

ABInBev



Maverick Hackathon

...

Shubham Bhardwaj
Saptarshi Ratna

Team Name - ALPHA 1

Automating Invoice Digitization Problem

“ Develop an intelligent system that can translate unstructured data (manual + system based text, images) into a machine-readable format.

There are intelligent character reading systems in the market but their efficiencies are quite low. We need you to create an intelligent bot that can be trained to identify key data points like invoice data, vendor name, unit of currency, tax amount, PO or contract reference number from the document. ”

Approach


- Using Grayscale, Thresholding and Word Bounds
- Using pytesseract for OCR
- Using Google Vision API
- Named Entity Recognition
- Using Machine Learning to auto tag labels

Conventional Image Processing

LETTERPRESS DESIGNS

INVOICE

Number 76 - 981102



ITEM DESCRIPTION	QTY	PRICE	TOTAL
Graphic Design	25	\$ 20.00	\$ 500.00
Web Design (per page)	5	\$ 25.00	\$ 125.00
Web Development	5	\$ 50.00	\$ 250.00
Brand Bible	1	\$ 30.00	\$ 30.00
Logo Studies	3	\$ 30.00	\$ 90.00
SUB TOTAL			\$ 995.00
TAX (15%)			\$ 149.25
GRAND TOTAL			\$ 1144.25

PAYABLE TO

Stephanie Chan
25 Wood St, West Boylston,
MA, 01583

BANK DETAILS

Stephanie Chan
United Fremont Bank
Acct. No. 229 - 2091 - 1092 - 01


THANK YOU

Letterpress Designs - (04) 298 1092 4095 - +76 209 2872 0192 - letterpress@info.com

LETTERPRESS DESIGNS

INVOICE

Number 76 - 981102



ITEM DESCRIPTION	QTY	PRICE	TOTAL
Graphic Design	25	\$ 20.00	\$ 500.00
Web Design (per page)	5	\$ 25.00	\$ 125.00
Web Development	5	\$ 50.00	\$ 250.00
Brand Bible	1	\$ 30.00	\$ 30.00
Logo Studies	3	\$ 30.00	\$ 90.00
SUB TOTAL			\$ 995.00
TAX (15%)			\$ 149.25
GRAND TOTAL			\$ 1144.25

PAYABLE TO

Stephanie Chan
25 Wood St, West Boylston,
MA, 01583

BANK DETAILS

Stephanie Chan
United Fremont Bank
Acct. No. 229 - 2091 - 1092 - 01


THANK YOU

Letterpress Designs - (04) 298 1092 4095 - +76 209 2872 0192 - letterpress@info.com

LETTERPRESS DESIGNS

INVOICE

Number 76 - 981102



ITEM DESCRIPTION	QTY	PRICE	TOTAL
GRAPHIC DESIGN	25	\$ 20.00	\$ 500.00
WEB DESIGN (PER PAGE)	5	\$ 25.00	\$ 125.00
WEB DEVELOPMENT	5	\$ 50.00	\$ 250.00
BRAND BIBLE	1	\$ 30.00	\$ 30.00
LOGO STUDIES	3	\$ 30.00	\$ 90.00
SUB TOTAL			\$ 995.00
TAX (15%)			\$ 149.25
GRAND TOTAL			\$ 1144.25

PAYABLE TO

Stephanie Chan
25 Wood St, West Boylston,
MA, 01583

BANK DETAILS

Stephanie Chan
United Fremont Bank
Acct. No. 229 - 2091 - 1092 - 01

THANK YOU

Letterpress Designs - (04) 298 1092 4095 - +76 209 2872 0192 - letterpress@info.com

pytesseract for OCR

[illegible]

LETTERPRESS DESIGNS

INVOICE

Number 76 - 981102

(TEM DESCRIPTION

Graphic Design

Web Design (per page)

Web Development

Brand Bible

Logo Studies

BANK DETAILS

Stephanie Chan

United Fremont Bank

Acct. No. 229-2091 1092 -01

Qiy PRICE TOTAL

25 \$20.00 \$500.00

5 \$25.00 \$125.00

5 \$50.00 \$ 250.00

1 \$30.00 \$30.00

3 \$30.00 \$90.00

SUB TOTAL \$995.00

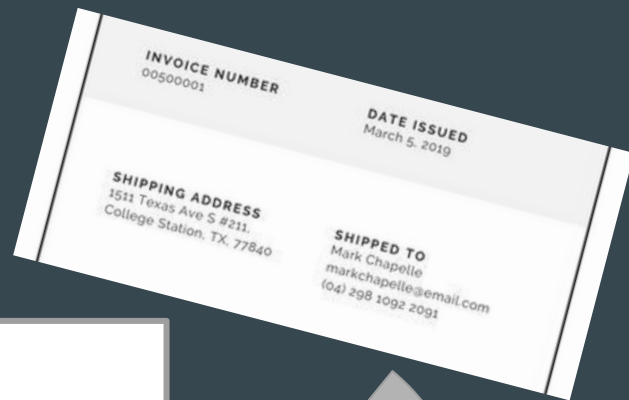
TAX (15%) \$149.25,

GRAND TOTAL \$ 144,25



Problems faced with Google Vision API

- An invoice is not constructed of coherent text lines, but mostly contains tables and small text blocks.
- The location and structure of these blocks will vary over all different invoice templates.



