OCR Tool (ko_screenshot)

Description

This tool is used to extract text from images and output customer data. The purpose is to assist the daily work of the business team by getting information in pictures. It uses the OCR (Azure AI Vision Audio) to perform the extraction by automation (Selenium).

Update Log

- V1.4 2023/10/24:
  - Files are sorted in order
  - Handle “NIL” cases in content
  - Chinese characters are seen as ‘fake numbers’, this will influence the output of the telephone number. It has been fixed.

- V2.0 2024/01/22 (Developing):
  - Open the EPRC website

Current EPRC scrapping solution (24/01/2024)

As it is not possible to scrap data from ECPR web, the current approach to perform automation would be done by the “pyautogui” library.

Steps:

1. The application would create a Google Chrome instance with the ECPR website.

2. There will be a pop-up message box. Please ignore the box, and then log in + search as usual.

3. After searching the query, please click the message box to continue.

4. The automation will start and finish ultimately.

5. The photos will be stored in the predefined folder address.

Current Problem (24/01/2024)

1. Web Scrapping issues

Somehow cannot 'web scrap' the data, maybe the EPRC web has some anti-web-scrap approaches. In this case, automation may not be applied easily. A Possible approach is to use the pyautogui library, but it is not ideal as the HTML elements are not guaranteed to be the same.

2. Resolution problems

The screen resolution does not matter. Since the contacts are generated as images. The images have fixed sizes and resolutions. After testing, when Google Chrome with a 50% zoom level on the desktop.

3. Null values

This problem can be solved if the images can be downloaded (Prior to solving problem 1). Otherwise, This problem may need to be checked manually.

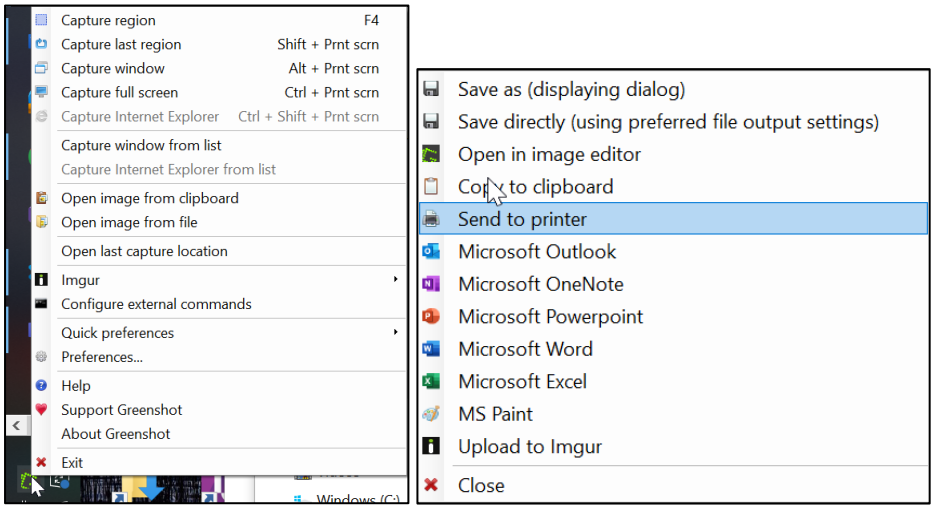
4. Hardcoded Pyautogui Code (A stupid way)

The potential "risk" of Pyautogui is that it cannot handle unexpected cases. Thus, please do not move the mouse unless there is a pop-up message box.

Preparation (Screenshots)

1. It is strongly suggested to use the Greenshot to capture the screenshots. The Greenshot is free and provides useful functions such as capturing the screenshots of the last region, magnifying glass, and saving directly in the preferred directory. It also provides the function of capturing the screenshots by pressing the 'Print Screen' button. Please make sure the Greenshots logo is on your toolbar before you start using the application.

   The Greenshot can be downloaded from the following link:
   https://getgreenshot.org



   (menu in the toolbar, possible output options)

2. When capturing the screenshots, please make sure that the screenshots are clear and that other objects do not cover the text (Please note that the mouse will stay at the screen when using GreenShots).

