关于灾害推文的自然语言处理

——预测哪个推文是真实灾害

小组成员信息：

智能 201 池福金 2006500036

智能 201 尤鸿兴 2006500032

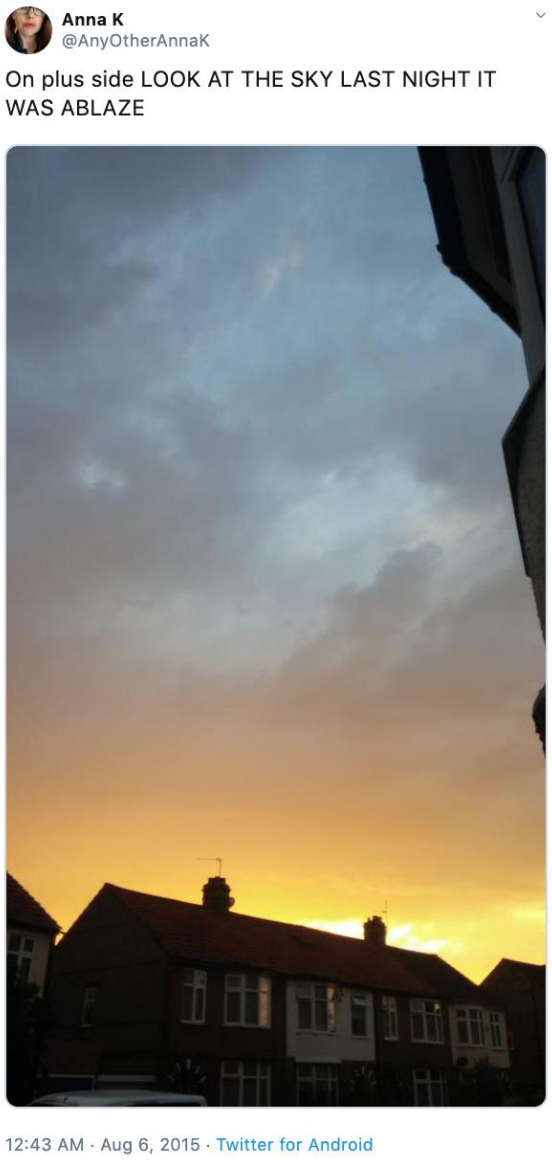
计科 204 张瑞鹏 2007200081

Kaggle 链接：[Natural Language Processing with Disaster Tweets | Kaggle]

研究目的：

现在越来越多的人通过社交媒体传播信息，有的信息是对自然灾害的描述，有的并不是，我们想通过制作一个模型识别这些文案来判断是否描述关于自然灾害，以此将这些信息筛选出来，方便人们查看关于自然灾害文案

推文示例：



在这里作者用的“ablaze”是燃烧的意思，不过明显是比喻，而不是某种灾难。对人来说这很容易分辩但是对于机器来说则不然。

所以我们的任务是构建一个机器学习模型，使用提供的 10000 条手动分类的推文组成的数据集，预测哪些是真实灾害哪些不是，。

数据集：

本次使用的数据集是 kaggle 中提供是数据，数据的内容包括推特中的文案，以及可能会提供该文案的发送位置和关键字，每条文案对应一个独有的 ID 号，和每条文案都有一个标签，用来标明是否为描述灾难的文案。

train.csv

数据连接：

数据集截图：

id	keyword	location	text	target
1			Our Deeds are the Reason of th	1
4			Forest fire near La Ronge Sask. (	1
5			All residents asked to 'shelter in	1
6			13,000 people receive #wildfires	1
7			Just got sent this photo from Ru	1

…

41			Do you like pasta?	0
44			The end!	0
48	ablaze	Birmingha	@bbcmtd Wholesale Markets ak	1
49	ablaze	Est. Septe	We always try to bring the heav	0
50	ablaze	AFRICA	#AFRICANBAZE: Breaking news:	1

如图中每个数据有以下特征：

- id- 每条推文的唯一标识符
- keyword- 推文中的特定关键字（可能为空）
- location- 发送推文的位置（可能为空）
- text- 推文文本
- target- 仅在 train.csv 中存在表示一条推文是关于真正的灾难（1）或不是（0）

使用的方法：

1，使用类似 torchtext 的自然语言数据预处理的库进行数据预处理。

2，预计使用基本 Transformer 模型中的 Encoder 层，然后对其输出取均值作为模型的输出。

3，或者经过实际操作时候，转换其他的方法。

预期目标：

1，输入测试的的推文（如果有 keyword 和 location 就添加），模型会输出概率，对应[0,0.5]视为 0，[0,1]视为 1，来表示推文是否是关于真正的灾难。

2，可以做到对全部的测试集进行预测，并且可以对自己写的英文句子进行预测，测试是否有描述灾难。