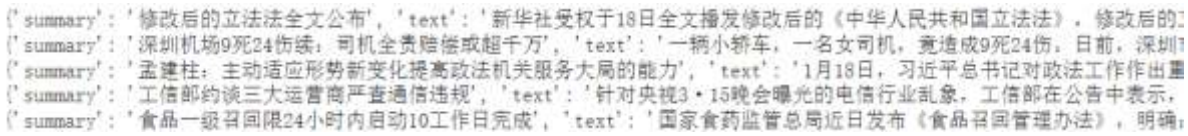
Dataset

huggingface 已經把 train, validation 的 dataset 弄好了，我只要整理一下就好，我把 dataset 整理成和英文文本一樣的 dataframe(有 text 和 summary)，我取的資料是前 2000 筆資料。T5 和 gpt2 都是用同樣的資料。



Model

T5 模型的 tokenizer 沒辦法處理中文，所以我換成其他 tokenizer，但是結果還是沒有很好。雖然 t5 tokenizer 的分數看起來最高，但是實際上他根本看不懂，把輸出印出來看就是一堆 Unknown。

這是 t5 模型用各種 tokenizer 的 evaluation

[t5 tokenizer]	[bert_fast tokenizer]
Rouge-L-P': 0.23259874230344826, 'Rouge-L-R': 0.31306666666666666, 'Rouge-L-F': 0.31306666666666666, 'Rouge-2-P': 0.0010158730158730158, 'Rouge-2-R': 0.0026666666666666666, 'Rouge-2-F': 0.0026666666666666666	Rouge-L-P': 0.020760127685004695, 'Rouge-L-R': 0.044308225295973246, 'Rouge-L-F': 0.044308225295973246, 'Rouge-2-P': 0.003104483688737341, 'Rouge-2-R': 0.0068518832392900075, 'Rouge-2-F': 0.0068518832392900075}
[bert tokenizer]	[Roberta tokenizer]
Rouge-L-P': 0.023500622810601363, 'Rouge-L-R': 0.04557549238840649, 'Rouge-L-F': 0.04557549238840649, 'Rouge-2-P': 0.003721591667025889, 'Rouge-2-R': 0.007925213132299378, 'Rouge-2-F': 0.007925213132299378}	'Rouge-L-P': 0.005650396093632591, 'Rouge-L-R': 0.009910578700485482, 'Rouge-L-F': 0.009910578700485482, 'Rouge-2-P': 0.00028707135027924215, 'Rouge-2-R': 0.0005839727195225917, 'Rouge-2-F': 0.0005839727195225917

T5 預訓練的資料有蠻多種的， 有 t5-small, t5-base, t5-large, t5-3b, t5-11b，但後面兩個太大了跑不動，所以我就用前面 3 個跑跑看

T5-small	T5-large
Rouge-L-P': 0.08999413719466974, 'Rouge-L-R': 0.18579441575105168, 'Rouge-L-F': 0.18579441575105168, 'Rouge-2-P': 0.04394062043989692, 'Rouge-2-R': 0.09282632838589598, 'Rouge-2-F': 0.09282632838589598	Rouge-L-P': 0.09628490606639933, 'Rouge-L-R': 0.16936987135680664, 'Rouge-L-F': 0.16936987135680664, 'Rouge-2-P': 0.06994522110493928, 'Rouge-2-R': 0.13168205444935036, 'Rouge-2-F': 0.13168205444935036

Train

Gpt2 模型我就直接用 gpt2 自己的 tokenizer，可是在訓練時，input batch size 和 target batch size 都會不合，所以我在 target 後面補零當作 padding，讓他們大小符合可以計算 loss，我還有加 attention mask，讓模型在訓練時不要過於注意填充的部分。

這是 gpt2 的分數

Average ROUGE-2 score: 0.018863725936945745

T5 和 GPT2

T5 和 gpt2 在架構上的差別是 t5 有 encoder 和 decoder，但 gpt2 只有 decoder。T5 的概念是文本到文本的轉移任務，輸入是文本，輸出也是文本；

gpt2 的輸出會由前面的一部分輸出來決定，這種機制叫做自回歸。