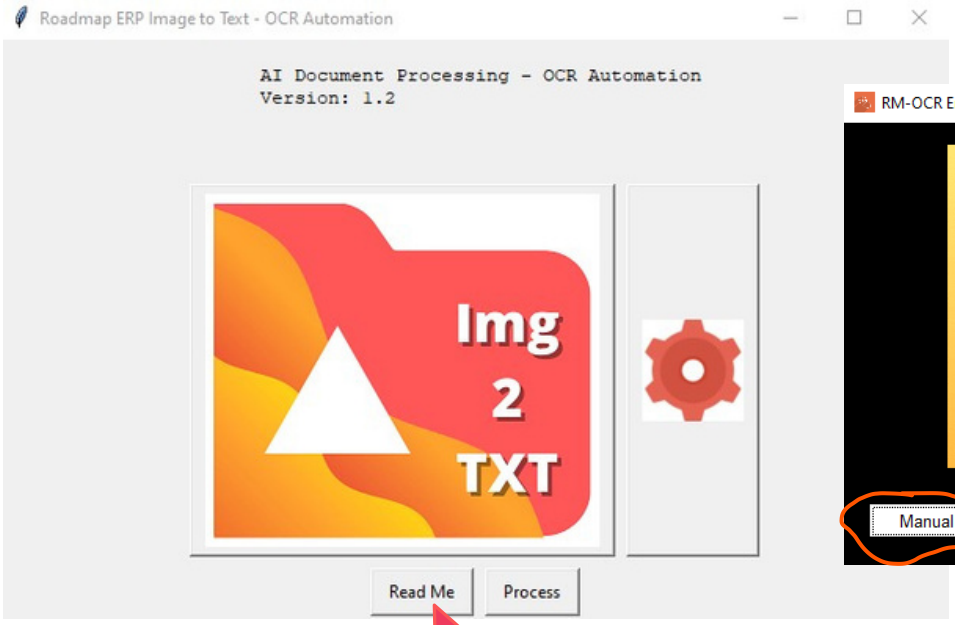
RM-I2T

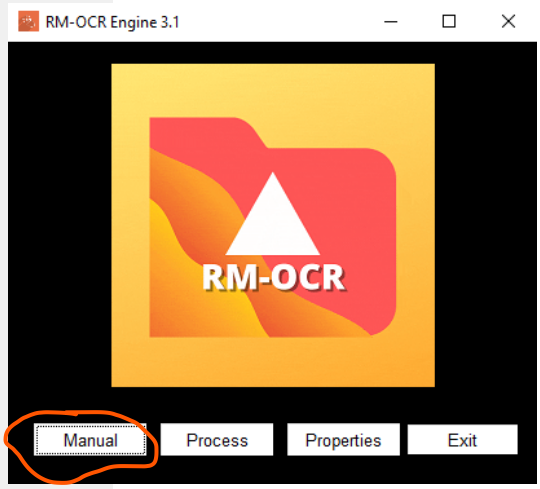
Guidelines

USER GUIDE

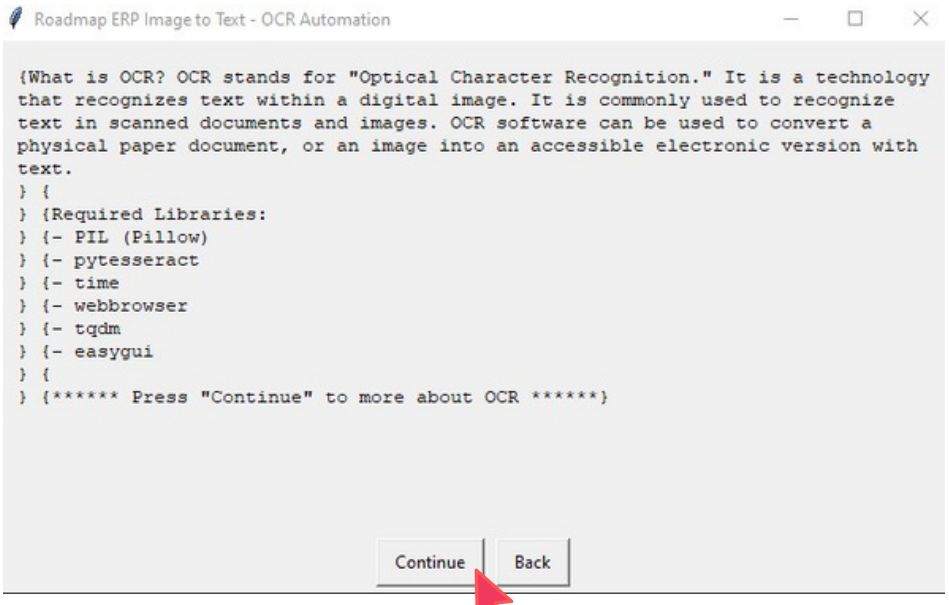
Roadmap ERP

Mainmenu:





Steps to read User Guide



These are some default libraries required to be installed in python environment.

