Text Recognition and processing using PyTesseact and OCR

# Steps to execute the program

1. Install the latest version of python if not already installed by going to <https://www.python.org/downloads/>
2. Make sure to check “Add Python X.X.X to PATH” during installation to python to PATH.
3. Verify the python installation:
   1. Open command prompt
   2. Type > *python –version*
   3. You should see an output such as: Python X.X.X where X.X.X is the version of python installed.
4. Install Tesseract using windows installer by downloading the same from <https://github.com/UB-Mannheim/tesseract/wiki>. You can choose a 32/64-bit installation depending on the system.
5. Make sure the installation path is set to “C:\Program Files\Tesseract-OCR” and install the software.
6. Download the zip file from GitHub repository.
7. Extract the file to a specific directory. Keep a note of the directory where the zip file is extracted.
8. Open a new command prompt and navigate to the directory where the zip file was extracted.
   1. For example, if the zip was extracted to Downloads, type in

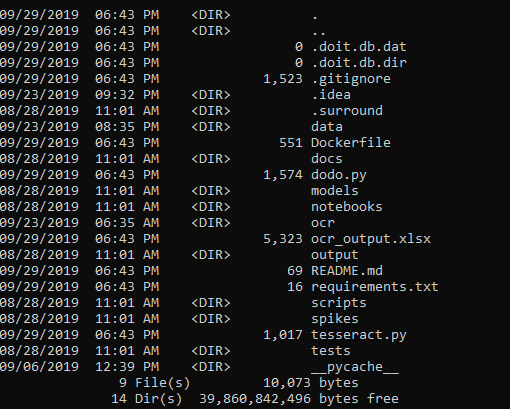
*> cd C:\Users\USERNAME\Downloads*

Insert the username of your system in place of USERNAME. If your system username is Deakin, the path becomes C:\Users\Deakin\Documents.

1. Once there, type in below command and hit enter.

*> cd ocr*

1. Type in > *dir* and hit enter. You should see a file structure similar to below:



1. Depending upon how the zip file was extracted, you could instead only see a single folder ocr. In that case, type in > *cd ocr* again in the command prompt and hit enter followed by *> dir* command to get the output same as above.
2. Type in the below command to install the required dependencies. Wait for the installations to finish.

*> pip install -r requirements.txt*

1. Once all dependencies are installed, type *> python -m ocr*
2. You should see a small interface with Browse button.
3. Click on the button to open the file dialog
4. Navigate to path source code directory and go to data/bills folder. It contains sample bill images.
   1. In this case, the path would be C:\Users\Deakin\Documents\ocr\data\bills
5. Choose one or many images to pass to the program by pressing and holding control button and choosing multiple images or dragging the cursor across multiple images.
6. Click Ok once all the images are selected.
7. The algorithm scans the images, extracts the bill totals and writes the output to an excel sheet which is saved in the same folder as the source code. The path of the excel is displayed on the interface. In this case it would be C:\Users\Deakin\Documents\ocr\ocr\_output.xlsx