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A组 | | | | | | |
| 姓 名 | 王靖婷 | 专业 | 人工智能 | 班级 | 091181 | 学号 | 09118140 |
| 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**  实现一个基于web的高考志愿填报系统，旨在为广大考生提供各高校近几年的录取信息并基于成绩排名、兴趣爱好等为其推荐合适的学校。 | | | | | | | |
| **二、小组任务和个人任务**  小组任务：建立以高校为主体的知识图谱，并完成相应的页面设计。任务可细分为  1）对专业名称进行消歧处理，即将每个学校招生的专业归类到所属的一级学科；  2）知识图谱的设计；  3）数据处理，提取实体与关系并生成编号；  4）创建可以导入Neo4j的csv文件，生成数据库；  5）前端页面设计。  个人任务：参与知识图谱的设计，进一步修改预处理后的数据，并将其导入neo4j中。 | | | | | | | |
| **三、个人任务需求分析**  1、在设计阶段，主要工作就是确定实体和各实体间的关系。既然是建立一个以高校为主体的知识图谱，那么实体必然包括学校、专业、省份，关系必然包括学校与省份间（学校位于哪个省份）、学校与专业间（学校开设哪些专业）的关系。此外，增加一级学科的实体，对应地增加专业与一级学科之间的从属关系（专业属于哪个一级学科）。而对于录取分数，实体与关系的设立就不是非常清晰明了。由于专业、分数、省份、年份之间不是简单的一对多关系，而是混杂的多对多关系，分数又由专业、省份、年份、科类共同决定，且各个高校专业名称叫法五花八门，如何理清它们之间的关系，是设计过程中的一大难点。  2、在数据处理阶段，根据Neo4j官网给出的例子，实体类的表头应包括ID，:LABEL，关系类的表头应包括START\_ID，END\_ID和:TYPE三个域，此外可根据需要增加必要的属性如name等。我们需要做的是在预处理数据的基础上进一步加工，规范表头，增加label，必要时可手动修改一些细小的错误，从而创建可以直接导入图数据库的csv文件。  3、在导入阶段，参考Neo4j官网文档，利用neo4j-admin import命令将csv文件导入自带的图数据库中。 | | | | | | | |
| **四、实验过程（需附上关键代码及相关说明）**  上一小组的同学初步设计了知识图谱并进行了初步的数据处理，按照他们的想法，我画出了下图：    上一组对各实体进行了编号并抽取出了学校与省份、学校与专业、专业与一级学科之间的关系。但我们发现这样设计稍有欠妥，如major表中已有的实体做了专业的属性，subject和ranking与其他实体都毫无关系，在此基础上，我们组做了进一步的改进。经组内成员讨论商定，设计了如下图谱：    实体包括   * College（CollegeID:ID, CollegeName, 985:int, 211:int, Top:int, :LABEL） //学校 * First-Level Discipline（DisciplineID:ID, DisciplineName, :LABEL） //一级学科 * Major（MajorID:ID, MajorName, Contributor, :LABEL） //专业 * Year（YearID:ID, YearName, :LABEL） //年份，2017/2018/2019 * Subject（SubjectID:ID, SubjectName, :LABEL） //科类，文科/理科/全科 * Province（ProvinceID:ID, ProvinceName, :LABEL） //省份   关系包括   * Located\_in（:START\_ID, :END\_ID, :TYPE）//学校->省份，学校位于哪个省份 * Established\_in（:START\_ID, :END\_ID, :TYPE）//专业->省份，专业在哪个省份招生 * Belong\_to（:START\_ID, :END\_ID, :TYPE）//专业->一级学科，专业属于哪个一级学科 * Is\_subject（:START\_ID, :END\_ID, :TYPE）//专业->科类，专业招收文科生还是理科生 * Has（:START\_ID, :END\_ID, :TYPE）//学校->专业，学校开设哪些专业 * Need\_score（:START\_ID, Score:int, :END\_ID, :TYPE）//专业->年份，该专业在这一年的录取分数   为了在图谱中显示分数信息，我们增加了年份实体，将分数作为专业与年份之间关系的一个属性，表明该专业在该年份的录取分数。