

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 影像預處理

2.1.1 二值化

`threshold(thresh: f64)` 函式可將灰階影像轉換為純黑與純白的二值化影像。然而，由於中文字筆劃粗細不一，閾值 (`thresh`) 的設定若不恰當，可能會降低文字辨識的準確度。

| 檢定標準 | 檢定標準 | 檢定標準 |
|------------------|------------------|------------------|
| Figure 1: 閾值=100 | Figure 2: 閾值=150 | Figure 3: 閾值=200 |

因此，雖然使用 Figure 2 的處理結果尚屬可接受，但在本次專案中，最終決定於文字辨識的預處理階段不採用二值化。

2.2 文字辨識

2.2.1 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。

人文社會學院學士班

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

011012

2.2.2 PaddleOCR

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 4 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 5 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 6 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 4: | 011012 Figure 5: | (英文+數學A)25 Figure 6: |
|------------------------------|---------------------|-------------------------|

綜上所述，將 **tesseract** 用於單元格的純數字辨識，而將 PaddleOCR 作為主要文字辨識工具，似乎為最佳組合。

3、操作方法

根據 Figure 7 所示，大考中心提供的篩選標準圖表可能包含多個區塊。因此，需要先辨識各區域的**外圍邊線**並進行分割，再辨識表格內的**分割線**，將表格切分成單元格，最後進行**文字辨識**，並將結果整理成結構化的 JSON 格式。

