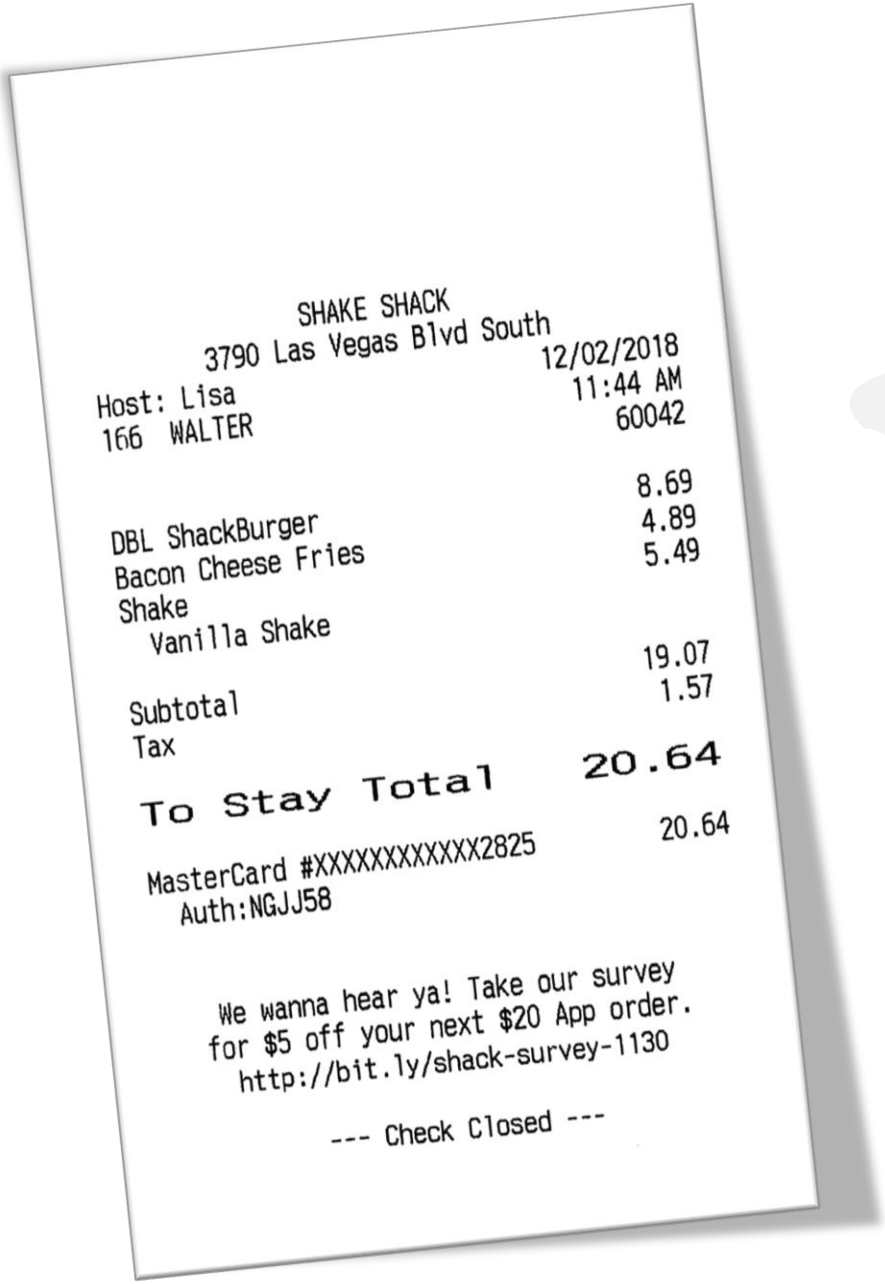


SHAKE SHACK
3790 Las Vegas Blvd South
Host: Lisa 12/02/2018
11:44 AM
166 WALTER 60042

8.69
DBL ShackBurger 4.89
Bacon Cheese Fries 5.49
Shake
Vanilla Shake

19.07
Subtotal 1.57
Tax

To Stay Total 20.64

20.64
MasterCard #XXXXXXXXXXXX2825
Auth:NGJJ58

We wanna hear ya! Take our survey
for $5 off your next $20 App order.
http://bit.ly/shack-survey-1130

