影像表格文字化的實作與學習

作者：薛詠謙

1、 動機

在尋找合適校系的過程中，我發現 University TW 所提供的分數搜尋功能非常實用。不過，該網站的資料更新速度並不一定與官方同步，而 **大學甄選入學委員會** 所公布的入取分數資料又多以 **圖片表格** 的形式呈現，難以直接整理或分析。

因此，我產生了想要將影像表格轉換成可數位化利用文字資料的想法，因此，我開始學習 **OCR**（光學文字辨識）技術，並嘗試實現影像中文字資料的自動化擷取與轉換。

2、 資料蒐集

根據 [1]，影像表格分析流程可分為以下幾個步驟：

1. **影像前處理**：對原始影像進行去噪、二值化或對比度等調整。
2. **表格分割**：偵測並標註表格的行列結構，將整體表格拆分為單元格。
3. **文字辨識**：對每個單元格進行 OCR 辨識，取得對應文字內容。
4. **轉換輸出格式**：將最終文字結果整理成結構化資料或表格文字格式。

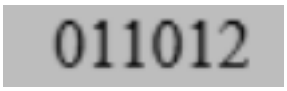
2.1 純文字辨識

2.1.1 `tesseract`

將中文文字影像直接輸入至 `tesseract` 進行辨識時，即便經過去除表格邊框等前處理步驟，仍未能有效提升辨識效果。

`tesseract figure-1.png - -l chi_tra`
無法 辨識出文字

不過，`tesseract` 在 **數字** 的辨識上表現十分精準，能穩定且無誤地擷取出 **校系代碼** 欄位的資料。



`tesseract figure-2.png - -c tessedit_char_whitelist=0123456789`
成功辨識出 011012

2.1.2 PaddleOCR

PaddleOCR 是百度開發的文字辨識工具，相較於 `tesseract`，在中文文字的擷取上具有顯著的提升。

Bibliography

[1] 蔡桓銘, “用 Tesseract 結合 LSTM 模型實作手填表格辨識,” Jan. 15, 2021. doi: 10.6846/TKU.2021.00596.