在与上一组同学沟通交流后，他们做了一次修改，我们在这次改动后的数据基础上实现后续操作。具体的操作包括修改表头，增加标签，规范数据，创建可以直接导入Neo4j的csv文件等。以下是分配给我的任务。  1、抽取出专业与省份的established\_in关系  原major表如下，需提取出第一列与第三列，并增加:TYPE列。      2、修改major表的格式  修改表头，增加:LABEL列    其他表的修改由我们组其他成员完成，我也写了两个处理的函数供于参考。注意写入文件要用utf-8编码，否则导入图谱后会乱码。    3、整合组员处理后的csv文件将其导入Neo4j中  根据官网给出的文档，使用neo4j-admin import命令将所有的结点与关系导入neo4j数据库中。  第一次导入时由于数据出现错误而失败，根据提示是由于has表中collegeID出现如下图所示错误，  修改后进行第二次导入又出现错误，此次错误如下图所示，established\_in中出现了不存在的END\_ID，经排查是在前期的数据处理过程中丢掉了省份p36全国联招的结点。    加上丢失的节点后，第三次终于导入成功！ | | | | | | | |
| **五、实验结果与分析**  按以上设计，构建好的知识图谱如下：    从左边可以看到该数据库的信息概览。当我们点击College标签，进入页面可以看到，每个结点代表一所学校，双击该结点可以看到这个学校所在的省份及包含的专业。      当我们双击某个专业的结点后，会显示出该专业从属的一级学科、科类、招生的省份以及对应年份的分数信息。    如果从省份出发，双击某个省份结点，会显示位于该省份的高校和其他学校在该省份招收的专业。    综上，我们利用neo4j图数据库基本完成了以高校为主体的知识图谱的构建，用户可以在页面上直观地找到高校的一些信息。但我们成品仍存在着缺陷与不足。  1、由于专业的名称叫法五花八门，本质上相同的专业各个学校都有不同的叫法，相同的专业每年招生计划也有变动，可能2017是按专业招生，2019年则变成大类招生，目前我们只做到把它们归类到一级学科的层面，没有更细地划分到二级学科。  2、同样是由于上述问题，在对专业进行编号时，我们对第一组整合的数据表中每一条记录都进行了编号。也就是说，我们把本该是多对多的关系拆分成了多个一对多的关系，例如，专业与省份之间应该是多对多的关系，一个专业在多个省份招生，一个省份的学生也可以报考多个专业，我们的做法是重复专业，使一条专业实体的数据与只与一个省份关联，这样就实现了省份与专业的一对多关系。但这样做就造成了专业数据的大量冗余，图谱上会出现许多名称相同的专业结点，视觉上的观感不好，用户体验也不友好，缺少了点人性化。  针对上述问题，我们组也提供了相应的解决方法。  1、既然专业编号已经冗余，索性将省份、科类、年份、分数当作一个大的实体与某个专业结点相连。这样就省去了专业与四者分别关联。  2、改变专业的编号，实现多对多关系。  3、由于要将分数信息体现在里面，而分数是由专业、科类、省份、年份共同决定，经上述操作，每一个专业都对应唯一的省份和科类，在此基础上我们增加了年份实体，将分数作为专业与年份关系的一个属性。还有一种做法是将分数作为实体，将年份作为专业与分数关系的一个属性。 | | | | | | | | |
| **六、实验总结与心得体会**  1、数据处理。数据的清洗与处理关乎后面各组任务的进行。如果第一步出了问题，会连锁影响后面工作的推进。所以扣好第一颗扣子很重要。  2、协作交流，团队合作。100+人共同完成一个项目，需要组与组之间，小组内部充分的交流协作。组与组之间的工作会有重叠、衔接的部分，这时候就需要组长来协调与其他组的工作，并将任务下派给各组员，同时，组长也要整合各个组员的工作汇总到班级项目中。  3、通过本次课程，我对Neo4j图知识库有了深入的了解，为以后知识图谱的运用与实践打下了一定的基础。同时我也深刻地体会到设计知识图谱的重要性，比起数据处理，如何设计好图谱才是我们组的关键任务。由于前期配合交流的不够充分，导致图谱的设计环节出现了差错，历经了好几次的更迭修改。吃一堑长一智，下次就会避免犯类似的错误。 | | | | | | | | |

2020年9月制