3. Also, the addresses should be shown as a whole in the screenshot. (e.g. if the address is shown in two lines in the screenshot, the output data will have problems).

4. When there is empty space, it will affect the whole output. To prevent that, please add some text (like 'Not yet tested' in the left picture above). The screenshots should be in the format of '.png' or '.jpg'.

   Example of valid screenshots:

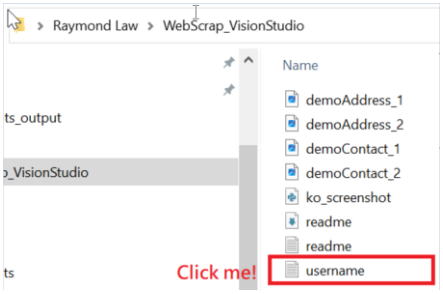
黎屋村(石崗)村屋/丁屋 / LAI UK TSUEN(SHEK KONG) VILLAGE HSE	
YOHO HUB (第B期) (第01座) / YOHO HUB PH B TWR 01	
坪輋村(坪輋)平房 / PING CHE TSUEN(PING CHE) BUNGALOW	
樟樹灘村(樟樹灘)村屋/丁屋 / CHEUNG SHUE TAN TSUEN(CHEUNG SHUE TAN) VILLAGE HSE	
峻巒 / PARK VISTA	
德寶大廈 / TAK PO BLDG	
青磚圍 村屋/丁屋 / TSING CHUEN WAI VILLAGE HSE	
文閣村(上水圍)村屋/丁屋 / MAN KOK VILLAGE(SHEUNG SHUI WAI) VILLAGE HSE	
運通洋樓 / WINNING HTS	
愛琴灣 / AEGEAN	
瓏門 (第一期) (第05座) / CENTURY GATEWAY PH 01 TWR 05	

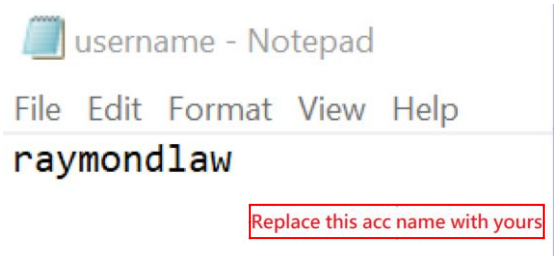
黎小姐/MISS LAI	93603100
黎太太/MRS LAI	90476567
盧小姐/MISS LO	60938071
盧先生/MR LO	91606151
蕭小姐/MISS SIU	61818627
謝小姐/MISS TSE	96603240
魏先生/MR NGAI	55238078
羅小姐/MISS LAW	65336886
羅先生/MR LAW	94618449
嚴小姐/MISS YIM	94127713
Not yet tested	62212759

5. You may input any number of screenshots in the folders ‘Greenshots_address’ and ‘Greenshots_contact’. However the output CSV file will leave the part as null if the screenshot does not provide the complete information. (e.g. if you do not input any photos in the folder ‘Greenshots_address’, the output csv file will only contain the information of ‘Contact’, ‘Telephone Number’ and ‘Page’.)

Preparation (Others)

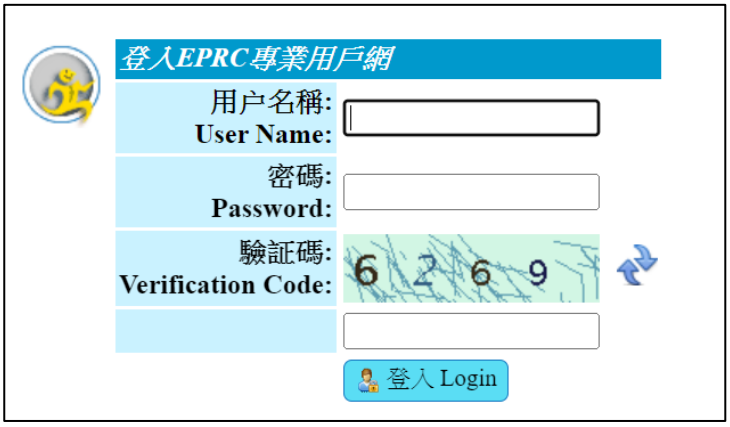
1. Please edit the text in the username.txt. Replace the ‘raymondlaw’ and input your username, which is the account name when you log in to the computer.



