暑期学校实验项目：高考志愿填报助手

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 小组名称 | 知识推理与数据可视化 | | | | | | |
| 姓 名 | 李英泰 | 专业 | 计算机科学与技术（人工智能） | 班级 | 091181 | 学号 | 09118128 |
| 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**  实验背景：   1. 在大二的上学期，我们已经学习了Python语言的基本使用。Python作为目前人工智能领域使用最广泛的语言，有必要得到很好的掌握。而这种掌握最好的方式就是通过练习。 2. 在学习过数据库概论之后，我们对结合数据库进行网页开发已经有了一定的了解。部分同学已经有过使用Django框架进行开发的经验。整体上具备完成此次项目开发的背景知识和相应能力。 3. 知识表示与推理作为人工智能专业的一门专业主干课，将在下一学期由漆桂林老师开设。课程的整体内容基本完全围绕知识图谱。在本次实践课当中融入构建图谱、使用图谱进行推理等内容，有利于获得对知识图谱的感性认识和基本了解，从人工智能专业的培养方案来看是十分合理的。 4. 在原定的方案中，此次实验将在高考志愿填报之前完成。因此安排开发高考志愿填报助手这样的工具，其效果能够得到及时有效的反馈。   目的：   1. 增强对于 Python使用的掌握。 2. 掌握基于Django框架进行web开发的技术。 3. 了解人工智能算法和知识图谱的应用 | | | | | | | |
| **二、小组任务和个人任务**  小组任务： 1、依据高校分数线进行统计分析，生成相应图表，展示变化趋势； 2、根据用户输入的动态个人信息进行相应的可视化； 3、以地图为依据生成可视化界面，展示高校信息； 4、基于知识图谱的推理和问答； 5、对应的django框架编写。  个人任务：  和小组内的廖涛同学一起，构建一个基于知识图谱的自然语言问答系统。该系统需要能够正确理解使用者通过输入文字所表达的问题意图，将回答问题所需要的查询的内容转化成恰当的cypher查询语句，在neo4j当中查询到所需的数据。最终将这些数据转化为自然语言语言的形式，以对话的方式返回给用户。 | | | | | | | |
| **三、个人任务需求分析**  在参考图谱组和小组内陈震寰等同学的总结之后，我们认为我们的自然语言问答系统需要回答以下问题：  XX省的一本线/二本线是多少  XX年XX分在XX省能排多少名  XX学校的哪些专业较好  XX专业有哪些学校较好  XX学校在哪个省份  XX省份有哪些学校  XX省考生考了XX分能不能上XX学校  XX学校某专业在XX省需要多少分  XX学校是不是985  XX学校是不是211  XX学校是不是双一流  以上我们设定回答的问题中，并非所有问题都具有实际意义。这主要是由于以下两个原因造成的：   1. 我和廖涛同学缺乏在图谱中进行复杂推理的相关技术，不能保证我们所开发出的系统 对一些复杂的问题作出正确的、甚至是有意义的回答。 2. 现有的数据相对较少，数据之间的关系也较为简单。本身不支持进行复杂的推理，也不能够对很多问题进行回答。事实上，从kgB组预留的查询信息的接口函数来看，我们能够在图数据库中进行的查找非常有限。我们也和kgA组沟通过查找相关的内容，但目前（截止写课程报告的时间）还没有得到相应的访问neo4j的方式。   对于同样的问题，不管使用者使用什么样的问句进行询问，我们都应该能够正确的判断提问者的真实意图。这需要我们对一个问题的各种表述方式进行总结。（因为我们所采用的是不基于人工智能模型的、简单的模板匹配的方式）模板覆盖的全面性对问答的性能起到了决定性的作用。  对于我们的工作来说，除了需求分析，另一点很重要的是对于我们能做什么的分析。从图谱组预留的访问图谱的接口来看，我们能够从图谱中获得以下信息：                因此我们实际上能且仅能回答的问题有：  XX省有哪些大学  XX省的一本线/二本线是多少  XX年XX分在XX省能排多少名  尽管如此，我们仍然为回答其他问题的需求编写了相关的代码。当然，这些问题并不能得到真正有效的回复，这是比较遗憾的。或许以后的某个时候，在获取更多的数据，对数据之间的关系有了更加丰富的描述之后，我们能够把这个问答系统变成一个真正实用、有用的系统。  