Continue to read entire user guide this will open-up the following Pdf file in default web-browser of your system

RM-I2T
Guidelines

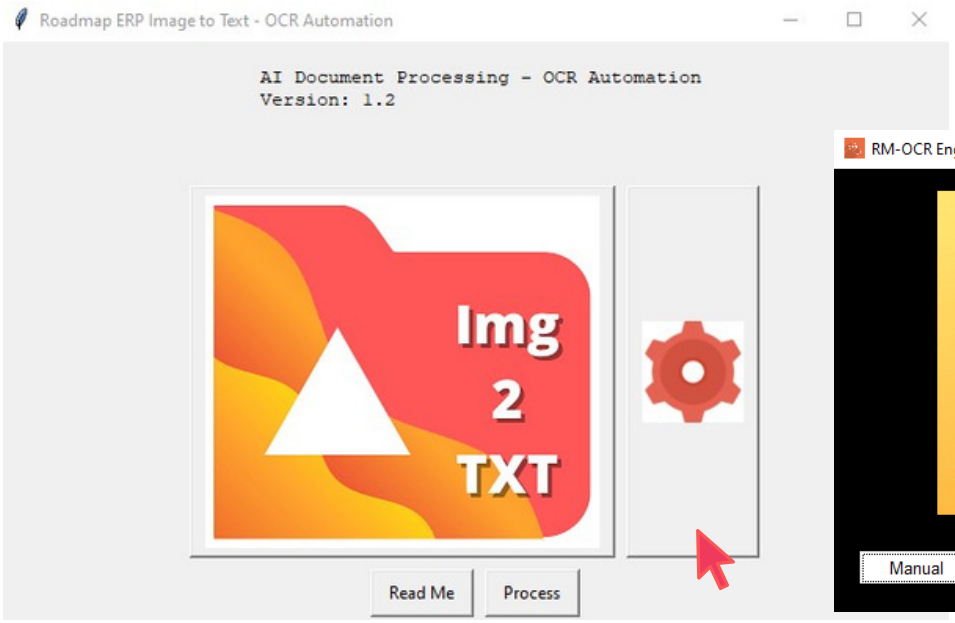
USER GUIDE Roadmap ERP

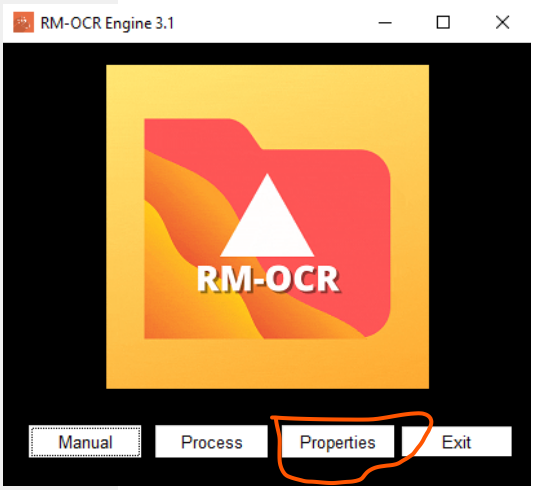
Update Tesseract

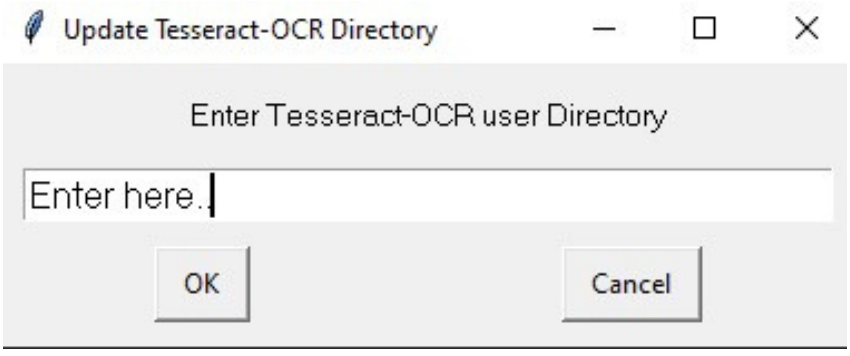
tesseract-ocr-w64-setup-v5.0.1.20220118

Install this application in c drive :

