

# SUBEX HACKATHON SUBMISSION

**Team: The Improbables**

**Members: Andrea Solomon, Dhiganth Rao**

**Problem Statement:** Electronic Reading and extraction of useful information from Invoices.

There are two parts to this problem statement: First, drawing boxes to pinpoint the location of items present in the invoice. Second, to actually extract the items into a textual format for future use.

Our solution:

Initially, OpenCV was used to properly enhance the features of the image, such as tables, and the text itself in the invoice to make it more clear. This was done using methods such as converting to grayscale, thresholding, and vertical and horizontal line detectors. After this, we wrote a function to extract all the contours that contain text in the image.

After extraction of the contours present in the image, we iterated through these contours. During each iteration, we cropped the original image to only the contour, and then proceeded to use PyTesseract, converting each cropped image to a string. This string contains all the textual information present in that cropped image.

We then run a number of preprocessing methods on this string, such as splitting the whole string into substrings of sentences, and removing escape sequences from the sentences.

After this, we implemented a function to differentiate sentences into **items** and **non-items**. This is done by calculating a score for each sentence. The score is calculated by assigning weightage to every component present in the sentence;

for example, a numerical component in the sentence is assigned more importance in a sentence, as it could denote either quantity or price. The weightage for each component is decided by finding the figure of speech each component has, and multiplying the number of occurrences of that figure of speech.

Then, after calculating the score of every sentence, we decided to keep a threshold score; sentences which have a score greater than the threshold, would be classified as items.

However, there are a few drawbacks associated with this method:

- There is no way to determine which is the contour we are looking for, unless we manually look through the output for each image.
- In the case of images that do not contain well defined tabular columns, we don't get any proper contours containing the information we need
- We could not find a proper threshold to classify a word into an item or a non item, as many of the times the output from Pytesseract was unsatisfactory and we could not find any python packages to appropriately tag the words. (We were experimenting with nltk's wordnet.synsets)

So we decided to alter our approach!

Apart from the image preprocessing listed above, we also made use of OpenCv's Inter-Cubic Interpolation function to further enhance the features of the image. Instead of using line detectors to detect contours, we used Kraken, a turn-key OCR system optimized for historical and non-Latin script material to detect and return the coordinates of boxes that contain text. These boxes were rearranged to represent the reading order, from left to right.

We iterated through these boxes, cropped the box from the image, enlarged the images and ran basic preprocessing steps on it using OpenCV before running it through Tesseract, an OCR engine to get the text present in the image.

Our assumption is that a product or an item, when present in an invoice will at least have an item name and the price of the item all in order on a single line.

For example: "Lifebuoy soap 1qty Rs.30.00"

Using this idea we iterated through each sentence in the text derived using Tesseract and checked the last word of the sentence. Using spaCy's Named Entity Recognition, we tagged the word as

1. "QUANTITY": Indicating it is a numerical
2. "MONEY": Indicating it is a numerical containing a unit of currency
3. "PERCENT": Indicating it is a numerical containing the "%" operator, probably referring to either a discount or a tax percentage

If the word belonged to either of these three categories, and the sentence did not contain any obvious key words that could mean it wasn't a product or an item name (such as "invoice", "bill" etc. as these could be followed by numbers too), it was stored into a new list for further processing.


In the second step, spaCy was again used to classify the words in the sentence as to whether it could represent a Discount term, an item, the Price or Quantity and based on that the final result dictionary was updated. Using this final list of items and their prices, the coordinates of the final box required was derived from the bounding boxes calculated before, and the final contour was drawn.

Our approach performs decently well in terms of drawing the contours, however in terms of detecting the data from the invoice provided it doesn't perform very well. Much of this can be attributed to Tesseract itself and the fact that many of the times the headers of the table cannot be distinguished clearly (which is why we decided to use predefined headers and not rely on Tesseract). Also it could be attributed to the poor quality of the images given.

