農業文章文字標註及辨識
報告說明文件

一. 環境

(1) 環境
使用 kaggle 提供的線上雲端環境，使用 kaggle 提供之免費 GPU 等硬體

(2) 語言
Python

(3) 套件(函式庫)
使用到的函示庫: pandas、re、os、openpyxl、matplotlib.pyplot、transformer、torch、random，其餘 kaggle 環境中安裝函式庫請詳閱 requirement.txt 檔案

(4) 預訓練模型
預訓練模型使用 hugging-face, bert-base-chinese

(5) 額外資料集
本次競賽使用之資料集僅使用主辦方提供之 Train、Public 以及 Private 資料集，沒有使用額外資料集

二. 資料處理

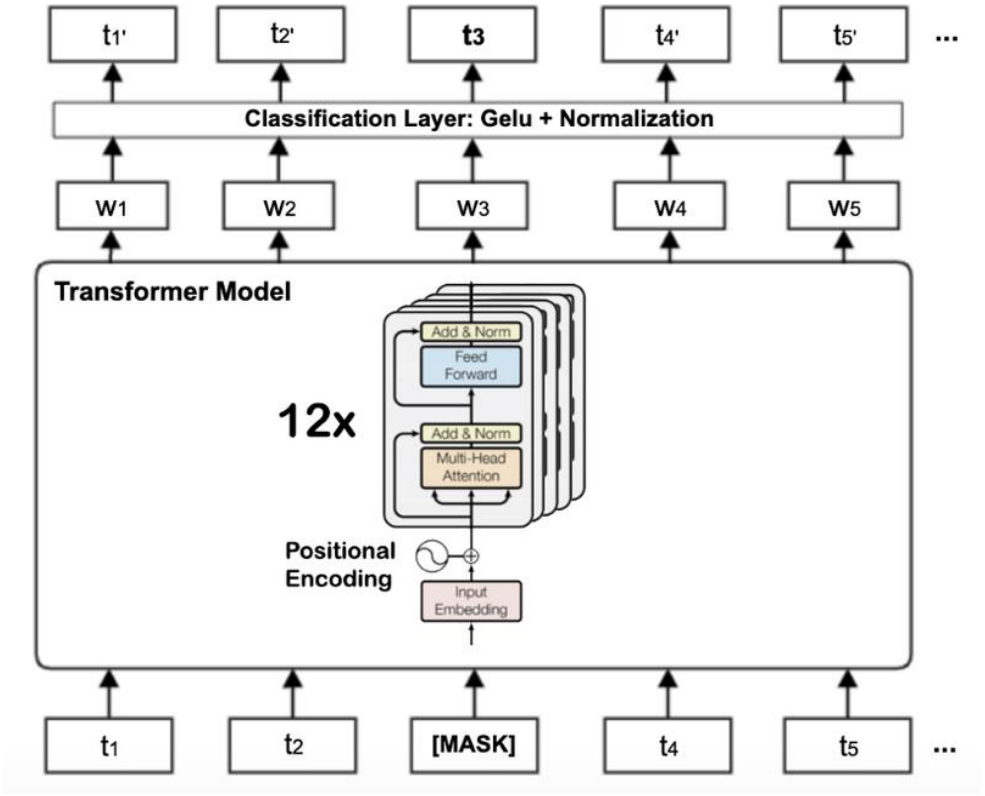
(1) 文章清理
使用 Regular expression operations 套件清理輸入文章，其中會將所有數字取代為 1，並刪除所有標點符號。

(2) 關鍵字處理
據關鍵字集合篩選出文章中包含的關鍵字，並將該文章的關鍵字集合併入輸入。範例如下，訓練集第 3 篇文章將處理為：
「性費洛蒙、蘇力菌、微生物製劑、青蔥、蔬菜、斜紋夜蛾、甜菜夜蛾、蛾類、夜蛾類。夏季為斜紋夜盜及甜菜夜蛾發生盛期.....（後略）」

(3) 模型輸入處理
最後配合 hugging-face 模型的輸入長度限制，擷取文章開頭片段傳入 hugging-face 預訓練的 BertTokenizer, bert-base-chinese 作詞嵌入。

三. 模型架構

(1)模型: hugging-face, bert-base-chinese
(2)使用 BERT 預訓練模型，並保留預訓練模型架構



圖一 BERT 架構圖

四. 訓練方式

(1) Learning-rate: 1e-5
(2) EPOCH: 3
(3) Optimizer: Adam
(4)訓練過程: 使用預訓練模型再做 Fine-Tune

五. 分析&結論

(1) 選擇模型和參數的理由

本小組參考並測試了 hugging-face 和 github 上提供的多個預訓練模型，包括 bert-chinese-mmt, albert, bart, roberta,chinese-macbert-base 等等，大多數模型在該競賽提供的農業文章資料集中，在上述學習率和 epoch 數量下的測試跑分都在 0.3-0.8 之間，而只有 bert-base 模型可以達到更高，因此本小組採用 bert-base-Chinese 模型。

(2) 改進方向

本組所做的 project 在資料處理階段無法處理過長的資料集，因此，若農業文章的關鍵訊息在大於加上關鍵字後的第 256 個字元以後的部分，那麼就會出現關鍵訊息的遺漏。所以本 project 的未來改進方法之一就是在模型的輸入部分作出改進，增加模型可以處理的文字長度。

六. 程式碼

請根據附檔程式第 3 儲存格中的變數 path 鏈結三個關鍵字 excel 檔案、訓練資料集輸入和其標籤、測試資料集輸入，直接執行該程式將輸出訓練後模型和測試集預測標籤。

七. 使用的外部資源與參考文獻

(1) bert-base-chinese, hugging-face,
Web site: https://huggingface.co/bert-base-chinese

聯絡資料

- **隊伍**

隊伍名稱	Privata Leaderboard 成績	Privata Leaderboard 名次
92NLP_107502508	**0.8100313**	5

- **隊員(隊長請填第一位。英文和信箱為獎狀製作所需，請確實填寫)**

姓名(中文)	姓名(英文)	學校、系所(中文)	學校、系所(英文)	電話	E-mail
黃印榕	HUANG,YIN-JUNG	國立中央大學資訊工程學系	Department of Computer Science & Information Engineering, National Central University	0912713029	eo1141495@gmail.com
黃曜駿	HUANG,YAO-JIUN	國立中央大學機械工程學系	Department of Mechanical Engineering, National Central University	0978876257	zxc22661560@gmail.com
陳哲安	CHEN, JHE-AN	國立中央大學電機工程學系	Department of Electrical Engineering, National Central University	0988400175	alanchen0226@g.ncu.edu.tw
馮智詮	FENG, ZHI-QUAN	國立中央大學資訊工程學系	Department of Computer Science & Information Engineering, National Central University	0913780656	fzq1999@qq.com

- **教授(英文和信箱為獎狀製作所需，請確實填寫)**

若為"連結課程"課堂作業或期末專題，請填授課教授，以利依連結課程彙整；

若不是"連結課程"，但某教授實際參與指導，請填該位教授

若以上兩者皆非，可不填

教授姓名(中文)	教授姓名(英文)	課程名稱(含課程代號)	學校系所(中文)	學校系所(英文)	E-mail
蔡宗翰	Richard Tzong-Han Tsai	自然語言處理 (CE7024)	國立中央大學資訊工程學系	Department of Computer Science & Information Engineering, National Central University	thtsai@g.ncu.edu.tw