第 8 小组开题报告

题目：

矛盾，我亲爱的华生

组员信息：

2006500021 雷昊
2006500031 韦瑞德
2006500038 申选贤

题目链接：

https://www.kaggle.com/competitions/contradictory-my-dear-watson/code

主要内容：

如果你有两个句子，它们可能通过三种方式相互关联：可能前者能推理出后者、或是前后矛盾、也可能是两者并不相关。自然语言推理 NLI 是一个流行的 NLP 问题，涉及确定句子对（由前提和假设组成）如何相关。

我们的任务是创建一个 NLI 模型，该模型将 0、1 或 2（对应于蕴涵、中性和矛盾）的标签分配给前提和假设对。为了使事情更有趣，训练和测试集包含十五种不同语言的文本！可以通过查看数据页面找到有关数据集的更多详细信息。

科学意义：

在自然语言中有时会出现一些有矛盾、有歧义的话语，如果不在预处理中筛选其中有矛盾的词、句子可能会导致后期机器难以理解语言意思、理解具有二义性等问题。

而研究该主题可以学习处理 NLP 中矛盾、不相关或是蕴含等关系的句子，可以帮助机器提高对后期自然语言处理的效率、并且能提高机器的语言逻辑能力。

数据集：

https://www.kaggle.com/competitions/contradictory-my-dear-watson/data

数据集说明：

原文：

- **train.csv**: This file contains the ID, premise, hypothesis, and label, as well as the language of the text and its two-letter abbreviation
- **test.csv**: This file contains the ID, premise, hypothesis, language, and language abbreviation, without labels.
- **sample_submission.csv**: This is a sample submission file in the correct format:

  : a unique identifier for each sample

  : the classification of the relationship between the premise and hypothesis (0 for entailment, 1 for neutral, 2 for contradiction)`id` `label`

译文：

- **train.csv**：此文件包含ID，前提，假设和标签，以及文本的语言及其两个字母的缩写
- **test.csv**：此文件包含 ID、前提、假设、语言和语言缩写，不带标签。
- **sample_submission.csv**：这是一个正确格式的示例提交文件：：每个样本

  的唯一标识符：

  前提和假设之间关系的分类（0表示蕴涵，1表示中性，2表示矛盾）`id` `label`

科学问题：

对数据集中的多种语言，我们要进行预处理并提升对前后文之间的关联判断，对预测的上下文给出蕴含、中性（不相关）、矛盾的评判

研究内容：

因为是因到多种国家的语言，故使用 Bert-base-multilingual-cased 进行数据预处理，后使用 CNN 卷积神经网络构建模型，对测试集进行预测

预期目标：

因为是基于三分类的自然语言逻辑推理问题，并设计到多个国家的语言，故目标正确率到达 80%以上，并且保证模型运行准确率的基础上、增强模型的健壮性和鲁棒性。

参考资料：

bert_encode 原理 伯特 (huggingface.co)

BERT 进行文本分类 保姆级教程，用 PyTorch 和 BERT 进行文本分类 - 知乎 (zhihu.com)