| 【续表：蔬菜产量表】 | | | | | | | | | | | | | | | |
|------------|--------|---------|-------|----------|---------|---------|-------|----------|---------|---------|-------|----------|---------|---------|-------|
| 【一】蔬菜类 | | 本市产量 | | 本市产量 | | | | 本市产量 | | | | 本市产量 | | 本市产量 | |
| 品名 | 单位 | 产量 | 比上年增长 | 品名 | 单位 | 产量 | 比上年增长 | 品名 | 单位 | 产量 | 比上年增长 | 品名 | 单位 | 产量 | 比上年增长 |
| 010000 | 蔬菜类 | 1000000 | 100.0 | 010100 | 叶类蔬菜 | 1000000 | 100.0 | 010200 | 茄果类蔬菜 | 1000000 | 100.0 | 010300 | 瓜豆类蔬菜 | 1000000 | 100.0 |
| 010100 | 叶类蔬菜 | 1000000 | 100.0 | 010101 | 白菜类 | 1000000 | 100.0 | 010102 | 甘蓝类 | 1000000 | 100.0 | 010103 | 芹菜类 | 1000000 | 100.0 |
| 010101 | 白菜类 | 1000000 | 100.0 | 01010101 | 大白菜 | 1000000 | 100.0 | 01010102 | 结球白菜 | 1000000 | 100.0 | 01010103 | 散叶白菜 | 1000000 | 100.0 |
| 010102 | 甘蓝类 | 1000000 | 100.0 | 01010201 | 花椰菜 | 1000000 | 100.0 | 01010202 | 西兰花 | 1000000 | 100.0 | 01010203 | 其他甘蓝类 | 1000000 | 100.0 |
| 010103 | 芹菜类 | 1000000 | 100.0 | 01010301 | 芹菜 | 1000000 | 100.0 | 01010302 | 其他芹菜类 | 1000000 | 100.0 | 01010303 | 其他芹菜类 | 1000000 | 100.0 |
| 010104 | 菠菜类 | 1000000 | 100.0 | 01010401 | 菠菜 | 1000000 | 100.0 | 01010402 | 其他菠菜类 | 1000000 | 100.0 | 01010403 | 其他菠菜类 | 1000000 | 100.0 |
| 010105 | 其他叶类蔬菜 | 1000000 | 100.0 | 01010501 | 茼蒿 | 1000000 | 100.0 | 01010502 | 茼蒿 | 1000000 | 100.0 | 01010503 | 茼蒿 | 1000000 | 100.0 |
| 010200 | 茄果类蔬菜 | 1000000 | 100.0 | 010201 | 茄果类蔬菜 | 1000000 | 100.0 | 010202 | 茄果类蔬菜 | 1000000 | 100.0 | 010203 | 茄果类蔬菜 | 1000000 | 100.0 |
| 010201 | 茄果类蔬菜 | 1000000 | 100.0 | 01020101 | 番茄 | 1000000 | 100.0 | 01020102 | 番茄 | 1000000 | 100.0 | 01020103 | 番茄 | 1000000 | 100.0 |
| 010202 | 茄果类蔬菜 | 1000000 | 100.0 | 01020201 | 茄子 | 1000000 | 100.0 | 01020202 | 茄子 | 1000000 | 100.0 | 01020203 | 茄子 | 1000000 | 100.0 |
| 010203 | 茄果类蔬菜 | 1000000 | 100.0 | 01020301 | 辣椒 | 1000000 | 100.0 | 01020302 | 辣椒 | 1000000 | 100.0 | 01020303 | 辣椒 | 1000000 | 100.0 |
| 010204 | 茄果类蔬菜 | 1000000 | 100.0 | 01020401 | 其他茄果类蔬菜 | 1000000 | 100.0 | 01020402 | 其他茄果类蔬菜 | 1000000 | 100.0 | 01020403 | 其他茄果类蔬菜 | 1000000 | 100.0 |
| 010205 | 茄果类蔬菜 | 1000000 | 100.0 | 01020501 | 其他茄果类蔬菜 | 1000000 | 100.0 | 01020502 | 其他茄果类蔬菜 | 1000000 | 100.0 | 01020503 | 其他茄果类蔬菜 | 1000000 | 100.0 |
| 010300 | 瓜豆类蔬菜 | 1000000 | 100.0 | 010301 | 瓜豆类蔬菜 | 1000000 | 100.0 | 010302 | 瓜豆类蔬菜 | 1000000 | 100.0 | 010303 | 瓜豆类蔬菜 | 1000000 | 100.0 |
| 010301 | 瓜豆类蔬菜 | 1000000 | 100.0 | 01030101 | 黄瓜 | 1000000 | 100.0 | 01030102 | 黄瓜 | 1000000 | 100.0 | 01030103 | 黄瓜 | 1000000 | 100.0 |
| 010302 | 瓜豆类蔬菜 | 1000000 | 100.0 | 01030201 | 冬瓜 | 1000000 | 100.0 | 01030202 | 冬瓜 | 1000000 | 100.0 | 01030203 | 冬瓜 | 1000000 | 100.0 |
