庞大的字典中查找数据，基于线性查找，效率非常低。现在基于单词有很多共同字符（共同前缀字母）特点，构造一棵字典树，可大大提高单词的查询效率。并能快速对单词排序、词频统计、单词拼写检查及全文搜索。  本次课设只要实现基于英文文章建一棵字典树。并能统计相关单词出现的词频。  英文字典树的结构图是这样的。按照树型结构存储字符串，每个结点存一个字符，自顶向下做标记的就是词的词尾，比如，app,apple,application,abstract,absorb,block,black,blake... 等等  IMG_256  **英文字典树的结点数据结构：**  1.词频 int型变量记录词频  2.flag标记，当前结点是否构成单词  3.结点值  4.孩子兄弟链结构  **【功能】**  字典树要求用孩子兄弟链或孩子链数组构建。   1. 按照指定15篇英文文章，展示相应的字典树。字典树的展示可以模仿目录树展示，以缩进方式层次展示目录树。   字典树构造过程中，如果发现ab？Cd，这样不合法字符，能提示不合法单词，并舍弃。   1. 输入一个字符串，能插入到字典树中，并展示。 2. 单词查询。能实现模糊查询。如输入字母a，ab，则能展示a或ab为前缀的所有单词。基于树的先序遍历展示。 3. 统计每个单词的词频，并能按照字母顺序展示所有单词的词频。   【扩展功能（**可选，加分项**）】   1. 基于英文字典words.txt（10万数据）建一个英文所有单词的字典树。 2. 对给定的15篇英文文章做英文拼写检查。如果有单词不在字典树，为不合法单词，程序能   指出不合理的英文单词，并给出正确的单词写法建议。   1. 输入一份英文文章，能在字典树查询单词，并统计单词的出现次数。并输出每个单词的词   频。  4.全文搜索。输入单词的几个字母，则能输出相应单词提示。  **题目5、豆瓣电影数据的知识图谱构建**  **【功能】**  一、基于豆瓣TOP250电影数据构建如下电影图结构，并能基于该图结构，做相关的路径搜索及推荐。  IMG_256  1.图的顶点数据有：  （1）电影结点（电影ID，电影名称，标签）  （2）导演结点（导演ID，导演名称，标签）  （3）演员结点（演员ID，演员名称，标签）  （4）类型结点（类型ID，类型名称，标签）  其中ID编号为代码自动生成。可以从元数据文件提取这些数据，单独存一个文本文件node.csv表示。标签为可选数据，一般自己定义颜色。可以在可视化展示。  2.边：顶点的关系。主要有如下：  （1）导演——电影关系（导演）  （2）演员——电影关系（出演）  （3）导演——演员关系（合作）  （4）电影——类型关系（属于）  这些关系，需要写代码自动从元数据文件提取，最后保存成movie\_relation.csv文件。格式如下：  肖申克的救赎 弗兰克·德拉邦特 导演  大话西游之大圣娶亲 周星驰 出演  大话西游之大圣娶亲 吴孟达 出演  周星驰 刘镇伟 合作  无间道2 动作 属于  **【注意】**  以上数据都要基于原数据自动抽取，如果用C语言处理比较复杂，建议学习python的语言的pandas库、numpy库处理，结合python语言的列表、字典、集合等数据结构操作。  **【建图要求】**  1. 能够基于原始数据，自动抽取图的顶点（实体）及边（实体关系）  2. 基于movie\_relation.csv数据，选择邻接表或邻接矩阵构建电影的知识图谱。  3. 基于Python的的networkx库展示输出  4. 由于顶点及边较多，可以选择展示部分数据。   1. **图路径搜索**   1. 输入某个演员、导演。能查询其参与作品、导演作品。  2. 查询关系。输入2个顶点，能查询其关系。如周星驰，无间道2，查询2个顶点路径关系。如果不存在关系，提示无。如果有多个关系，能展示最短路径，需要可视化展示。  3. 智能问答。输入无间道的导演是谁？基于图的搜索后，能给出答案。   1. **电影推荐设计（可选，加分项）**   输入某个电影。基于广度遍历，查询该电影和其他电影的路径数量，取数量最多的前10电影推荐，并能展示推荐路径。本题最好保存电影之间的所有路径，并保存文件。推荐路径如图所示展示：  电影“春光乍泄”和“重庆森林”共有3条共同路径。推荐理由：2部电影属于“爱情”类型、都是“王家卫”导演、都有演员“梁朝伟“参演。  **推荐路径**  **【数据要求】** | | | | | | | | |
| **四、考核内容和方式**  **答辩提交材料：**  1. 课程设计报告（打印稿）。**答辩当天交。**  2. 课程设计报告（电子版）+源程序（运行无误，电子版）。**提交课堂派。**  **考核方式：**答辩。  **答辩内容：**  1. 演示程序。程序稳定运行。  2. 答辩能讲调试中碰到问题，能回答教师提的问题，能讲清楚程序设计思路。  **注意事项：**  答辩中出现以下情况视为不及格：  1. 课设报告查重率超过60%。  2. 答辩当天不会回答问题，设计思路不清，碰到问题不会解决，**不及格**。 | | | | | | | | |
| **五、成绩评定标准**  1. 优：按要求完成题目，完成题目要求功能，软件可靠性好，功能有适当扩展。 答辩回答问题正确，对系统的演示流畅，源代码解释清晰。课设报告清晰直观。  2. 良：完成设计题目，完成题目要求功能，系统容错较好。答辩回答问题较好，对系统的演示较流畅，源代码解释较清晰。课设报告清晰直观。  3. 中：完成设计题目，完成题目要求基本功能，实现部分系统容错。答辩回答问题较好，对系统的演示较流畅，源代码解释较清晰。课设报告完成较好。  4. 及格：基本完成程序功能，系统可靠性较差，数据验证较少。答辩问题能正确回答，系统能正确演示，课设报告符合要求。  5. 不及格：答辩不会回答问题，代码判定不是自己完成，没有完成课设报告。 | | | | | | | | |

任选教师（课程负责人）签名： 教研室主任签名: 学院审批：

日期： 日期： 日期：