Steps to take before Implementation:

If you decide to run our files, kindly make sure that you have the necessary packages installed. We've listed this at the start of Final.py.

A sample screenshot of the output we obtain is present as well.

<b>GSTIN :-</b> 23CTOPS4492Q1ZX		<b>TAX INVOICE</b>		<b>FSSAI LIC. NO. :</b> 1234567890																																																												
		<b>RAJ SUPER WHOLESALE BAZAR</b> 45, AMBA PRASAD TIWARI MARG, DAULATGANJ UJJAIN-MP-456001 Tel No. : 0734-4060723, 9993736333																																																														
<b>Bill to / Ship to</b> <b>RAJ DATA PROCESSORS</b> 45, DAULATGANJ  <b>UJJAIN</b> State : Madhya Pradesh Code : (23) <b>GSTIN No.</b> <b>PAN No.</b>			Tax Is Payable On Reverse Charge: NO  Place of Supply <b>State :</b> Madhya Pradesh <b>Code :</b> (23)  Bill No.                : <b>2254</b> Bill Date    27/02/2019  Challan No.        :                                Date L.R. No.            :    0                                Date Transport         :																																																													
SNo	DESCRIPTION	QUANTITY	RATE RS.	AMOUNT RS.																																																												
1	<b>SWADIST SOYA OIL 1LTR (POUCH)</b> SPECIAL POUCH PACKING MFG BY : AVI AGRO MARKETED BY : DAMMANI	10	78.10	780.95																																																												
2	<b>BALAJI BASMATI RICE 1 KG</b>	12	45.71	548.57																																																												
3	<b>LIFEBUOY SOAP (LEMON) 125GM*4 94/-</b>	12	76.27	915.25																																																												
4	<b>CAMEL TEA 500GM</b>	6	94.29	565.72																																																												
5	<b>DABUR HONEY 50GM 37/-</b>	2	34.29	68.57																																																												
6	<b>GOLD COIN BREAD 33/-</b>	3	30.00	90.00																																																												
7	<b>KISSAN TOMATO KETCHUP REFILL 450GM 50/-2ND</b>	2	40.68	81.36																																																												
8	<b>BRITANIA MUFFILLS CAKE 10/-</b>	1	8.48	8.48																																																												
9	<b>KHARAK SAKHARIYA (SPECIAL) 1KG</b>	1	120.54	120.54																																																												
<table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="text-align: left;">HSN</td> <td style="text-align: right;">TAXABLE RS.</td> <td style="text-align: right;">IGST RS.</td> <td style="text-align: right;">CGST%</td> <td style="text-align: right;">CGST</td> <td style="text-align: right;">SGST%</td> <td style="text-align: right;">SGST</td> </tr> <tr> <td style="text-align: left;">- %</td> <td style="text-align: right;">90.0</td> <td style="text-align: right;">0.00</td> <td style="text-align: right;">0.00</td> <td style="text-align: right;">0.00</td> <td style="text-align: right;">0.00</td> <td style="text-align: right;">0.00</td> </tr> <tr> <td style="text-align: left;">5.00%</td> <td style="text-align: right;">1963.8</td> <td style="text-align: right;">0.00</td> <td style="text-align: right;">2.50</td> <td style="text-align: right;">49.10</td> <td style="text-align: right;">2.50</td> <td style="text-align: right;">49.10</td> </tr> <tr> <td style="text-align: left;">12.00%</td> <td style="text-align: right;">120.5</td> <td style="text-align: right;">0.00</td> <td style="text-align: right;">6.00</td> <td style="text-align: right;">7.23</td> <td style="text-align: right;">6.00</td> <td style="text-align: right;">7.23</td> </tr> <tr> <td style="text-align: left;">18.00%</td> <td style="text-align: right;">1005.1</td> <td style="text-align: right;">0.00</td> <td style="text-align: right;">9.00</td> <td style="text-align: right;">90.46</td> <td style="text-align: right;">9.00</td> <td style="text-align: right;">90.46</td> </tr> <tr> <td style="text-align: left;">Total</td> <td style="text-align: right;">3179.43</td> <td style="text-align: right;">0.00</td> <td></td> <td style="text-align: right;">146.78</td> <td></td> <td style="text-align: right;">146.78</td> </tr> </table>			HSN	TAXABLE RS.	