C:\Users\~~Ganga Babu.M~~\AppData\Local\Programs\Tesseract-OCR\tesseract.exe

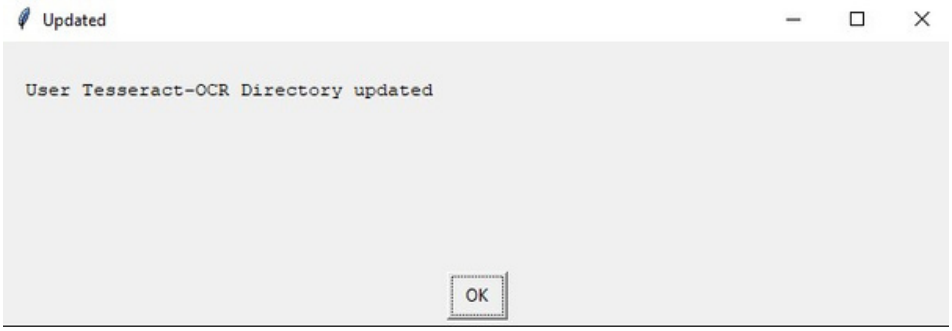




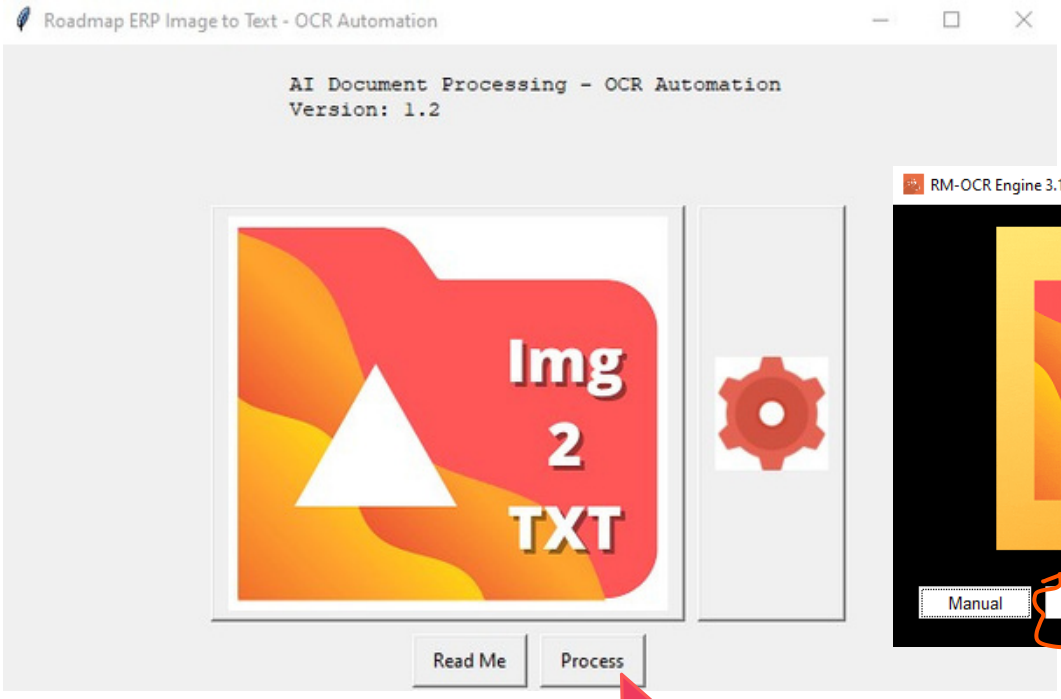


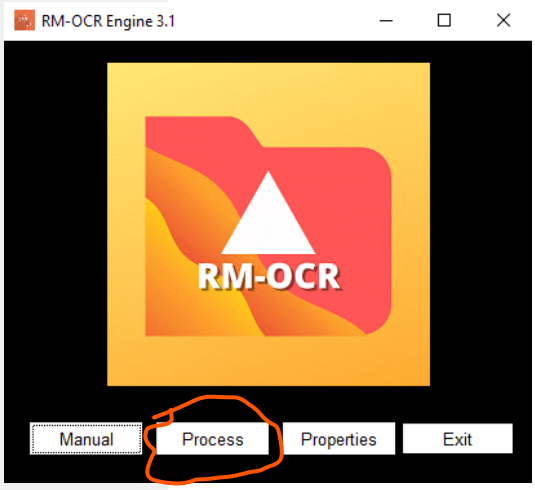
Replace the field with tesseract.exe path of your system

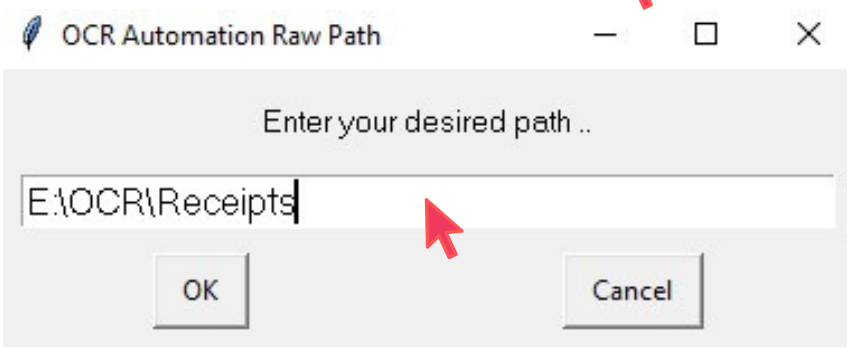
C:\Users\~~Ganga Babu.M~~\AppData\Local\Programs\Tesseract-OCR\tesseract.exe



