第十三届“华中杯”大学生数学建模挑战赛题目

B 题　技术问答社区重复问题识别

技术社区问答平台作为用户互相分享交流的社区平台，近年来逐步成为用户寻找技术类疑难解答的首要渠道。 各分类技术性问题的文本数据量不断攀升，给问答平台的日常运营维护带来了挑战。随着新用户的不断加入以及用户数量的增加，新用户提出的疑问可能已经在平台上被其他用户提出并解答过，但由于技术性问题的复杂性，各个用户提问的切入角度不同，用问题标题关键词匹配的搜索系统无法指引新用户至现有的问题。于是，新用户会提出重复的问题，而这些问题会进一步增加平台上的文本量，导致用户重复响应相同的问题。对于这种现象，通常的做法是及时找到新增的重复问题并打上标签，然后在搜索结果中隐藏该类重复问题，保证对应已解决问题出现的优先度。所以，建立一个基于自然语言处理技术的自动标重系统会对问答平台的日常维护起到极大帮助。

目前，问答平台上的问题标重主要依靠用户人工辨别。平台用户会对疑似重复的问题进行投票标记，然后平台内的管理员和资深用户（平台等级高的用户）对该问题是否被重复提问进行核实，若确认重复则打上重复标签。该过程较为繁琐，依赖用户主观判断，存在时间跨度大、工作量大、效率低等问题，增加了用户的工作量且延长了新用户寻求答案所需的时间。因而，如能建立一个检测问题重复度的模型，通过配对新提出问题与文本库中现存问题，找出重复的问题组合，就能提高重复问题标记效率，提高平台问题的文本质量，减少问题冗余。同时，平台用户也能及时地根据重复标签提示找到相关问题并查看已有的回复。

附件给出了问答平台上问题的文本内容记录，以及比较两个问题之间是否重复的数据集。请根据附件给出的问题文本数据及问题配对信息，建立一个能判断问题是否重复的分类模型，并解决：

1）输出样本问题组为重复问题的概率；

通常使用 F1-score 对分类模型进行评价：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i} \ ,$$

其中P_i为第i类的查准率，R_i为第i类的查全率；

2）从附件问题列表中，给出与目标问题重复概率最大的前 10 个问题的编号；

对于每个问题的预测结果采用 top K 列表对其进行评估，评估公式如下：

$$R = \frac{N_{detected}}{N_{total}},$$

其中$N_{detected}$ 为在 top K 列表结果中正确检测到的重复问题编号数量，N_{total}为该样本实际拥有的重复问题数量。评估时 K 取 10，若样本中无重复问题则不会计分。

附 数据说明

每个问题类别对应两个附件。

附件 1 为问题编号、对应问题内容和该问题分类的数据。具体表结构示例如下：

问题编号	问题内容
86333	我有印度行政区的形状文件。和海岸线的折线形状文件。 如何输出一张表格告诉我哪些地区有海岸线的表格? 谢谢!
68897	我选择了一些地址点，我想基于选择中的行更新地址字段。如何弄清楚选择 Taable 的名称以传递给 ArcPy.updateCursor（）？
…	…

附件 2 为问题两两组合组成的问题组。每个问题组对应了标示该组内两个问题是否重复的标签数据。具体表结构示例如下：

问题组	问题编号 1	问题编号 2	问题组是否重复 （0、1 分别代表不重复和重复）
1	86333	68897	0
2	50415	25518	1
…	…	…	…