| 010303 | 瓜豆类蔬菜 | 1000000 | 100.0 | 01030301 | 苦瓜 | 1000000 | 100.0 | 01030302 | 苦瓜 | 1000000 | 100.0 | 01030303 | 苦瓜 | 1000000 | 100.0 |
| 010304 | 瓜豆类蔬菜 | 1000000 | 100.0 | 01030401 | 其他瓜豆类蔬菜 | 1000000 | 100.0 | 01030402 | 其他瓜豆类蔬菜 | 1000000 | 100.0 | 01030403 | 其他瓜豆类蔬菜 | 1000000 | 100.0 |
| 010305 | 瓜豆类蔬菜 | 1000000 | 100.0 | 01030501 | 其他瓜豆类蔬菜 | 1000000 | 100.0 | 01030502 | 其他瓜豆类蔬菜 | 1000000 | 100.0 | 01030503 | 其他瓜豆类蔬菜 | 1000000 | 100.0 |
| 010400 | 水生蔬菜 | 1000000 | 100.0 | 010401 | 水生蔬菜 | 1000000 | 100.0 | 010402 | 水生蔬菜 | 1000000 | 100.0 | 010403 | 水生蔬菜 | 1000000 | 100.0 |
| 010401 | 水生蔬菜 | 1000000 | 100.0 | 01040101 | 茭白 | 1000000 | 100.0 | 01040102 | 茭白 | 1000000 | 100.0 | 01040103 | 茭白 | 1000000 | 100.0 |
| 010402 | 水生蔬菜 | 1000000 | 100.0 | 01040201 | 茭白 | 1000000 | 100.0 | 01040202 | 茭白 | 1000000 | 100.0 | 01040203 | 茭白 | 1000000 | 100.0 |
| 010403 | 水生蔬菜 | 1000000 | 100.0 | 01040301 | 茭白 | 1000000 | 100.0 | 01040302 | 茭白 | 1000000 | 100.0 | 01040303 | 茭白 | 1000000 | 100.0 |
| 010404 | 水生蔬菜 | 1000000 | 100.0 | 01040401 | 茭白 | 1000000 | 100.0 | 01040402 | 茭白 | 1000000 | 100.0 | 01040403 | 茭白 | 1000000 | 100.0 |
| 010405 | 水生蔬菜 | 1000000 | 100.0 | 01040501 | 茭白 | 1000000 | 100.0 | 01040502 | 茭白 | 1000000 | 100.0 | 01040503 | 茭白 | 1000000 | 100.0 |
| 010406 | 水生蔬菜 | 1000000 | 100.0 | 01040601 | 茭白 | 1000000 | 100.0 | 01040602 | 茭白 | 1000000 | 100.0 | 01040603 | 茭白 | 1000000 | 100.0 |
| 010407 | 水生蔬菜 | 1000000 | 100.0 | 01040701 | 茭白 | 1000000 | 100.0 | 01040702 | 茭白 | 1000000 | 100.0 | 01040703 | 茭白 | 1000000 | 100.0 |
| 010408 | 水生蔬菜 | 1000000 | 100.0 | 01040801 | 茭白 | 1000000 | 100.0 | 01040802 | 茭白 | 1000000 | 100.0 | 01040803 | 茭白 | 1000000 | 100.0 |
| 010409 | 水生蔬菜 | 1000000 | 100.0 | 01040901 | 茭白 | 1000000 | 100.0 | 01040902 | 茭白 | 1000000 | 100.0 | 01040903 | 茭白 | 1000000 | 100.0 |
| 010410 | 水生蔬菜 | 1000000 | 100.0 | 01041001 | 茭白 | 1000000 | 100.0 | 01041002 | 茭白 | 1000000 | 100.0 | 01041003 | 茭白 | 1000000 | 100.0 |
| 010411 | 水生蔬菜 | 1000000 | 100.0 | 01041101 | 茭白 | 1000000 | 100.0 | 01041102 | 茭白 | 1000000 | 100.0 | 01041103 | 茭白 | 1000000 | 100.0 |
| 010412 | 水生蔬菜 | 1000000 | 100.0 | 01041201 | 茭白 | 1000000 | 100.0 | 01041202 | 茭白 | 1000000 | 100.0 | 01041203 | 茭白 | 1000000 | 100.0 |
| 010413 | 水生蔬菜 | 1000000 | 100.0 | 01041301 | 茭白 | 1000000 | 100.0 | 01041302 | 茭白 | 1000000 | 100.0 | 01041303 | 茭白 | 1000000 | 100.0 |
| 010414 | 水生蔬菜 | 1000000 | 100.0 | 01041401 | 茭白 | 1000000 | 100.0 | 01041402 | 茭白 | 1000000 | 100.0 | 01041403 | 茭白 | 1000000 | 100.0 |
| 010415 | 水生蔬菜 | 1000000 | 100.0 | 01041501 | 茭白 | 1000000 | 100.0 | 01041502 | 茭白 | 1000000 | 100.0 | 01041503 | 茭白 | 1000000 | 100.0 |
