数据科学基础研究报告

1. 小组成员信息

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 姓名 | 学号 | 邮箱 | Python题目完成数量 | 职责 |
|  |  |  |  |  |
|  |  |  |  |  |
| 陈子合 | 181250019 | 843731853@qq.com |  |  |

1. 研究问题

对题目难度进行分类，将题目分成为容易，较易，中等，较难，困难五个等级。

1. 代码开源地址

代码地址：https://github.com/lxzssfd/DataScience

代码主要分为三部分：第一部分是对课程所提供的json文件进行解析处理，得出我们所需要的数据生成新的json文件；第二部分对上一部分得出的数据进行标准化处理，利用熵权法得出各个数据对于最终结果的影响权重；第三部分对数据进行聚类处理，得出五种不同难度的分类。

1. 研究方法
2. 对于原始数据，求出每道题目最终得分，提交代码长度以及python代码占比的算术平均值。
3. 对于求出来三种均值，对数据进行标准化处理，以消除数据范围大小不同所带来的结果偏差 //TODO
4. 利用熵权法求出三种数据对于题目难度的影响权重。//TODO
5. 利用聚类的方法讲题目根据难度大致分为五类。//TODO
6. 案例分析

根据课程所提供的数据，我们提出了三个能够衡量题目难易程度的指标，分别是题目平均得分，提交代码平均长度，提交代码python代码占比。在上学期的编程过程中我们发现，在完成部分困难的题目时，同学们会倾向于复制网上查找到的代码，且这些代码大部分都是java或是c++代码，所以可以推测某个题目越困难，其提交代码的python代码占比就会越小，平均得分就会越低，代码平均长度会越长。

根据所提供的数据test\_data.json，我们求出每道题目的平均得分，代码平均长度以及python代码占比。由于这三种数据的的范围不同，而且python代码占比与题目的难度成反比，平均代码长度与题目难度成正比，这样的情况对最后的结果会有影响，所以我们先使用了数据标准化的方法，将三组数据都变为值在0-1的区间内，且都与题目的难度成正比，方便了后续工作的进行。

对于标准化后的数据，我们使用熵权法，先求出各项指标的信息熵，将其比重作为数据对题目难度影响所占的比重，经过计算可以得出python占比，代码长度以及最终得分所占的比重分别为：0.599692，0.070708，0.329599701.

有了上面各项数据所占的比重以及各项数据标准化后的结果，我们就可以得到对于每道题目难度的数值评价，将这些数据都放在以为坐标上，根据它们之间的距离，也就是他们的相似程度，我们就可以使用聚类的处理方法，将相似的题目归为一类，也就可以由此将题目分为五种不同的类型。最终的结果为：最简单题目有606道，较简单有158道，中等难度有76道，较难有27道，极难有15道，详细结果请参考附录。

1. 课程意见

由于慕测平台的一些局限性可能会导致本次作业中所使用可供用于题目难度分析的数据较少，得出的结果可能会跟实际有些小小的偏差，是否可以改进慕测系统方便以后提供更加全面的数据，例如可以记录在网页上运行的记录，这样就可以提供更多可以分析的数据。

1. 附录