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** 所公布的錄取分數資料又多以 **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] 與 [2] ，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，如 `csv` 或 `json` 。

2.1 純文字辨識

2.1.1 `Tesseract`

將中文文字影像直接輸入 `tesseract` 進行辨識時，即使經過灰階化等前處理，辨識效果仍未明顯提升。推測原因可能在於大考中心提供的檔案 **解析度** 過低[3]，導致模型無法正確辨識文字。

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-1.png - -l chi_tra`
無法 辨識出文字

`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。

在大多數情況下，PaddleOCR 能夠成功辨識中文文字，其特色包括：

- 少數情況下可能漏字，例如在 Figure 1 中僅辨識出「(華語文教學組)」。
- 純數字辨識有時也會漏字，如在 Figure 2 中僅辨識出「1012」。
- 對中英文混合及標點符號具有良好容錯性，大部分情況下都能正確辨識，例如在 Figure 3 成功辨識出「(英文+數學 A)25」。

中國文學系乙組(華語文教學組) Figure 1:	011012 Figure 2:	(英文+數學A)25 Figure 3:

文獻參考

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

[2] Johnny Chang, “使用 python 萃取掃描文件中的表格(一)切豆腐篇.” [Online]. Available: https://between2058.medium.com/%E4%BD%BF%E7%94%A8python%E8%90%83%E5%8F%96%E6%8E%83%E6%8F%8F%E6%96%87%E4%BB%B6%E4%B8%AD%E7%9A%84%E8%A1%A8%E6%A0%BC-%E4%B8%80-%E5%88%87%E8%B1%86%E8%85%90%E7%AF%87-d5b65b7ec320

[3] Willus Dotkom, “Optimal image resolution (dpi/ppi) for Tesseract 4.0.0 and eng.traineddata?.” [Online]. Available: https://groups.google.com/g/tesseract-ocr/c/Wdh_JJwnw94/m/24JHDYQbBQAJ?pli=1