摘要

在本文之中，我们分析了每天提交的报告数量随时间变化的趋势，并设计了一个模型来预测给定日期的提交的报告数量。我们考察了单词的使用频率、首字母、元音字母数量等属性，并采用一定的方式将这些属性量化。借助这些属性，我们训练了一个BP神经网络来预测未来某一日期给定的单词在困难模式下尝试次数的占比分布。同时，依据单词在困难模式下各种尝试次数的占比分布，我们设计了一个得分模型来量化一个单词的难度。依据这些量化的难度，我们可以人为地设置难度等级。在此基础上，我们就可以实现对单词的难度分类。我们还设计了一个分类模型，用来预测给定单词的难度。在论文的最后，我们还在给定的数据集中发现了一些有趣的特性，我们将其可视化地表现了出来。

在问题一中，为了预估给定的日期的提交的报告数量，我们基于指数拟合的方式构建出了一个预测公式。将日期所对应的单词编号作为自变量输入即可得出对应的预测结果。我们还选择了一些属性和特征并对它们做了一定程度的量化，通过这些量化的结果，我们可以判断一个单词困难模式下尝试次数的分布是否与某一个属性具有很强的关联性。在此基础上，我们选择那些关联性强的属性，舍弃那些对分布结果影响不大的属性，并将这些属性作为我们下一问题的数据输入。

在问题二中，我们训练了一个BP神经网络，将问题一中得到的量化的属性作为输入来预测一个给定的单词在困难模式下尝试次数的分布占比。借助训练后的BP神经网络，我们预测了EERIE这个单词对应的尝试次数占比。我们还对该模型进行了误差分析，用来量化其不确定性。同时，我们还展示了在不同参数（隐藏层等）情况下的训练结果，与我们所选择的训练结果进行对比，详细分析了我们选择特定参数的理由。

在问题三中，我们根据单词的量化属性建立了一个得分模型来表示一个单词的难易程度，单词的得分越高，表示单词越简单。以得分模型为基础，我们对难度进行了不同等级的划分，作为我们分类时候的类别。我们建立了一个随机森林分类的模型来进行详细的分类，将我们已经依据得分模型分类好的数据集作为输入，对我们的模型进行了训练，模型经过训练后在测试集的表现在可接受的范围之内：准确率80.4%，召回率80.4%，精确率为87.4%，F1的值为83.7%。通过训练之后的模型，我们可以分析给定的单词的难度。我们分析了EERIE这个单词的难度。以得分模型为桥梁，我们甚至可以与问题二的模型建立联系，最终我们发现问题二与问题三的结果是可以互相佐证的。

在问题四中，我们探究给定数据集的一些比较有趣的特性。我们分析了困难模式得分报告数量与总报告数量的比例关系并分析了困难模式下不同的尝试次数平均占比的分布，并给出了我们对这些特性的一些看法。

在论文的最后，我们对模型的灵敏性进行了分析，并给出了调整参数之后我们模型的结果变化。