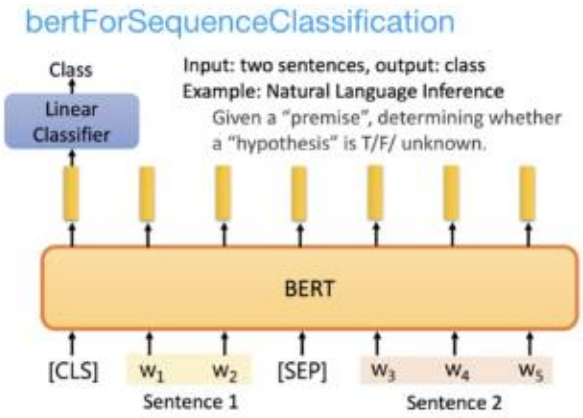
OASIS ML Group TRAINING 07

◆ Natural Language Processing with BERT

In this practice, you are going to build a BERT model which is usually used in Natural Language Processing(NLP). Because the number of parameters of BERT is very big, you just need to load the pretrained model and fine tune the model to fit the specific downstream task. Please refer to the website which provided in the last page. This dataset is from Kaggle and the task you need to do is Fake New Classification. You need to compare the two news titles and distinguish among **agreed and disagreed and unrelated.** This is a sequence for multi-classification task.

Follow the hints below to finish it.



▫ Practice：

■ Analysis datasets：

The dataset consists of two files：train.csv, test.csv

The train.csv file contains the 320,767 training examples and labels. You need to predict the result with “title1_zh” and “title2_zh”, then compare the result and fine tune the model with “label”. The test.csv file contains 80,126 test examples. There is no “label”, you just need to make predictions.

Hint 01：Import library you needed.

Hint 02：choose a version of BERT pretrained model and load it.

- bert-base-chinese
- bert-base-uncased
- bert-base-cased
- bert-base-german-cased
- bert-base-multilingual-uncased
- bert-base-multilingual-cased
- bert-large-cased
- bert-large-uncased
- bert-large-uncased-whole-word-masking
- bert-large-cased-whole-word-masking

```
PRETRAINED_MODEL_NAME = "bert-base-chinese"  # 指定繁簡中文 BERT-BASE 預訓練模型

# 取得此預訓練模型所使用的 tokenizer
tokenizer = BertTokenizer.from_pretrained(PRETRAINED_MODEL_NAME)
```

Hint 03：Prepare data with features which you need.

	text_a	text_b	label
0	苏有朋要结婚了，但网友觉得他还是和林心如比较合适	好闺蜜结婚给不婚族的素岚扔花球，倒霉的素岚掉水里笑哭苏有朋！	unrelated
1	爆料李小璐要成前妻了贾乃亮模仿王宝强一步到位、快刀斩乱麻！	李小璐要变前妻了？贾乃亮可能效仿王宝强当机立断，快刀斩乱麻！	agreed
2	为彩礼，母亲把女儿嫁给陌生男子，十年后再见面，母亲湿了眼眶	阿姨，不要彩礼是觉得你家穷，给你台阶下，不要以为我嫁不出去！	unrelated
3	猪油是个宝，一勺猪油等于十副药，先备起来再说	传承千百的猪油为何变得人人唯恐避之不及？揭开猪油的四大谣言！	unrelated
4	剖析：香椿，为什么会致癌？	香椿含亚硝酸盐多吃会致癌？测完发现是谣言	disagreed

Hint 04：Show the analysis of the data, you can also visualize the information.

```
unrelated    0.678674
agreed       0.294626
disagreed    0.026700
Name: label, dtype: float64
```

Hint 05：Build dataset and **dataloader** to make data fit for BERT model.

Hint 06：Load the model for this downstream task(recommended), or build it by yourself.

```
model = BertForSequenceClassification.from_pretrained(
    PRETRAINED_MODEL_NAME, num_labels=NUM_LABELS)
```

Hint 07：Implement train and predict function, you can follow the reference website to accomplish.

```
def get_predictions(model, dataloader, compute_acc=False):
    ???
    ...
    ...
```

Hint 08：Define Hyper-parameters by yourself.

```
optimizer = torch.optim.Adam(model.parameters(), lr=1e-5)
EPOCHS = 20
```

(You can tune by yourself.)

Hint 09：Train the model and show the trends of training loss and accuracy(make the training results visualized or any method you come up with), ensure they will converge.

```
[epoch 1] loss: 608.239, acc: 0.904
[epoch 2] loss: 466.882, acc: 0.924
[epoch 3] loss: 391.834, acc: 0.938
[epoch 4] loss: 329.941, acc: 0.946
[epoch 5] loss: 277.181, acc: 0.956
[epoch 6] loss: 234.651, acc: 0.960
[epoch 7] loss: 201.066, acc: 0.965
[epoch 8] loss: 174.458, acc: 0.969
[epoch 9] loss: 150.795, acc: 0.972
[epoch 10] loss: 133.343, acc: 0.973
[epoch 11] loss: 119.180, acc: 0.972
[epoch 12] loss: 106.254, acc: 0.978
[epoch 13] loss: 95.489, acc: 0.980
[epoch 14] loss: 86.657, acc: 0.981
[epoch 15] loss: 79.388, acc: 0.982
[epoch 16] loss: 73.053, acc: 0.982
```

Hint 10：Export the test results as “bert_pred_testing_samples.csv”.

Id	Category
321187	unrelated
321190	unrelated
321189	unrelated
321193	unrelated
321191	unrelated
321194	unrelated
321192	unrelated
321197	unrelated
321195	unrelated
321199	unrelated
321198	agreed
321201	agreed
321196	agreed
321200	agreed
321203	unrelated
321202	agreed
321206	agreed
321204	agreed
321207	unrelated
321208	unrelated
321205	unrelated
321209	unrelated
321211	unrelated

Reference: [進擊的 BERT: NLP 界的巨人之力與遷移學習]