資料分析與學習基石 個人專題期末報告

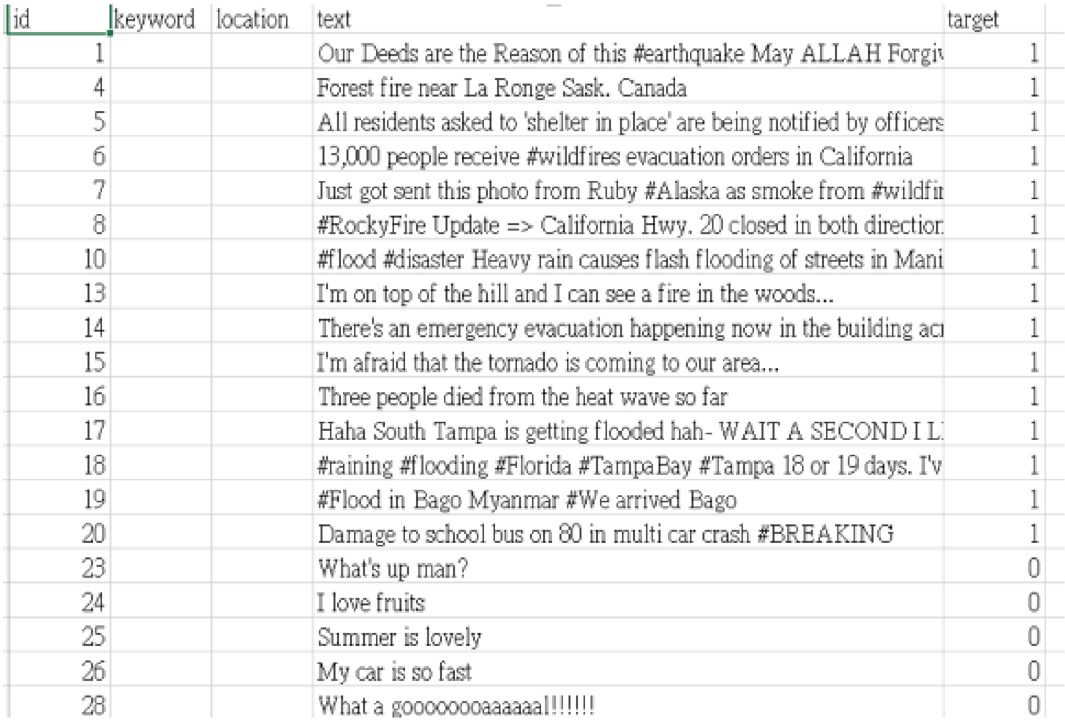
楊智翔 F74086187

- 問題定義:
  分辨一篇文章(或句子)是否跟災難有關

- 資料取得與收集:
  kaggle 上面有一個 competition
  名稱是: Natural Language Processing with Disaster Tweets

- 競賽介紹:
  Twitter 已成為緊急情況下的重要溝通渠道。智能手機的無處不在使人們能夠即時宣布他們現在遇到的各種狀況。正因為如此，更多的機構對以編程方式監控 Twitter 感興趣（即救災組織和新聞機構)。但是，一個人的話是否真的在宣布災有時候並不清楚。因此在這次比賽中，我的任務是建立一個模型，讓此模型可以預測哪些推文是關於真實災難的，哪些不是。

- 資料一小部分樣貌:



id	keyword	location	text	target
1			Our Deeds are the Reason of this #earthquake May ALLAH Forgiv	1
4			Forest fire near La Ronge Sask. Canada	1
5			All residents asked to 'shelter in place' are being notified by officers	1
6			13,000 people receive #wildfires evacuation orders in California	1
7			Just got sent this photo from Ruby #Alaska as smoke from #wildfir	1
8			#RockyFire Update => California Hwy. 20 closed in both direction	1
10			#flood #disaster Heavy rain causes flash flooding of streets in Mani	1
13			I'm on top of the hill and I can see a fire in the woods...	1
14			There's an emergency evacuation happening now in the building acı	1
15			I'm afraid that the tornado is coming to our area...	1
16			Three people died from the heat wave so far	1
17			Haha South Tampa is getting flooded hah- WAIT A SECOND I L	1
18			#raining #flooding #Florida #TampaBay #Tampa 18 or 19 days. I'v	1
19			#Flood in Bago Myanmar #We arrived Bago	1
20			Damage to school bus on 80 in multi car crash #BREAKING	1
23			What's up man?	0
24			I love fruits	0
25			Summer is lovely	0
26			My car is so fast	0
28			What a goooooooaaaaaal!!!!!!	0

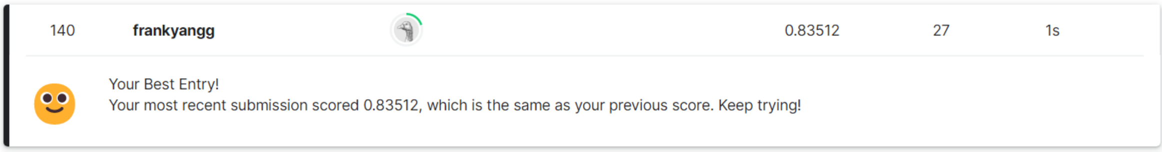
- 任務:
  利用他給的資訊(keyword, location, text)去預測 target(0 或 1)

- 目標:
  讓 model 可以更好的分辨文章，盡量提升準確率(f1_score)

- 競賽延伸(可能的利用方式):
  分辨假新聞或者對文章進行分類 如新聞等等

- 使用的模型與嘗試:
  模型:bert 預訓練: bert-base-uncased(曾嘗試的有 bert-base-cased, bert-large-uncased)。Fine-tune 階段:BertForSequenceClassification(本身預設 loss function 是 CrossEntropyLoss，有試試看用 focal loss，不過沒比較好)更改其他參數如:epochs, learning_rate, 取的 train test 資料大小與分布等等

- 總結:
  1. 上次報告的預期成果與目標: 總共在 leaderboard 上面的隊伍有 900 多隊，預期目標分數達到 0.81 並且進前 300 名。
  2. 最終成果:分數達到 0.835 排名達到第 140 名，有成功達到我的目標並且更進一步。



140 frankyangg 0.83512 27 1s

Your Best Entry!
Your most recent submission scored 0.83512, which is the same as your previous score. Keep trying!