暑期学校实验项目：高考志愿填报助手

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 小组名称 | aiB组 | | | | | | |
| 姓 名 | 王贵涛 | 专业 | 人工智能 | 班级 | 615184 | 学号 | 61518424 |
| 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**  高考是我国人才选拔的重要途经，而高考志愿填报则对于每一名考生来说又是除考试外最重要的一项事务。本次实验以2017-2019三年的各大高校录取分数和各省市的分数段数据为基础，尝试开发一套基于历史数据的高考志愿填报助手的系统。 | | | | | | | |
| **二、小组任务和个人任务**  本小组的具体工作是进行聚类算法的设计与实现，并和其他小组进行对接。我们的聚类算法需要对多年的分数、排名及高校进行分析，得出分段分布模型，以及专业的分布模型。此外，我们还实现了学科分类评级的功能，由于种种原因，没有选择呈现在最后的展示中。  我们小部分成员负责的任务是对每个学校的专业进行评级，这就需要将每个专业的所有学校统计出来，但各个学校的专业名称不尽相同，同一个专业可能有完全不同的叫法。我的个人任务是将各个学校的专业名称划分到一级学科名称中。 | | | | | | | |
| 1. **个人任务需求分析**   我的主要任务是将各个学校的专业名称划分到一级学科名称中。首先在数据库中统计所有的专业名称，保存至一个集合；然后导出一级学科名称的列表。使用相似度匹配的算法将专业名称对应到与它最为相似的一级学科名称，并将其对应的一级学科代码填入数据中，导出csv文件供后续同学使用。 | | | | | | | |
| 1. **实验过程（需附上关键代码及相关说明）**   首先从数据库内导出一级学科，在Firstlevel内，共有83个一级学科。      再从数据库的Majors内导出专业名称至一个集合，去掉重复的部分和括号里的内容，共有602个不同的专业。    之后对每一个专业和一级学科进行相似度比较，将最相似的一级学科和对应的分数存入以专业名称为键值的字典中。    结果发现准确率不够高，因此手动修改优化了部分数据，使之准确率提高。  将中国语言与外国语言顺序对调，这样遇到某语言会优先匹配外语。在一级学科中添加“无”学科，作为那些与所有一级学科的匹配度都为0的专业使用。    去掉一级学科中影响匹配准确率的字，如“学”，留下代表本一级学科的关键字。    部分输出结果如下：    最后根据此预测结果将一级学科对应的id填入专业之后，转出存为csv表格。 | | | | | | | |
| 1. **实验结果与分析**     上图为最终输出的csv表格，作为下一个同学对专业进行评级使用。其中的firstlevelID为majorName栏里专业预测的对应一级学科的序号，对于无法预测的专业和“all”专业此栏为None。 | | | | | | | |
| **六、实验总结与心得体会**  通过本次实验，我更多的是熟悉了码云上团队开发项目时任务和代码管理的操作，相比于以往我们通过私人的方式（QQ等）互相传送代码时更加有条理，所有代码存在云端，可以记录实验过程中的每一次改动，看见每一次的提升和其他成员的工作进展，和同学彼此交流变得方便快捷。  另一方面，本次实验督促我们提前设计好函数的接口，方便队员直接对接，这样每个人的工作虽然分开但又通过准确的接口彼此联系在一起。代码在给其他人使用以及在自己后续做更改时才能更加高效、快捷。 | | | | | | | |

2020年9月制