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 | 杜熙源 | 专业 | 计算机科学与技术（人工智能） | 班级 | 091181 | 学号 | 09118144 |
| 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**  高考是大部分中国家庭必须面对的难题，除去考生本人成绩的优劣，志愿填报对考生最终考学结果也有着很大的影响，由此衍生出了辅助报考的产业。在这个人工智能和大数据广为应用的时代，人工智能在高考报考上发挥着重要的作用。  本项目旨意于利用人工智能技术和网站设计结合，为用户提供专业、可靠的报考助手网站，助力广大学子进入心仪的学府。 | | | | | | | |
| **二、小组任务和个人任务**  本小组负责构建有关省份分布的知识图谱，其内容应该包括省份内的学校分布、一本线、二本线、专科线、一分一档表等信息，图谱构建完成后，应能完成对应的可视化，使信息更加清晰明了。  本组将工作分为前端和后端，前端负责html页面设计和知识图谱在html页面上的可视化工作，后端负责将整理好的数据导入neo4j数据库中。  本人负责后端的部分工作，需要将整理好的数据导入neo4j中。 | | | | | | | |
| **三、个人任务需求分析**  本人需要将1组提供的json文件导入neo4j中，通过使用py2neo库来操作文件，生成图的点和边。  当时处理数据时，已有的数据为json格式，需要将json读取为字典格式并将其数据导入图数据库中。  由于1组提供了两种格式（json，csv）的数据，为了充分利用这两种格式的数据，组内安排了不同的人处理这两种数据。  本人负责处理json文件的数据。 | | | | | | | |
| **四、实验过程（需附上关键代码及相关说明）**  所需要的库如下：    其中，py2neo是操作neo4j的库  获取json文件夹的地址，对接neo4j数据库，清空其内部已构造的关系：    构建省份节点：    由于1组提供的数据部分存在问题，所以这里使用try和except包装，以跳过错误文件。    上述代码为构建关系的代码，以列表存储从json文件中读取的数据，其中包括获取图数据库中的省份节点，构建省份、年份、专业、排名等关系，鉴于1组提供的数据存在的问题，本处仍然使用try和except进行包装。 | | | | | | | |
| **五、实验结果与分析**  实验的结果图片示例如下：    很不幸的是，1组提供的数据有很多存在问题，在第二次更新前，存在问题的数据如下：  2017吉林文科.json, 2017吉林理科.json,2017山东理科.json, 2017广西文科.json  2017广西理科.json, 2018山东文科（后半段人数有引号）.json, 2018青海文科.json  2019河北理科.json, 2017上海all.json  图中限制了节点的数量为125个，可以看到显示已经非常密集，在后续的可视化工作中，可能需要将分数和排名节点设置为不可见。  不幸的是，后续的数据更新并没有更新json格式的数据文件，所以本组最后使用了csv文件进行图数据库的构建。 | | | | | | | | |
| **六、实验总结与心得体会**  本次实验利用了neo4j工具构建图数据库，使用py2neo库对neo4j进行操作，在本次实验中复习了json文件的读写，首次尝试了图数据库的构建，对neo4j进行了简答的操作，是一次非常珍贵的体验。  可惜的是，1组的错误json数据后续并没有更新，后续只能放弃我这一部分的代码，这一点还是非常可惜的。  本次项目设计开始时，组内并不清楚知识图谱是如何构建的，对此进行了很长时间的讨论，在有了基本的印象后，又进行了很长时间关于如何构建图数据库的讨论。  在最终和其他组进行交流，反复的思考讨论后，最终选择了neo4j库，为了充分利用数据，准备了两种代码读取不同的文件构建图数据库。  可惜的是，我所完成的读取json格式数据的代码由于多种原因最后没有被使用，这也令我十分遗憾。  总之，这次实验让我增强了和他人交流合作的能力，对以后学习工作上的合作会有很大的帮助。 | | | | | | | | |

2020年9月制