Text Mining HW3 Report R12723006 財金碩一 林姝延

- 執行環境：vs code
- 執行語言：python 3.11.5 (anaconda3)
- 使用 Model：L-2_H-128_A-2, BERT-Tiny
- 執行方式：將存有 1095 個 txt.檔的 PA1-data 資料夾以及 training_new.txt 檔案存放在與當前執行程式相同的路徑下，並直接讀取同樣位於此路徑的資料夾與檔案。而最終輸出的預測結果 csv 檔案同樣會存到與程式碼相同路徑的位置。
- 邏輯說明：
    - 首先讀取 1095 個 txt 檔案，建構一個 包含文本和標籤資訊的 data frame。並讀取 training_new.txt 檔案，取得各個 class 作為 training dataset 的文章編號，建構 training dataset。
    - Load pre-trained BERT 模型，並提取 training data 的 CLS embeddings。
    - 訓練 SVM 模型。
    - 批次處理 testing data，建構 CLS embeddings。
    - 以 SVM 模型分批次進行預測。
    - 匯總所有批次的預測結果到 data frame 中。
    - 把預測結果以 id 排序並存成 csv 檔。
    - 輸出預測結果的 precision, recall, f1。
- 註解及特殊處理：
    - 程式執行過及中自動下載 Hugging Face Transformers 資料庫的 pre-trained BERT based model, google/bert_uncased_L-2_H-128_A-2, 為與原 uncased_L-2_H-128_A-2 model 兼容、PyTorch 版本的 pretrained BERT model。
    - 為了避免執行過程出現 Kernel crash，在 testing 時按批次處理，也就是將 batch size 設定為 32，每次進行 CLS embeddings 的建構以及預測都以 32 筆為單位。
- 執行結果：
  所得預測結果如下：

result_df_svm

✓ 0.0s

	Id	Value
111	17	4
1073	18	2
842	20	5
869	21	8
939	22	5
...	...	...
60	1091	3
120	1092	1
74	1093	7
248	1094	11
194	1095	5

900 rows × 2 columns