IGST RS.	CGST%	CGST	SGST%	SGST	- %	90.0	0.00	0.00	0.00	0.00	0.00	5.00%	1963.8	0.00	2.50	49.10	2.50	49.10	12.00%	120.5	0.00	6.00	7.23	6.00	7.23	18.00%	1005.1	0.00	9.00	90.46	9.00	90.46	Total	3179.43	0.00		146.78		146.78	<table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="text-align: right;">GROSS AMOUNT RS.</td> <td style="text-align: right;"><b>3179.43</b></td> </tr> <tr> <td style="text-align: right;">LESS DISCOUNT RS.</td> <td style="text-align: right;">0.00</td> </tr> <tr> <td style="text-align: right;">FRIEGHT +/- RS.</td> <td style="text-align: right;">0.00</td> </tr> <tr> <td style="text-align: right;">TAXABLE RS.</td> <td style="text-align: right;"><b>3179.43</b></td> </tr> <tr> <td style="text-align: right;">I.G.S.T. Rs.</td> <td style="text-align: right;">0.00</td> </tr> <tr> <td style="text-align: right;">C.G.S.T. Rs.</td> <td style="text-align: right;">146.78</td> </tr> <tr> <td style="text-align: right;">S.G.S.T. Rs.</td> <td style="text-align: right;">146.78</td> </tr> <tr> <td style="text-align: right;">CESS Rs.</td> <td style="text-align: right;">0.00</td> </tr> <tr> <td style="text-align: right;"><b>Net Amount Rs.</b></td> <td style="text-align: right;"><b>3473.00</b></td> </tr> </table>		GROSS AMOUNT RS.	<b>3179.43</b>	LESS DISCOUNT RS.	0.00	FRIEGHT +/- RS.	0.00	TAXABLE RS.	<b>3179.43</b>	I.G.S.T. Rs.	0.00	C.G.S.T. Rs.	146.78	S.G.S.T. Rs.	146.78	CESS Rs.	0.00	<b>Net Amount Rs.</b>	<b>3473.00</b>
HSN	TAXABLE RS.	IGST RS.	CGST%	CGST	SGST%	SGST																																																										
- %	90.0	0.00	0.00	0.00	0.00	0.00																																																										
5.00%	1963.8	0.00	2.50	49.10	2.50	49.10																																																										
12.00%	120.5	0.00	6.00	7.23	6.00	7.23																																																										
18.00%	1005.1	0.00	9.00	90.46	9.00	90.46																																																										
Total	3179.43	0.00		146.78		146.78																																																										
GROSS AMOUNT RS.	<b>3179.43</b>																																																															
LESS DISCOUNT RS.	0.00																																																															
FRIEGHT +/- RS.	0.00																																																															
TAXABLE RS.	<b>3179.43</b>																																																															
I.G.S.T. Rs.	0.00																																																															
C.G.S.T. Rs.	146.78																																																															
S.G.S.T. Rs.	146.78																																																															
CESS Rs.	0.00																																																															
<b>Net Amount Rs.</b>	<b>3473.00</b>																																																															
BANK NAME: BANK OF INDIA A/C NO. : 9100123456456 IFSC : BKID00001901			Rupees Three Thousand Four Hundred Seventy-Three Only																																																													

Terms & Conditions : 1. We warranted hereby to certify be of that the foods nature mentioned and quality in which this invoice these are purpose to be articles of food being of perishable nature must be stored in cool & dry place. 2. Subject to Ujjain Jurisdiction 3. E. & O. E.

For : RAJ SUPER WHOLESALE BAZAR

(Authorised Signatory)

Certified that the Particulars given above are true and correct