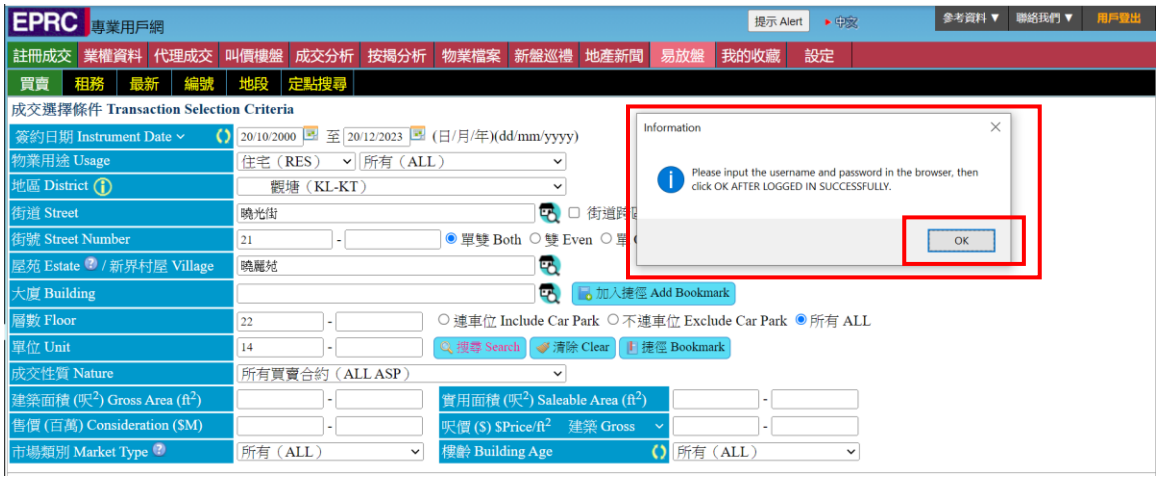


HOW TO USE

1. The application would open the EPRC login page on Google Chrome. It requires the user to **<u>manually input</u>** the account and password (Fig. 1a). The reason is that the verification code cannot be filled by automation. After successfully logged in, please click the 'OK' for the alert dialog.



(Fig. 1a)



(Fig. 1b)

2. The application will check the files in the pre-defined locations.If this is the first time to run the application, or the predefined directories has not been created, the application will automatically add following directories in 'C:{username}' with some **demo pictures** for testing:

   - ../Pictures/Greenshots_input
     - ../Pictures/Greenshots_input/Greenshots_address
     - ../Pictures/Greenshots_input/Greenshots_address
   - ../Pictures/Greenshots_output

- **(Please do not move the mouse cursor and input by keyboard before the browser is closed.)**

3. Before running the application, please follow the instructions in the 'Preparation' section above. The system will also check the files in the pre-defined locations.

4. The application will run the OCR (Azure AI Vision Audio) in the website to perform the extraction by autiomation (Selenium). Please do not move the mouse cursor and input by keyboard before the browser is closed.

5. The application will save the output file in the '../Pictures/Greenshots_output' and it will be opened after application is finished.

6. The pictures in the Greenshots_address and Greenshots_contact folders will be removed and duplicated to the new folders.

Example Output

- The ouput file has 5 columns: 'Chinese Address','English Address', Contact, 'Telephone Number', 'Page'
- The output file will be saved as a csv file in the '../Pictures/Greenshots_output'
- The output file will be opened after application is finished.
- **Please note that the Chinese address result is not accurate. The result is only for reference and the user should check the result by themselves.**

Error Handling

- If you see any error, please refer to the instructions of this file and check the following points:
  - The screenshots are prepared in the right folders
  - The screenshots are in the format of '.png' or '.jpg'.
  - The screenshots are in the folder 'Greenshots_address' and 'Greenshots_contact'.
- **Please do not move the mouse cursor and input by keyboard before the browser is closed.**
- If you find any bugs, please contact the developer (Raymond Law).