TABLE DETECTION AND DATA EXTRACTION

Background: For a long time, humans have relied on paper invoices to process payments and maintaining accounts. Extraction of information from invoice tables manually can be tedious and time consuming. The solution to this industry wide problem is to automate the process of table detection and data extraction from such invoices to keep all records on computer systems.

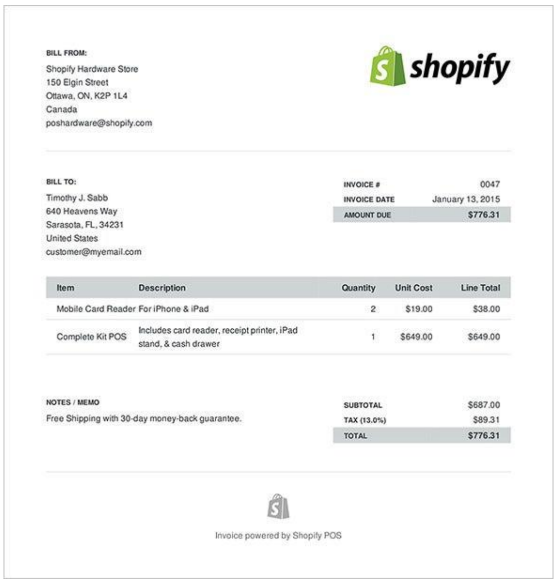
Relevance of the Problem: People spend huge amount of time in manually browsing through various invoices and collecting important information for ledger. In order to solve this problem and reduce wastage of paper, human time and labour, automation of this process is necessary.

Problem Statement: To Detect Tables from the invoices provided and extract data from those into data frames. This requires both OCR to most efficiently 'recognise' and 'read' the data off the invoice and extract information.

- Optical Character Recognition - Recognizing the text and numbers present in the documents.
- Information Extraction - Once the Process of OCR is complete it's important to identify which piece of text corresponds to which extracted field. If a field is the total, subtotal, date of invoice, vendor etc.

For Example:

Step 1: Table Detection



BILL FROM:
Shopify Hardware Store
150 Elgin Street
Ottawa, ON, K2P 1L4
Canada
poshardware@shopify.com

BILL TO:
Timothy J. Sabb
640 Heavens Way
Sarasota, FL, 34231
United States
customer@myemail.com

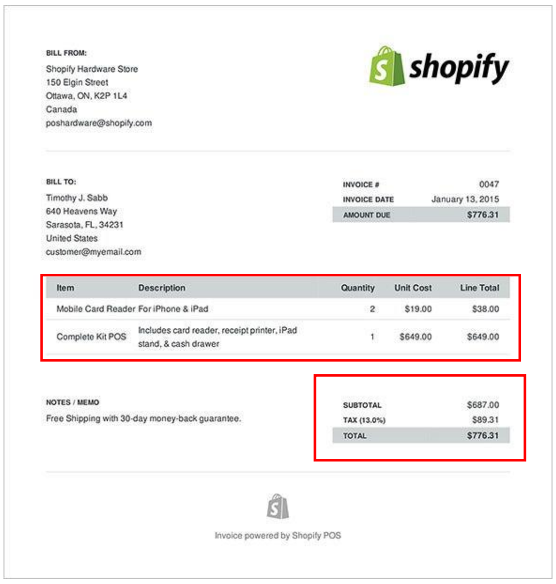
INVOICE #	0047
INVOICE DATE	January 13, 2015
AMOUNT DUE	$776.31

Item	Description	Quantity	Unit Cost	Line Total
Mobile Card Reader	For iPhone & iPad	2	$19.00	$38.00
Complete Kit POS	Includes card reader, receipt printer, iPad stand, & cash drawer	1	$649.00	$649.00

NOTES / MEMO
Free Shipping with 30-day money-back guarantee.

SUBTOTAL	$687.00
TAX (13.0%)	$89.31
TOTAL	$776.31

Invoice powered by Shopify POS



BILL FROM:
Shopify Hardware Store
150 Elgin Street
Ottawa, ON, K2P 1L4
Canada
poshardware@shopify.com

BILL TO:
Timothy J. Sabb
640 Heavens Way
Sarasota, FL, 34231
United States
customer@myemail.com

INVOICE # 0047
INVOICE DATE January 13, 2015
AMOUNT DUE $776.31

Item | Description | Quantity | Unit Cost | Line Total
Mobile Card Reader | For iPhone & iPad | 2 | $19.00 | $38.00
Complete Kit POS | Includes card reader, receipt printer, iPad stand, & cash drawer | 1 | $649.00 | $649.00

NOTES / MEMO
Free Shipping with 30-day money-back guarantee.