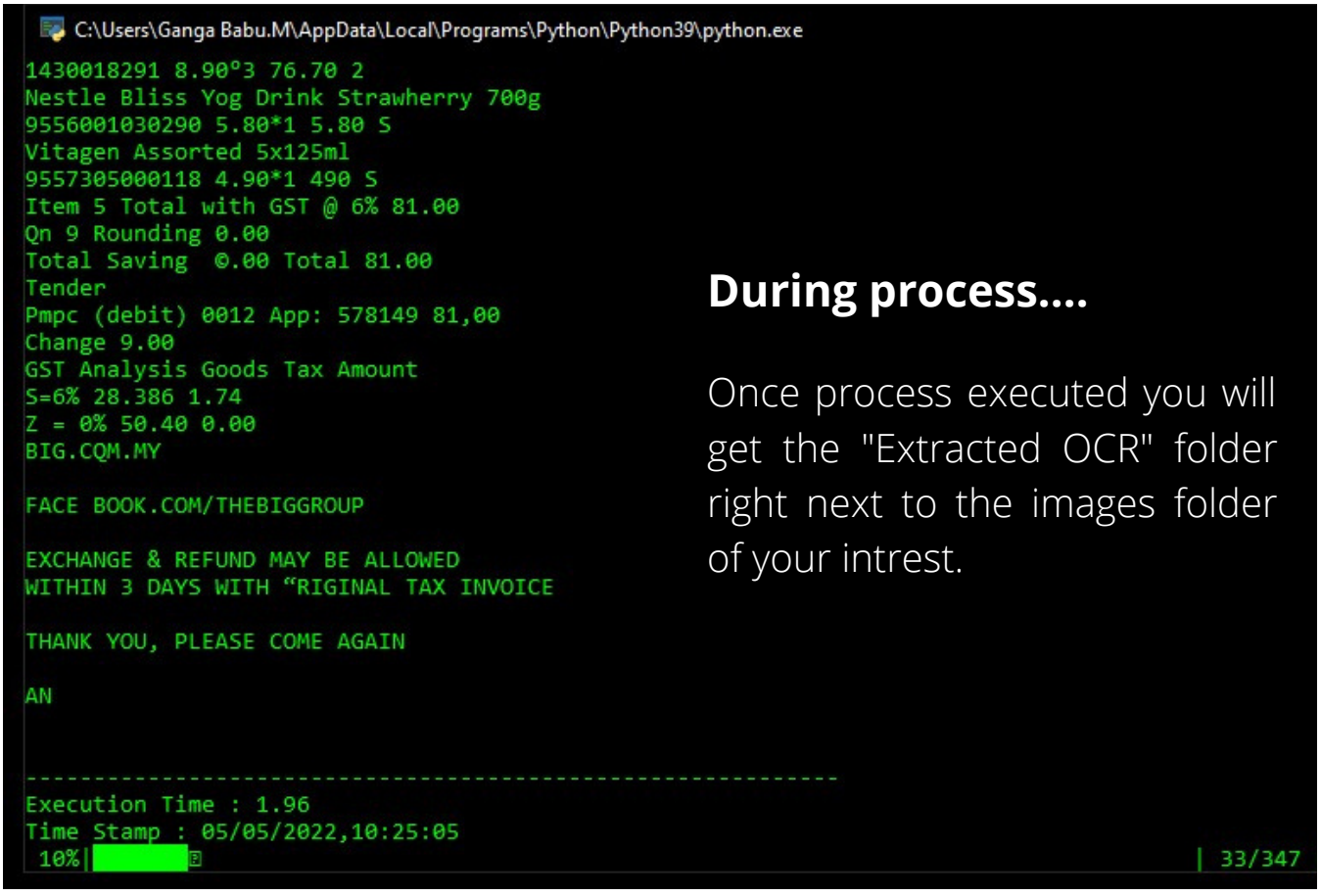
Processing Image to Text :