| 010416 | 水生蔬菜 | 1000000 | 100.0 | 01041601 | 茭白 | 1000000 | 100.0 | 01041602 | 茭白 | 1000000 | 100.0 | 01041603 | 茭白 | 1000000 | 100.0 |
| 010417 | 水生蔬菜 | 1000000 | 100.0 | 01041701 | 茭白 | 1000000 | 100.0 | 01041702 | 茭白 | 1000000 | 100.0 | 01041703 | 茭白 | 1000000 | 100.0 |
| 010418 | 水生蔬菜 | 1000000 | 100.0 | 01041801 | 茭白 | 1000000 | 100.0 | 01041802 | 茭白 | 1000000 | 100.0 | 01041803 | 茭白 | 1000000 | 100.0 |
| 010419 | 水生蔬菜 | 1000000 | 100.0 | 01041901 | 茭白 | 1000000 | 100.0 | 01041902 | 茭白 | 1000000 | 100.0 | 01041903 | 茭白 | 1000000 | 100.0 |
| 010420 | 水生蔬菜 | 1000000 | 100.0 | 01042001 | 茭白 | 1000000 | 100.0 | 01042002 | 茭白 | 1000000 | 100.0 | 01042003 | 茭白 | 1000000 | 100.0 |
| 010421 | 水生蔬菜 | 1000000 | 100.0 | 01042101 | 茭白 | 1000000 | 100.0 | 01042102 | 茭白 | 1000000 | 100.0 | 01042103 | 茭白 | 1000000 | 100.0 |
| 010422 | 水生蔬菜 | 1000000 | 100.0 | 01042201 | 茭白 | 1000000 | 100.0 | 01042202 | 茭白 | 1000000 | 100.0 | 01042203 | 茭白 | 1000000 | 100.0 |
| 010423 | 水生蔬菜 | 1000000 | 100.0 | 01042301 | 茭白 | 1000000 | 100.0 | 01042302 | 茭白 | 1000000 | 100.0 | 01042303 | 茭白 | 1000000 | 100.0 |
| 010424 | 水生蔬菜 | 1000000 | 100.0 | 01042401 | 茭白 | 1000000 | 100.0 | 01042402 | 茭白 | 1000000 | 100.0 | 01042403 | 茭白 | 1000000 | 100.0 |
| 010425 | 水生蔬菜 | 1000000 | 100.0 | 01042501 | 茭白 | 1000000 | 100.0 | 01042502 | 茭白 | 1000000 | 100.0 | 01042503 | 茭白 | 1000000 | 100.0 |
| 010426 | 水生蔬菜 | 1000000 | 100.0 | 01042601 | 茭白 | 1000000 | 100.0 | 01042602 | 茭白 | 1000000 | 100.0 | 01042603 | 茭白 | 1000000 | 100.0 |
| 010427 | 水生蔬菜 | 1000000 | 100.0 | 01042701 | 茭白 | 1000000 | 100.0 | 01042702 | 茭白 | 1000000 | 100.0 | 01042703 | 茭白 | 1000000 | 100.0 |
| 010428 | 水生蔬菜 | 1000000 | 100.0 | 01042801 | 茭白 | 1000000 | 100.0 | 01042802 | 茭白 | 1000000 | 100.0 | 01042803 | 茭白 | 1000000 | 100.0 |
| 010429 | 水生蔬菜 | 1000000 | 100.0 | 01042901 | 茭白 | 1000000 | 100.0 | 01042902 | 茭白 | 1000000 | 100.0 | 01042903 | 茭白 | 1000000 | 100.0 |
| 010430 | 水生蔬菜 | 1000000 | 100.0 | 01043001 | 茭白 | 1000000 | 100.0 | 01043002 | 茭白 | 1000000 | 100.0 | 01043003 | 茭白 | 1000000 | 100.0 |
| 010431 | 水生蔬菜 | 1000000 | 100.0 | 01043101 | 茭白 | 1000000 | 100.0 | 01043102 | 茭白 | 1000000 | 100.0 | 01043103 | 茭白 | 1000000 | 100.0 |
| 010432 | 水生蔬菜 | 1000000 | 100.0 | 01043201 | 茭白 | 1000000 | 100.0 | 01043202 | 茭白 | 1000000 | 100.0 | 01043203 | 茭白 | 1000000 | 100.0 |
| 010433 | 水生蔬菜 | 1000000 | 100.0 | 01043301 | 茭白 | 1000000 | 100.0 | 01043302 | 茭白 | 1000000 | 100.0 | 01043303 | 茭白 | 1000000 | 100.0 |
| 010434 | 水生蔬菜 | 1000000 | 100.0 | 01043401 | 茭白 | 1000000 | 100.0 | 01043402 | 茭白 | 1000000 | 100.0 | 01043403 | 茭白 | 1000000 | 100.0 |
