

1. 题目：使用灾难推文进行自然语言处理

2.成员：(1) 黄志伦 2006500014

(2) 吴佳远 2006500010

3.Kaggle 链接：<https://www.kaggle.com/competitions/nlp-getting-started>

4.科学意义：如今是网络和信息时代，使得人们能够实时宣布他们正在观察到的情况，正因如此，越来越多的新闻组织和救灾机构需要程序化的对灾难推文进行，可以更好的帮助政府和组织及时了解公众对灾难事件的反应和情感，从而更有效地为受灾群众提供支援和帮助。除此之外，这种技术还可以预测和监测灾难的发生，并提供实时的灾难相应和救援。

5.科学问题：预测哪些推文是关于真实灾难，哪些不是

6.研究内容：

(1) 数据集介绍：

数据下载链接：<https://www.kaggle.com/competitions/nlp-getting-started/data>

文件

- **train.csv** - the training set

由于用的是比赛的数据集，**test.csv** 没有给出 **target**，所以打算将 **train.csv** 分出部分样本作为测试集，进行模型预测。

列

- `id` - 每条推文的唯一标识符
- `text` - 推文文本
- `location` - 推文发送的位置（可能为空）
- `keyword` - 推文中的特定关键字（可能为空）
- `target` - 是否为灾难推文（1/0）

(2) 算法介绍:

计划使用 transformer 模型来进行灾难推文的预测

Transformer 模型是一种基于注意力机制的神经网络模型，主要使用了自注意力机制和多头注意力机制对输入序列进行编码和解码。

Transformer 模型在处理自然语言任务时表现优异，相对于 RNN 和 CNN 模型，Transformer 模型特有的自注意力机制不仅无需考虑序列顺序，还能更好的捕捉上下文信息。除此之外，还能并行计算，提高模型的表达和泛化能力。

7.预测目标：灾难推文预测准确率 90%以上