update the field with desired path to extract the text from the images



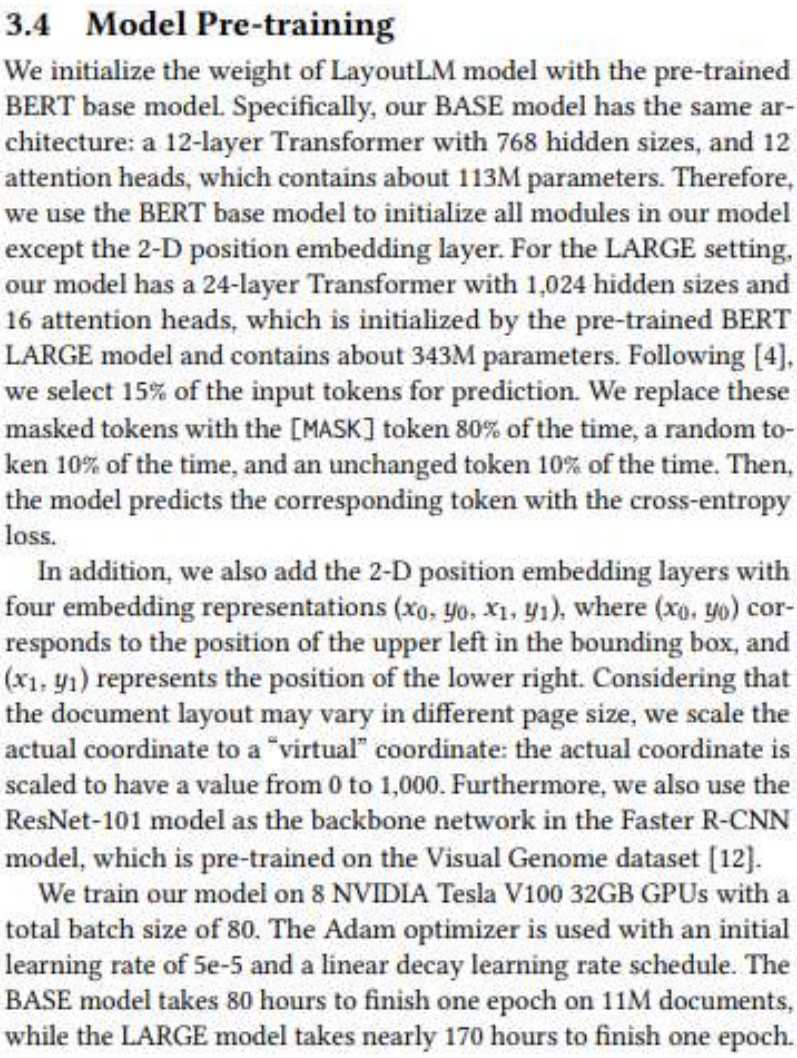
During process....

Once process executed you will get the "Extracted OCR" folder right next to the images folder of your intrest.

Note:

There are few procedures to select the images for OCR and other text extraction from various bill, invoice or document images. To use pytesseract the minimum DPI (Dots per Inch) should be at least 300.

Good samples:

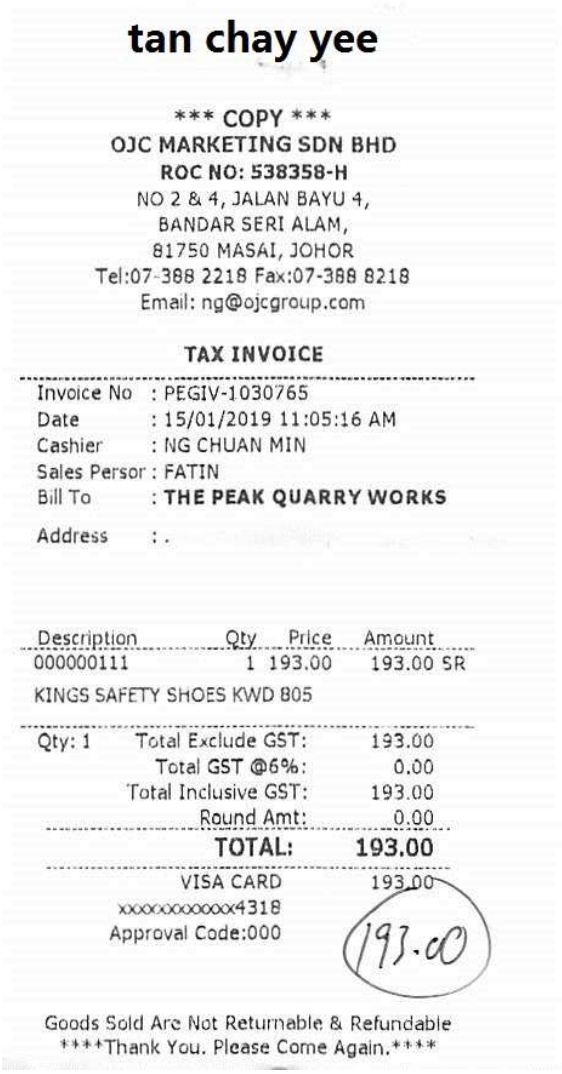


3.4 Model Pre-training

We initialize the weight of LayoutLM model with the pre-trained BERT base model. Specifically, our BASE model has the same architecture: a 12-layer Transformer with 768 hidden sizes, and 12 attention heads, which contains about 113M parameters. Therefore, we use the BERT base model to initialize all modules in our model except the 2-D position embedding layer. For the LARGE setting, our model has a 24-layer Transformer with 1,024 hidden sizes and 16 attention heads, which is initialized by the pre-trained BERT LARGE model and contains about 343M parameters. Following [4], we select 15% of the input tokens for prediction. We replace these masked tokens with the [MASK] token 80% of the time, a random token 10% of the time, and an unchanged token 10% of the time. Then, the model predicts the corresponding token with the cross-entropy loss.

In addition, we also add the 2-D position embedding layers with four embedding representations (x_0, y_0, x_1, y_1), where (x_0, y_0) corresponds to the position of the upper left in the bounding box, and (x_1, y_1) represents the position of the lower right. Considering that the document layout may vary in different page size, we scale the actual coordinate to a "virtual" coordinate: the actual coordinate is scaled to have a value from 0 to 1,000. Furthermore, we also use the ResNet-101 model as the backbone network in the Faster R-CNN model, which is pre-trained on the Visual Genome dataset [12].

We train our model on 8 NVIDIA Tesla V100 32GB GPUs with a total batch size of 80. The Adam optimizer is used with an initial learning rate of 5e-5 and a linear decay learning rate schedule. The BASE model takes 80 hours to finish one epoch on 11M documents, while the LARGE model takes nearly 170 hours to finish one epoch.



tan chay yee

*** COPY ***
OJC MARKETING SDN BHD
ROC NO: 538358-H
NO 2 & 4, JALAN BAYU 4,
BANDAR SERI ALAM,
81750 MASAI, JOHOR
Tel:07-388 2218 Fax:07-388 8218
Email: ng@ojcgroup.com

TAX INVOICE

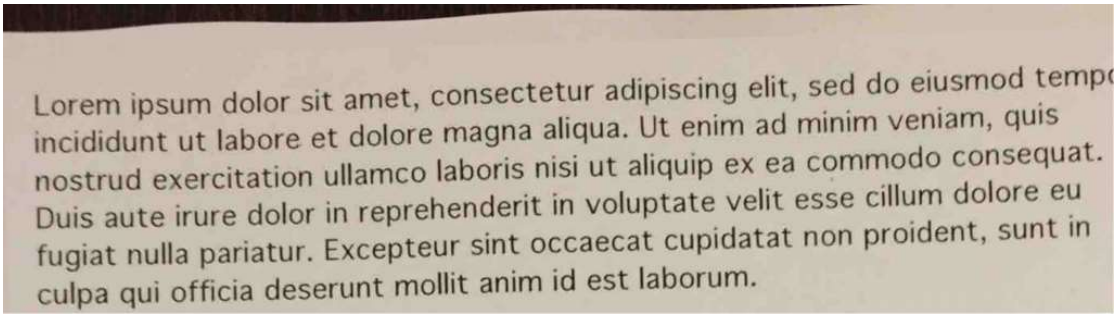
Invoice No : PEGIV-1030765
Date : 15/01/2019 11:05:16 AM
Cashier : NG CHUAN MIN
Sales Persor : FATIN
Bill To : **THE PEAK QUARRY WORKS**
Address : .

Description	Qty	Price	Amount
000000111	1	193.00	193.00 SR
KINGS SAFETY SHOES KWD 805			

Qty: 1	Total Exclude GST:	193.00
	Total GST @6%:	0.00
	Total Inclusive GST:	193.00
	Round Amt:	0.00
	TOTAL:	**193.00**
	VISA CARD	193.00
	xxxxxxxxxxxx4318	
	Approval Code:000	

193.00

Goods Sold Are Not Returnable & Refundable
****Thank You. Please Come Again.****



Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempo incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

an input object image into a sequenc
usually essential. For example, Grav
set of geometrical or image features f
while Su and Lu [33] convert word i
HOG features. The preprocessing s
the subsequent components in the pip
systems based on RNN can not be tra

Bad samples:

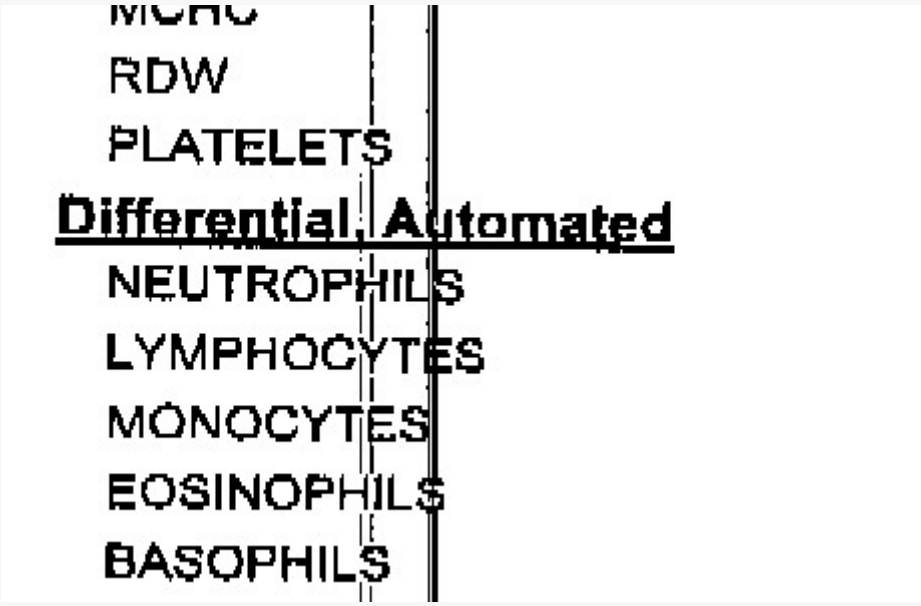
I've run into a few quirks with the 64 bit version of the LMDE (respin) RC and found the solutions as well either from other threads, comments on the .iso test posts on the Commun website or ones I discovered myself. Not all might experience every one of these quirks bu for those that do:

[i]NOTE: All the below solutions have worked me without a hitch. Just thought I ought to p them all in one post. Will update if I found more.

EXAMPLE 1: VERTICAL SCAN LINES

The image below is a section of a scanned image that has a set of vertical lines all the way down the left side. One of the data elements can be used to demonstrate the effect these vertical line has on the whole image. In bold below is the raw OCR data of what was captured where NEUTROPHILS is present in the image.

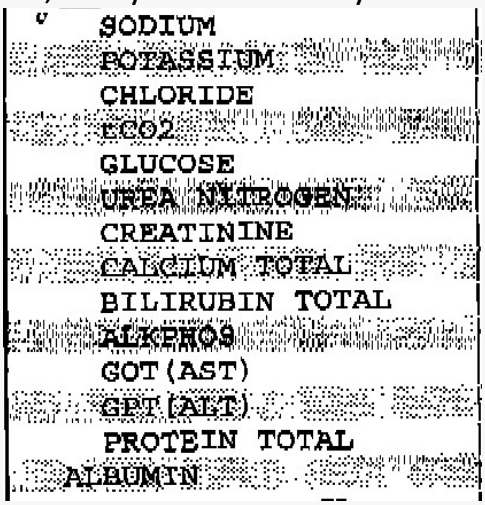
OCR output: NEUTROP I /111



This problem is generally attributed to a poor scanner and can be remedied by either fixing or replacing the device.

EXAMPLE 2: SHADED LINES

Shaded lines are great for creating contrast between lines in a spreadsheet. They make reading the data incredibly easy for a human being, unfortunately as easy as they make it for a human they make it difficult for a computer. Take the example below, each alternating line is easy to determine from another, and you could easily follow one row across the sheet.



When zooming in on the line containing UREA NITROGEN it becomes obvious why an OCR engine might have a problem determining what letters are contained within. All of the little lines used to create the grey effect overlap with and get in the way of the letters themselves.

Digging in to the OCR output from that line and the line below perfectly highlights how this can have an effect on capture accuracy. Creatinine is easily recognized but there is no way to determine that the value above is urea nitrogen. See below for the OCR output of those two lines.

OCR output:

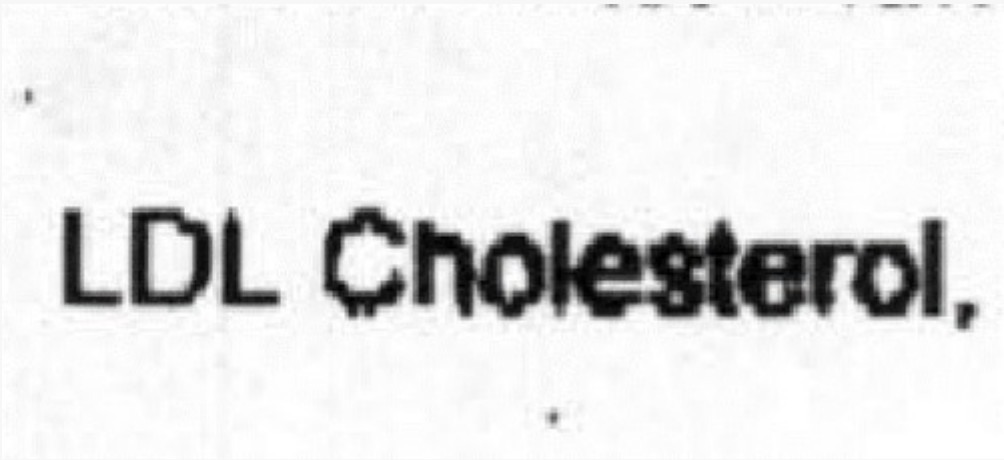
°I kg%#.0000K05*

CREATININE

This particular capture issue can be very difficult to work around, but increasing the resolution of incoming documents or your scanner can go a long way to making this text easier for an OCR engine to read.