| 010435 | 水生蔬菜 | 1000000 | 100.0 | 01043501 | 茭白 | 1000000 | 100.0 | 01043502 | 茭白 | 1000000 | 100.0 | 01043503 | 茭白 | 1000000 | 100.0 |
| 010436 | 水生蔬菜 | 1000000 | 100.0 | 01043601 | 茭白 | 1000000 | 100.0 | 01043602 | 茭白 | 1000000 | 100.0 | 01043603 | 茭白 | 1000000 | 100.0 |
| 010437 | 水生蔬菜 | 1000000 | 100.0 | 01043701 | 茭白 | 1000000 | 100.0 | 01043702 | 茭白 | 1000000 | 100.0 | 01043703 | 茭白 | 1000000 | 100.0 |
| 010438 | 水生蔬菜 | 1000000 | 100.0 | 01043801 | 茭白 | 1000000 | 100.0 | 01043802 | 茭白 | 1000000 | 100.0 | 01043803 | 茭白 | 1000000 | 100.0 |
| 010439 | 水生蔬菜 | 1000000 | 100.0 | 01043901 | 茭白 | 1000000 | 100.0 | 01043902 | 茭白 | 1000000 | 100.0 | 01043903 | 茭白 | 1000000 | 100.0 |
| 010440 | 水生蔬菜 | 1000000 | 100.0 | 01044001 | 茭白 | 1000000 | 100.0 | 01044002 | 茭白 | 1000000 | 100.0 | 01044003 | 茭白 | 1000000 | 100.0 |
| 010441 | 水生蔬菜 | 1000000 | 100.0 | 01044101 | 茭白 | 1000000 | 100.0 | 01044102 | 茭白 | 1000000 | 100.0 | 01044103 | 茭白 | 1000000 | 100.0 |
| 010442 | 水生蔬菜 | 1000000 | 100.0 | 01044201 | 茭白 | 1000000 | 100.0 | 01044202 | 茭白 | 1000000 | 100.0 | 01044203 | 茭白 | 1000000 | 100.0 |
| 010443 | 水生蔬菜 | 1000000 | 100.0 | 01044301 | 茭白 | 1000000 | 100.0 | 01044302 | 茭白 | 1000000 | 100.0 | 01044303 | 茭白 | 1000000 | 100.0 |
| 010444 | 水生蔬菜 | 1000000 | 100.0 | 01044401 | 茭白 | 1000000 | 100.0 | 01044402 | 茭白 | 1000000 | 100.0 | 01044403 | 茭白 | 1000000 | 100.0 |
| 010445 | 水生蔬菜 | 1000000 | 100.0 | 01044501 | 茭白 | 1000000 | 100.0 | 01044502 | 茭白 | 1000000 | 100.0 | 01044503 | 茭白 | 1000000 | 100.0 |
| 010446 | 水生蔬菜 | 1000000 | 100.0 | 01044601 | 茭白 | 1000000 | 100.0 | 01044602 | 茭白 | 1000000 | 100.0 | 01044603 | 茭白 | 1000000 | 100.0 |
| 010447 | 水生蔬菜 | 1000000 | 100.0 | 01044701 | 茭白 | 1000000 | 100.0 | 01044702 | 茭白 | 1000000 | 100.0 | 01044703 | 茭白 | 1000000 | 100.0 |
| 010448 | 水生蔬菜 | 1000000 | 100.0 | 01044801 | 茭白 | 1000000 | 100.0 | 01044802 | 茭白 | 1000000 | 100.0 | 01044803 | 茭白 | 1000000 | 100.0 |
| 010449 | 水生蔬菜 | 1000000 | 100.0 | 01044901 | 茭白 | 1000000 | 100.0 | 01044902 | 茭白 | 1000000 | 100.0 | 01044903 | 茭白 | 1000000 | 100.0 |
| 010450 | 水生蔬菜 | 1000000 | 100.0 | 01045001 | 茭白 | 1000000 | 100.0 | 01045002 | 茭白 | 1000000 | 100.0 | 01045003 | 茭白 | 1000000 | 100.0 |
| 010451 | 水生蔬菜 | 1000000 | 100.0 | 01045101 | 茭白 | 1000000 | 100.0 | 01045102 | 茭白 | 1000 | | | | | |

Figure 7:

處理流程可表述如下：

外圍邊線辨識 → 區域分割 → 表格邊線辨識 → 單元格分割 → 文字辨識 → 將資料結構化

3.1 外圍邊線辨識與分割

3.2 表格邊線辨識

3.3 單元格分割

3.4 資料結構化

參考文獻

- [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](https://doi.org/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