利用知识图谱进行因果推理也是一个常见的应用。如果有更多的数据可用，比如一些新闻之类的事件数据，我们或许可以通过事件抽取等方式，结构化的推理出某一年分数线变化的原因。例如今年的新冠疫情之后，武汉大学和华中科技大学的分数线上升。通过这样的方式，我们甚至有可能预测一所学校之后的分数线变化当然未必准确，（但是即使是那些“专业”的志愿填报机构，难道就能做到准确了吗？）但对考生应该是有参考意义的。长学期漆桂林老师主讲的知识表示与推理课程有16课时的实践时间，倘若有合适的问题和机会，我将选择把短学期的工作做得更完善一些。 | | | | | | | |
| **四、实验过程（需附上关键代码及相关说明）**  1. 首先建立一个对问题意图的分类器。  1.1建立和问题相关的关键词词库    1.2建立和问题意图相关的疑问词库    1.3结合用户输入的问题，进行关键词匹配    1.4通过结合问题当中的关键词和疑问词，为问题添加question\_type的标签。    2.根据问题的意图，匹配关键字，生成对应的cypher查询语句。      3.根据从图谱中查询到的内容，结合回复模板，生成自然语言回复。  3.1根据2中生成的cypher，在neo4j中完成相应的查询    3.2使用预先设定的模板，将查询结果组织成自然语言的形式 | | | | | | | |
| **五、实验结果与分析**  因为事实上能做的东西非常有限，这里只放一些“成功”的实验结果，没有展示不能回答的问题。（事实上和实际场景中的需求相比，不能回答的问题占多数） | | | | | | | |
| **六、实验总结与心得体会**  实验总结：  本次实验中我们组的工作是知识推理和图谱可视化。一开始选择参与这个方向的原因主要是希望通过这次实践课能够进行一些知识推理方面的实践，但后来发现根据现有的数据其实做不了什么推理，只能做查询。  孔老师曾经在第一节课上说过我们组的自由度是比较高的，但我的实际体验却与之相反，由于许多工作都需要建立在其他组工作的基础上，实际上许多方面都是受限的。  就我和廖涛同学所负责的这部分来说，因为问答的答案需要在数据库中有相关的数据才能回答，我们能够满足的需求实际上取决于所掌握的数据和所建立的知识图谱的设计。但是出于各种原因，我们在中段之后才参与到图谱组对图谱的设计当中，有些概念设计已经完成了，不便于修改。因此最终呈现的结果并不是非常有利于进行问答的设计。这是我们工作的一个比较大的失误。  因为我在上个长学期徐立臻老师布置的数据库大作业上已经使用Django结合MySQL开发过一些东西，所以这门课程在这方面给我的新收获其实是在UI设计方面。虽然我并没有做相关的工作，但仍然收获了一些了解。孔老师在课上讲到的一些关于网络的知识也是额外的收获，尽管现在才结束一周就忘得差不多了。  在这次实践课中我和廖涛同学一起完成了一个自然语言的问答系统。我们所完成的问答是完全基于模板的。采用这种方式的主要原因是基于模板的方式比较简单，而且能够以较高的准确率回答问题。但缺点是对我们对模板定义的质量依赖很大，有很多的问题涵盖不到。在初期我们曾考虑过使用一些自然语言处理方面的神经网络模型来完成这项工作，但由于图谱建成已经是9月底了，最终放弃了这一点。尽管如此，这项开发仍然给我带来了很多的收获。  心得体会：  我认为我在这门课程中最大的收获其实在于那些大家做的不是很好的地方。比如数据组——>数据库组——>图谱组——>知识推理组，几乎每个环节发现了对数据的新需求的反馈都要延迟1天。这样每次新需求都要一周才能得到解决。从某种程度上来说，正是这样的体验引发我对于什么样的组织形式是成功的这个问题的思考。从更长的时间尺度上，这些也许比具体的代码练习或许能给我带来更多的收获。  就像在同时进行的另一门实践课上王贝伦老师所说的：那些机器学习的知识都会被忘掉，presentation的能力确实伴随一生的东西。  关于开发过程中对于图谱的使用，其实还有很多的不够好的地方。比如我们对于图谱的利用还处于比较浅层的阶段。如果融合一些目前知识图谱中常用的技术：如事件抽取等，我们或许可以对某年某所学校分数线的变化进行因果推理，或许能取得更好的效果，甚至可以用来预测之后的发展趋势。这样我们的知识推理才能算的上是知识推理，才是真正有实用价值的。 | | | | | | | |
|  | | | | | | | |

2020年9月制