SUBTOTAL $687.00
TAX (13.0%) $89.31
TOTAL $776.31

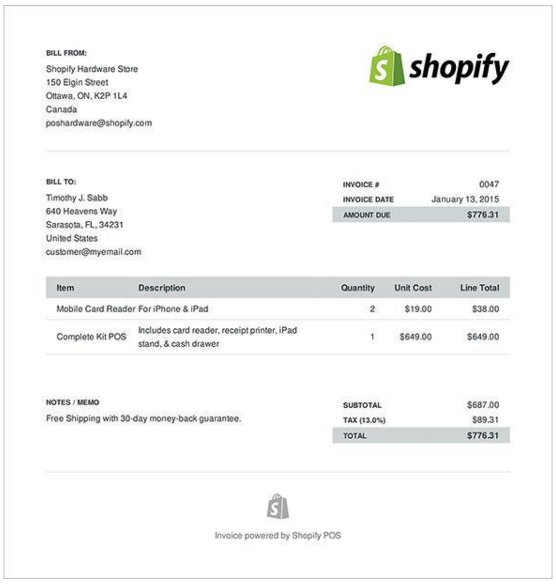
Invoice powered by Shopify POS

Output Format: Bounding box coordinates in the form [x1, y1, x2, y2], (x1,y1) is the top-left corner coordinate and (x2, y2) is the bottom right corner coordinates. For more than one bounding boxes, output would be list of bounding boxes, [[x1, y1, x2, y2]...]

For Example: For the above image, the original boxes = [[45, 290, 537, 468], [350, 184, 539, 230]]

Evaluation Metrics Required: Intersection over Union (IoU) is a measure of the magnitude of overlap between 2 bounding boxes or 2 objects. The problem would be evaluated on the mean average precision at different IoU thresholds. The IoU of a set of predicted bounding boxes and ground truth bounding boxes is calculated as:
IoU (A,B) = $(A \cap B)/(A \cup B)$

Step 2: Information Extraction



BILL FROM:
Shopify Hardware Store
150 Elgin Street
Ottawa, ON, K2P 1L4
Canada
poshardware@shopify.com

BILL TO:
Timothy J. Sabb
640 Heavens Way
Sarasota, FL, 34231
United States
customer@myemail.com

INVOICE #	0047
INVOICE DATE	January 13, 2015
AMOUNT DUE	$776.31

Item	Description	Quantity	Unit Cost	Line Total
Mobile Card Reader	For iPhone & iPad	2	$19.00	$38.00
Complete Kit POS	Includes card reader, receipt printer, iPad stand, & cash drawer	1	$649.00	$649.00

NOTES / MEMO
Free Shipping with 30-day money-back guarantee.

SUBTOTAL	$687.00
TAX (13.0%)	$89.31
TOTAL	$776.31

Invoice powered by Shopify POS

EXTRACTION OUTPUT

Row_1: {“item”: “Mobile Card Reader”, “Description”: “For iPhone & iPad”, “Quantity”: 2, “Unit Cost”: $19.00, “Line Total”: $38.00}

Row_2: {“item”: “Complete Kit POS”, “Description”: “Includes card reader, receipt printer, iPad stand, & case drawer”, “Quantity”: 1, “Unit Cost”: $649.00, “Line Total”: $649.00},

“SUBTOTAL”: $687.00,

“TAX (13.0%)”: $89.31,

“TOTAL”: $776.31

Evaluation Metrics Required: Word Error Rate (WER) is the number of errors divided by the total words. To put it in a simple formula, **Word Error Rate = (Substitutions + Insertions + Deletions) / Number of Words Spoken**
Where,

- A **substitution** occurs when a word gets replaced (for example, “noose” is transcribed as “moose”)
- An **insertion** is when a word is added that wasn’t said (for example, “SAT” becomes “essay tea”)
- A **deletion** happens when a word is left out of the transcript completely (for example, “turn it around” becomes “turn around”)

Reproducible Code:
Participants should upload reproducible code, instructions file on how to run the code and create the dev environment; for Python submission include a Conda environment.yml file and requirements.txt file for dev environment creation.