暑期学校实验项目：高考志愿填报助手

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 小组名称 | AI 算法应用 C 组 | | | | | | |
| 姓 名 | 闻浩 | 专业 | 人工智能专业 | 班级 | 091181 | 学号 | 09118121 |
| 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**  在新冠疫情影响下，经党中央、国务院同意，2020年全国普通高等学校招生统一考试（即高考）延期一个月举行，考试时间为7月7日至8日。而软件实践的课程原定于暑期前进行，因此在孔祥龙老师和同学们的商量下，决定开展对高考志愿填报助手的系统开发，并期望在应届高考生完成考试后，对系统进行测试。  在2020上半年春季课程中，我们完成了对一百多所高校的专业录取分数以及排名等相关数据进行了收集与整合，以便于在短学期的课程中进行使用。 | | | | | | | |
| **二、小组任务和个人任务**  小组任务：  **语义消歧**：确定各专业的语义表示，统一不同高校对于同一个专业的不同称呼，便于横向比较、提高数据利用率以及大类的推荐。列出所有高校的所有专业名称，找到与每个专业最接近的一个或多个一级学科，建立对照表。  **基于总成绩的专业推荐**：在不考虑特别爱好的前提下，成绩是学生选择专业的重要依据。因此，在给使用者推荐学之后，还需要依据成绩给他们推荐合适的专业。不同的高校录取分数线不同，同一高校不同专业的录取分数线也不同。数据量不足。可以根据算法A组给出的学校和用户输入的成绩与排名，给出该学校各专业的推荐度与风险值。  **基于能力的专业推荐**：如何评估用户的能力，以及根据其能力推荐专业，是十分关键的问题。这里我们参考了孔老师提供的数据，以下是其中一部分。在孔老师提供数据的基础上，我们引入了学群这一概念。最终总计有18个学群，19项能力。  个人任务：  **语义消歧**：确定各专业的语义表示，统一不同高校对于同一个专业的不同称呼，便于横向比较、提高数据利用率以及大类的推荐。列出所有高校的所有专业名称，找到与每个专业最接近的一个或多个一级学科，建立对照表。  **连接数据库**：根据第十组同学构建的数据库，编写利用数据库读取数据的接口，生成对应的数据格式，并将数据运用到本组的代码上进行运行。 | | | | | | | |
| **三、个人任务需求分析**  **语义消岐**：  确定各专业的语义表示，统一不同高校对于同一个专业的不同称呼，便于横向比较、提高数据利用率以及大类的推荐。  已有数据：  高校录取分数线整合（省份名字统一）.csv    一级学科改进.csv    二级学科改进.csv    实验过程：  1.在高校录取分数线整合表中获取所有高校所有专业的名称  2.在二级学科改进表中获取所有学科门类、一级学科和二级学科名称  3.使用Jaccard距离计算专业名称与学科名称的相似度，找出每个专业最接近的学科名称  4.人工检查并纠错  5.补充自定义的一级学科  6.将学科名称统一为一个或多个一级学科  7.删除特殊的专业名称  8.人工检查并纠错  **连接数据库：**  已有数据：  由第十组的同学构建的db.sqlite3（使用SQLiteStudio进行可视化）    本组的其他代码需要使用两个文件，一个为rank.csv，即高校专业的排名等数据，另一个为result.csv，即所有高校专业语义消岐的映射列表。因此，需要从数据库中生成这两张表以便后续的专业推荐。 | | | | | | | |
| **四、实验过程（需附上关键代码及相关说明）**  语义消岐：  将学科门类和二级学科统一为一个或多个一级学科，多个一级学科用’;’隔开    添加一级学科对应的ID：    删除特殊的专业名称：    人工检查并纠错：      **连接数据库：**  定义得到文件的函数，以便于后续添加到主程序中，可以不影响其他代码，使得主程序的更改得以最小化。使用sqlite3对数据库进行连接。    编写sql语句，对数据库中的数据进行连接，形成原文件格式的属性名，并存入dataframe类型中。    按照原文件中的属性名，定义并生成文件，    最终形成了从数据库得出的文件格式，可以应用到后续的专业推荐代码中。 | | | | | | | |
| **五、实验结果与分析**  专业与一级学科映射：    分析：缺点为需要人工的消岐检查，否则映射关系过于简单，从而语义上有一定的偏差。  解决方案：编写端到端无需人工消岐而效果优秀的语义消岐程序，使得新增的数据可以得到一级学科的映射。  数据库连接：  生成与原有rank.csv和result.csv属性名相同的文件，且可以运行后续的代码而不依靠原来的外部文件。    分析：尝试其中的部分高校，会有错误，原因为第十组的同学后续更新了一些新增高校的数据，但这一部分的数据专业名称未经过消岐处理，因此没有这一部分的专业映射到一级学科的数据，从而造成数据的缺失。  解决方案：编写端到端无需人工消岐而效果优秀的语义消岐程序，使得新增的数据可以得到一级学科的映射。 | | | | | | | |
| **六、实验总结与心得体会**  本次软件实践的课程设计，让我初次体验到了共同合作开发一个系统是怎样的。从上学期开始收集数据，大家都被分配了一些目标的院系，利用爬虫技术在各类高校的官网上进行收集数据，后又对数据进行了格式上的处理。再到这个学期分组进行开发。我加入的是AI算法组，主要负责对目标高校进行专业的推荐。我们这一组是在前一AI算法组推荐的高校的基础上进行的，因此主要考虑对不同的专业进行推荐度的评级。我们首先列举了很多影响推荐度指标的数据，再逐一的进行筛选哪一些是合适并易于得到的，然后确立了三个组内的小组。一组负责对所有高校众多专业的消岐，一组负责基于成绩的推荐，一组负责基于能力的推荐。而我加入了对专业进行消歧的小组，我们也与知识图谱A组的同学进行合作完成了该任务，进行了初步的语义对应。但我们观察到有不少专业对应不太准确，可能是因为名称的原因，也可能是因为语义不同的原因，所以我们对所有的专业数据进行了检查，得到了较为准确的映射关系。  而在所有三个小组的任务完成后，我们的代码都是利用了rank.csv和result.csv这两个文件进行的，但是实际的系统中并不能依靠文件进行推荐，而是应该与实时的数据库进行连接，利用数据库中不断更新的数据进行专业推荐。所以我主动负责了这一部分连接数据库的工作，与第十组负责django开发的同学进行对接，编写了利用数据库生成数据格式的代码，从而使得我们的代码可以不断的随着数据库的更新，扩展高校的数量和专业的变更等信息变动。  此次的软件实践课程让我学习到了除知识与编程能力以外的很多东西，我深刻体会到一个系统的开发需要多少人工的通力协作，我们很多组之间都紧密联系，需要及时的进行对接和商讨，才能将开发任务进行下去，失去任何一个环节都可能造成不可弥补的损失。多人协作的模式也是现在开发软件和完成未来工作的主要方式，让我们都体验到了交流与盒作的重要性，为未来进入工作岗位有着不可或缺的作用。而对于自己负责的任务而言，也让我感到个人的责任在多人协作中的作用。不能因为自己的原因而耽误了整个组的进度，所有开发工作需要每个同学都紧紧地跟上其他同学的步伐。我们需要锻炼这种个人与集体感的平衡，这样才能让整个开发任务效率得到提升，并有利于最后做出一个具有影响力的工程。 | | | | | | | |

2020年9月制