EXAMPLE 3: DITHERING

Dithering is intentionally applied noise used to randomize quantization error, intended to prevent patterns such as color banding in images. This works well for human vision as it smooth's out transitions between colors, but has extremely negative effects on OCR quality and therefore accuracy. In the example below the word Cholesterol is clearly visible:



Zooming into the 'tero' section shows why dithering can be so problematic. The letters begin to blend into one another and it's hard for the OCR engine to determine when one letter stops and another beings.

The raw OCR output can be seen below, in this instance it would be very difficult for a rules engine to correctly capture the value.

OCR Output: LDL Ct101eNara

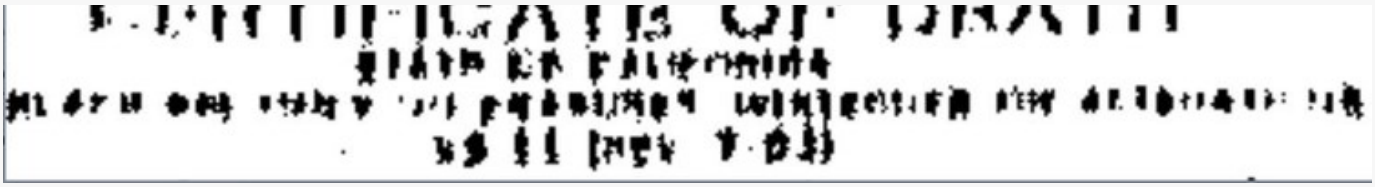
Like the above example, often the best way around this problem is to ensure the resolution settings of your scanner or incoming fax are set as high as they can be.

EXAMPLE 4: THIRD AND FOURTH GENERATION DOCUMENTS

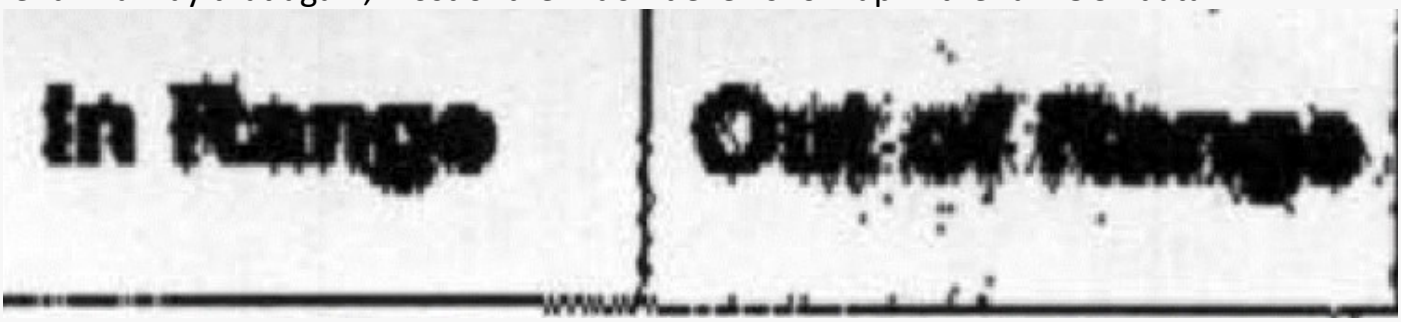
Each time a document is printed and rescanned image quality is lost. The third or fourth time this process is repeated can make a document illegible. While the overall effect can differ depending on the font, type size, and printer/scanner the result is always an incredibly hard to read image.

The below example illustrates the effects of repeated re-scans.

In this example we see that the letters have been eroded to the point that only largest font could be interpreted by a human, and even then it's difficult without context. The lines below it are so degraded that they don't actually show up as OCR'd characters at all.



This example shows the effect of multiple generations on bold typeface. The letters here blend in a way that again, most of them don't even show up in the raw OCR data.



The easiest and most effective way to prevent capture issues with regards to multi-generational images is to revise document workflows. Many times it is possible to optimize a document workflow so that any individual image can be intercepted earlier and scanned before it needs to be printed and scanned.

EXAMPLE 5: FIRST GENERATION IMAGE

Below is simply an example of a first generation image (in this case, a PDF) that OCR's with 100% accuracy. While generally not possible, when a first generation image can be available it should always be used for OCR and data capture workflows.