暑期学校实验项目：高考志愿填报助手

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 小组名称 | 知识推理和图谱可视化 | | | | | | |
| 姓 名 | 张雨 | 专业 | 人工智能 | 班级 | 091182 | 学号 | 09118203 |
| 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**  背景：基于Web的高考志愿推荐系统。  目的：为高考结束的学子志愿填报提供建议。 | | | | | | | |
| **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：1.根据各省份一分一段表和各高校录取分数表推算除各省多少名可以申请哪些大学。  2.统计每个省有多少985，211，双一流。  3.计算用户于某省填报性价比较高，高地洼地。 | | | | | | | |
| **三、个人任务需求分析**  任务1:  目标：用户选择xx省xx名次能上xx学校。  实现思路：根据各高校录取分数表首先获得某年份某校在某省的最低录取分数，然后根据各省份的一分一段表，将最低分数对应的名次加入到DataFrame中，最后根据输入的名次和省份查找并输出大学名称。  任务2:  目标：统计在我们搜集的数据中各省份985，211，双一流的学校总数供地图组使用。  实现思路：根据学校类型的csv，进行统计并保存输出。  任务3:  目标：根据任务1生成的表格，对于用户生成建议在x省报x校性价比最高。  实现思路：根据任务1生成的表格，获得高校在各省录取的最低排名，然后除以各省当年高考的总人数，获得百分比。对比各省的录取百分比，得到最高性价比。 | | | | | | | |
| **四、实验过程（需附上关键代码及相关说明）**  任务1：  清理数据  屏幕快照 2020-09-25 下午2.49.06  获得一个新的DataFrame，组成为年份，高校，省份和最低分  5847D52D-0B6E-4023-B4AB-43C7DC2CE545获取最低分对应的名次  0C18CF7D-A834-4304-82D5-1B147022DC73屏幕快照 2020-09-25 下午2.45.26  任务3:  确定各省高考人数，获取用户输入的专业（文理科）和大学信息  6BEA331731B885B603E61A0CA3F704FB  获取数据库中该学校对应的所有省份的三年最低录取分数，获取对应分数的最低排名，再除以各省份各年高考的总人数，然后获得百分比，画图显示出来  88DB23702FA99E8A4965FF30D60E7A3D  4AF1068B0EA83D6036FBF6AC6C4E94CA  BA1E616F3F72C7B131E7A74350FD3110 | | | | | | | |
| **五、实验结果与分析**  任务1:  屏幕快照 2020-09-25 下午3.21.26结果节选：  分析：可以看到有些分数在某些省份的一分一段表中超出了，导致了没有名次的结果。这是比较遗憾的点。还可以看到由于某些艺体录取分数线的存在，导致清华大学录取名次偏低的结果，还是比较遗憾的。  任务2:  屏幕快照 2020-09-25 下午3.55.40  分析：由于统计的学校数量并不完全，所以我们基于此统计出的985，211，双一流从总数来说是不准确的  任务3:  BF721F3431BE73E8DF3B0734E81CE9EE  4A563FD5DD110EF927044EF06369C77B  D5851385DAB58A7C3A1167FEE3224A84  95AE7A32D5A2C06A09679A5DAE8F01F5  分析：可以看到南京大学三年来在各省的招生比例虽然有不同但差距并不大。福建省特别的高说明南京大学在该省所招专业最低分可能有艺体生而被拉低了。有几个省份出现0的原因在于数据缺失。 | | | | | | | | |
| **六、实验总结与心得体会**  总结：在本次项目中我们负责的是比较基础简单的统计工作。难度其实并不大，主要难点在于原始数据出错率较高。譬如，在统计可以申请的大学时，我们发现有些学校录取艺体生的分数明显低于平均标准，甚至出现了几十分这样的数据，这对于我们这种选取最低分对应的名次的想法是较为不妥的。我们也考虑过是否使用平均值会好一些，考虑到艺体分也会计入算平均值，所以平均值其实也是偏低的。所以为了比较真实的效果，我们删除了330及其以下的数据，确保了一定的真实性。而在统计各省985，211大学数时，比较遗憾的点是提供的大学并不全，只有89所，而我们的统计是基于这89所学校，所以很多省份其实数据与真实数据并不相符。譬如，最开始浙江大学、南京理工大学等知名大学都没有进入统计结果中。这是比较遗憾的点。最开始我们还准备计算一下高地洼地，即用户这个名次在哪个省报该学校性价比较高。最开始的想法是用该校在某省的最低分对应的名次除以某省高考总人次得到的百分比来对比，但后来考虑到各省其实有倾向招募的省份，譬如各省在自省招募人数肯定是最多的，而在别省会因为政策不同而有很大区别。因此，仅计算百分比其实是比较不真实的，最终结果大概率无法作为有效参考数据。 | | | | | | | | |

2020年9月制