暑期学校实验项目：高考志愿填报助手

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 小组名称 | 知识图谱构建B组 | | | | | | |
| 姓 名 | 何洋龙 | 专业 | 人工智能 | 班级 | 1班 | 学号 | 09118143 |
| 实验时间 | 2020.8.31-2020.9.23 | | 指导教师 | 孔祥龙 | | 成绩 |  |
| **一、实验背景和目的**  本项目的目的是解决高考生在高考志愿填报时遇到的“我能上哪个学校？”，“我能上哪个专业？”的问题。能在网页端根据用户输入的省份，高考成绩等信息给用户智能推荐若干所“冲”“保”“稳”学校，并在此基础上提供更多基于人工智能算法的智能推荐。 | | | | | | | |
| **二、小组任务和个人任务**  **小组任务：**  后端：构建与省份部分相关的知识图谱，省份包含的属性有：位于该省的高校，该省三年的文理分数线等。  前端：将后端建立的知识图谱内容展示出来，实现知识图谱的可视化。  **个人任务：**  参与后端开发任务，利用其他小组提供的数据和小组成员确定的数据库模式建立Neo4j图数据库。 | | | | | | | |
| **三、个人任务需求分析**   1. 学习知识图谱的相关概念，了解知识图谱的基本结构。 2. 确定知识图谱中的实体与属性有哪些。 3. 学习Neo4j图数据库的使用方法，学习py2neo库的使用方法。 4. 设计图数据库模式，确定有哪些实体，实体有哪些属性，实体与实体之间有哪些关系。 5. 利用.json和.csv文件的数据建立图数据库中的节点和边。 | | | | | | | |
| **四、实验过程（需附上关键代码及相关说明）**  1. 安装Neo4j图数据库  2. 引用py2neo库，在Python中操控Neo4j数据库  3. 链接到图数据库    （其中password由服务器端设置）  4. 构建建知识图谱中的节点和关系  4.1 建立省份节点    4.2 建立学校节点    4.3 建立分数线节点    4.4 建立关系 | | | | | | | |
| **五、实验结果与分析**  图谱可视化结果（以江苏为例）    可以看到，我们构建的知识图谱有3种节点：省份，学校，分数信息以及2种关系：学校与省份之间的从属关系，分数信息节点与省份之间的从属关系。  学校节点的属性有：所在省份的ID，学校ID，学校名称，是否为985高校，是否为211高校，是否为双一流高校。  省份节点的属性有：省份ID，省份名称。  分数节点的属性有：年份，省份ID，文理科，分数线类型，分数线，各段分数，累计人数。  由于各段分数数据量过大，未在图谱可视化中展示。  从可视化结果中可以看出节点与节点之间的关系建立正常，节点信息正确。 | | | | | | | | |
| **六、实验总结与心得体会**  此次实验我们组构建了与省份信息相关的知识图谱，并实现了图谱的可视化，我作为后端小组成员主要负责的是知识图谱的构建部分。在实验过程中我学会了很多新知识也遇到了许多意料之外的问题，总结如下：   1. 学习了知识图谱相关的概念，以及学会了如何使用Neo4j图数据库构建知识图谱，学会了如何使用py2neo库利用Python控制图数据库。 2. 在最初构建知识图谱的时候由于图数据库模式设计不当，导致后期有过更改，重新构建知识图谱，在一定程度上影响了开发进度，这让我意识到良好的系统设计的重要性。 3. 在构建知识图谱时由于数据格式问题和数据缺失问题导致知识图谱无法完全构建，等到负责数据清洗的小组把符合要求的数据提供给我们才能构建出整个知识图谱，这让我深刻认识数据清洗工作的重要性。 | | | | | | | | |

2020年9月制