暑期学校实验项目：高考志愿填报助手

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 小组名称 | AI算法A组 | | | | | | |
| 姓 名 | 周吾君 | 专业 | 人工智能 | 班级 | 一班 | 学号 | 09118114 |
| 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**  高考志愿填报助手意在帮助高考完的考生根据自己所在省份的预测自身适合的大学与专业，有助于考生更加了解自己、明确目标定位。上学期，我们收集了各个省份的一分一段表以及部分高校的专业录取分数线，短学期内，我们组的任务目标为帮助考生预测可选大学，确定其冲一冲，稳一稳，保一保的若干所大学。 | | | | | | | |
| **二、小组任务和个人任务**  小组任务：本组计划用AI算法实现学校推荐，根据学生的高考分数排名，以及所采集大学的录取分数线和录取人数，可以训练一个能够预测学生报考学校的神经网络分类器。  个人任务：  1、找出各个省份对应的经纬度  2、各个大学在所在省份内的排名  3、最后输出每个大学对应的编码：（经纬度+排名）【三维】或者对经纬度做pca成一维这样一共就是二维【如果出现大学空缺，则需要自动往前补位，以防止出现排名不连续的情况】 | | | | | | | |
| **三、个人任务需求分析**  一 数据准备：查询各个省份对应经纬度以及各个大学的全国排名  数据准备组即第一组直到任务布置第二周仍提供任何有效数据，只对之前全体作业结果进行最简单的拼接且并未数据清洗，导致几乎没有任何可以利用的数据，因此自己搜寻了数据。  因为考虑到如果将学校经纬度细分到市级不仅会增大数据量采集而且对于真正实践来说，省份的优先考虑等级要高于市级，因此最终决定每个省份省会的经纬度代表所在省份经纬度，同时用此经纬度作为大学所在地的经纬度，作为编码前两个维度。  对于大学排名这一编码，最开始是考虑每个学校在各自省份排名，数据采集完成后经过讨论，认为将其更改为全国排名更为合适，因此通过查询2020年最新的中国大学排名确定了每个学校全国的排名，作为编码的第三个维度。  二 数据处理  将获取的经纬度信息转换为键为省份，值为经纬度的字典存放进json文件中，之后将大学所在地与与其经纬度一一对应确定编码的前两位，之后将已有大学排名进行排序后，作为编码的第三位，最后输出为键为大学名称，值为编码的字典存入名为‘college code’的json 文件中。 | | | | | | | |
| **四、实验过程（需附上关键代码及相关说明）**  一 首选人工获取每个省份省会的经纬度，再通过json文件处理最终是输出以下格式的文件    二 人工获取大学所在地即相关排名信息  尝试通过多渠道收集，但大多数情况存在数据残缺的缺点，最终选择网页[2020中国大学排名](http://www.gaosan.com/gaokao/265440.html)提供的数据作为排名的信息来源  最初收集的数据为TXT文本，包含很多无用项，需要加以利用的是第2,3,以及最后一项内容    之后通过已经收集完毕的省份经纬度信息以及学校信息，通过对大学所在地及其排名编码  处理数据程序入下图所示：    程序处理完后生成’college\_code’ json 文件，工作完成。 | | | | | | | |
| **五、实验结果与分析**  实验结果：  1 各个省份经纬度：  输出为”省份所在经纬度.json”，其中键为省份名称，值为其省会所在经纬度。格式如下图所示：    2 大学信息编码:  输出为’college\_code.json’，其中键为大学名称，值为大学编码(经纬度，全国排名三个维度)，格式如下图所示：    结果分析：  最终结果的大学编码，每个大学都对应三个维度的编码，其中前两维作为其所在地的经纬度，保证了对于学生对于地理位置的需求同时地域所在位置也是衡量大学品质的一个标准，因为关系到之后的就业问题等等，位置编码是不可缺少的。  编码第三维的大学全国排名则是更直接的表现了一个学校的优劣好坏，虽然学校品质不能仅仅通过单一的排名比较，但鉴于一些经验之谈收集途径什么困难，最终只选择对于大学的综合排名作为大学实力的衡量，虽然相比于人为判断缺少了更多的辅助信息帮忙，但作为系统性的推荐仍然能是一个不错的选择。 | | | | | | | | |
| **六、实验总结与心得体会**  实验总结：  本次个人任务对大学编码，涉及到了数据分析，数据采集，数据准备等多个阶段工作，更重要的大学编码是作为神经网络的输入，因此对于最后神经网络的预测能力有很大的影响，在编码内容选取方面进行了多次的协商与讨论最终选择了最具有代表性的三个特征作为每一所大学的编码，作为最基本的输入部分由实现结果也证明了数据的有效性和可靠性。  心得体会：  本次作业本人工作主要部分是数据的采集与整理，虽然采集部分本应该是数据组队的工作，了解到数据采集的过程之困难以及繁琐，同时采用的是多人小组合作模式，相比于单人工作，最重要的就是及时的沟通与调节，在实验过程中组内多次协商更改大学编码的内容，最终确定了最具有代表性的特征，同时因为是多个小组合作，也体会到了其中的沟通不易，类似于数据准备组第三周才开始清理提供数据的行为也实在是无可奈何，团队合作还是需要每个人都明白自己的责任是什么才能高效的推进，对于以后毕业可能会参与的小组工作也算是打了预防针。 | | | | | | | | |
|  | | | | | | | | |

2020年9月制