INVOICE NUMBER
DATE ISSUED
00500001
March 5, 2019

SHIPPING ADDRESS
SHIPPED TO
Mark Chapelle
1511 Texas Ave S #211.
College Station, TX. 77840
markchappelleaemail.com
(04) 298 1092 2091



How we solved the problems

We first divided the receipt using spatial separation into various blocks and extracted these blocks rather than reading line by line.

Then data was extracted from each block categorically for better interpretation by the Machine Learning model.

MUZZY HILL
BESPOKE CLOTHING
INFO@MUZZYHILL.COM / WWW.MUZZYHILL.COM

INVOICE NUMBER
00500001

DATE ISSUED
March 5, 2019

SHIPPING ADDRESS
1511 Texas Ave S #211,
College Station, TX. 77840

SHIPPED TO
Mark Chapelle
markchapelle@email.com
(04) 298 1092 2091

SERVICE DESCRIPTION	QTY	PRICE	TOTAL
Custom Tuxedd	01	\$ 500	\$ 500
Custom Belt	01	\$ 300	\$ 300
Custom Cuff Links	01	\$ 400	\$ 400
			TOTAL \$ 1200

Using Google AutoML API for labelling key-value pairs

Google Cloud Platform

My First Project

Natural Language

Invoice_recog

VIEW LABEL STATS

EXPORT DATA

Dashboard

Datasets

Models

IMPORT

ITEMS

TRAIN

EVALUATE

TEST & USE

Entity Extraction

All items

7

Filter table

Items

Labels

LETTERPRESS DESIGNS INVOICE Number 76 - 981102 ITEM DESCRIPTION QTY PRICE TOTAL

Graphic Design 25 \$ 20.00 \$ 500.00 Web Design (per

bank_details, company_contact_email, company_contact_no, company_name, currency, invoice_no, item_description, price, quantity, tax, total_price

MUZZY HILL BESPOKE CLOTHING INFOSMUZZYHILL.COM / www.MUZZYHILLS.COM INVOICE

NUMBER DATE ISSUED 00500001 March 5, 2019 SHIPPING ADDRESS SHIPPED TO Mark

Chapelle 1511 Texas

billed_to, company_contact_email, company_name, currency, invoice_no, item_description, price, quantity, total_price

Pay Invoice PayPal VISA zsavelater SAa . - LEATHERMAN'S - INVOICE Leatherman's Supply 324

South Street San Jose, CA 95131 United States Invoice number RAL1002125 Invoice date

2/15/2014 Bill To Payment terms Net

billed_to, company_name, currency, invoice_no, item_description, price, quantity, tax, total_price

ZYLKAR CORPORATION PRO FORMA INVOICE 1561 Appleview Town Invoice: #00121 Bakers

Street July 21, 2015 Chicago, IL 60411 Bill to: Ship To:

billed_to, company_contact_no, company_name, currency, invoice_no, item_description, price, tax, total_price

CREDIT NOTE # CN-17 Credits Remaining \$562.750 Zyker 7455 Drew Court White City Kansas

66872 U.S.A 270-510-0585 Bill To Rob & Joe Traders 4141 Hacienda Drive Pleasanton Credit

billed_to, company_contact_no, company_name, currency, invoice_no, item_description, price, quantity, tax, total_price

INVOICE INVOICE NUMBER DATE OF ISSUE a0001 mm/ddyyy Your company name BILLED TO

123 Your Street City, State, Country, ZIP Client Name Street address Code City, State

billed_to, company_contact_email, company_name, currency, invoice_no, item_description, price, total_price

COMMERCIAL INVOICE Bill To INV-000380 02 Jul 2018 Rodney Sanders 1007 Mountain Drive

sebring NY10029 NY Ship To Winchester avenue Harrietsfield QC JOT-Y9P Canada Zyker

billed_to, company_contact_no, company_name, currency, invoice_no, item_description, price, quantity, tax, total_price

bank_details

1

billed_to

6

company_co...

3

ADD NEW LABEL

*Thank
you*