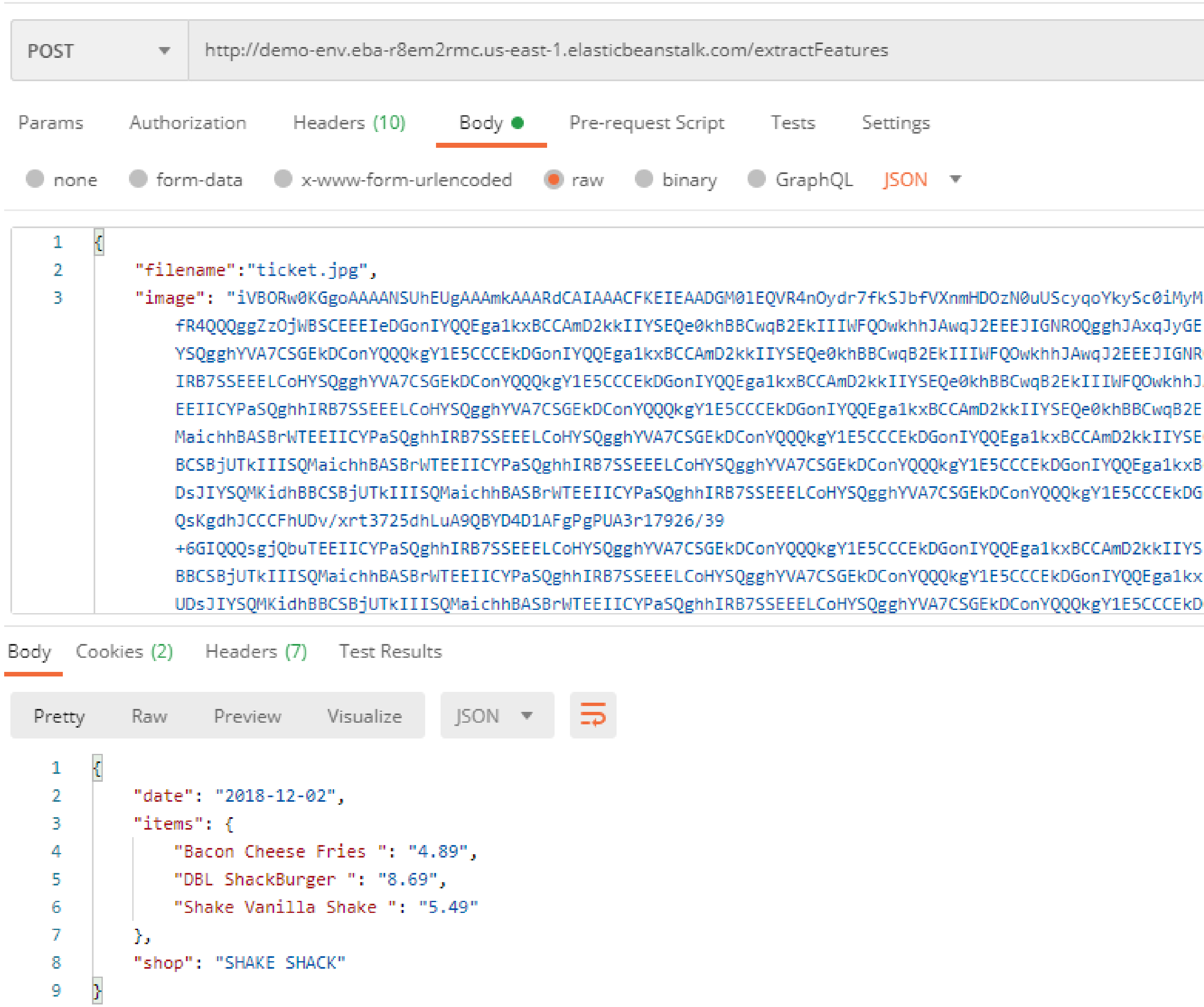
--- Check Closed ---

EXTRACCIÓN DE DATOS EN *RECIBOS*

- ***Pilar Madariaga Lasala***
- Estudiante Máster Inteligencia Artificial en Unir
- Pasantia en Grupo Bancolombia

OBJETIVO
Extracción de características

- *Servicio Rest en Python*
- **Input**: Imagen en base64
- **Output**:
  - Fecha compra
  - Establecimiento
  - Listado de producto y precio

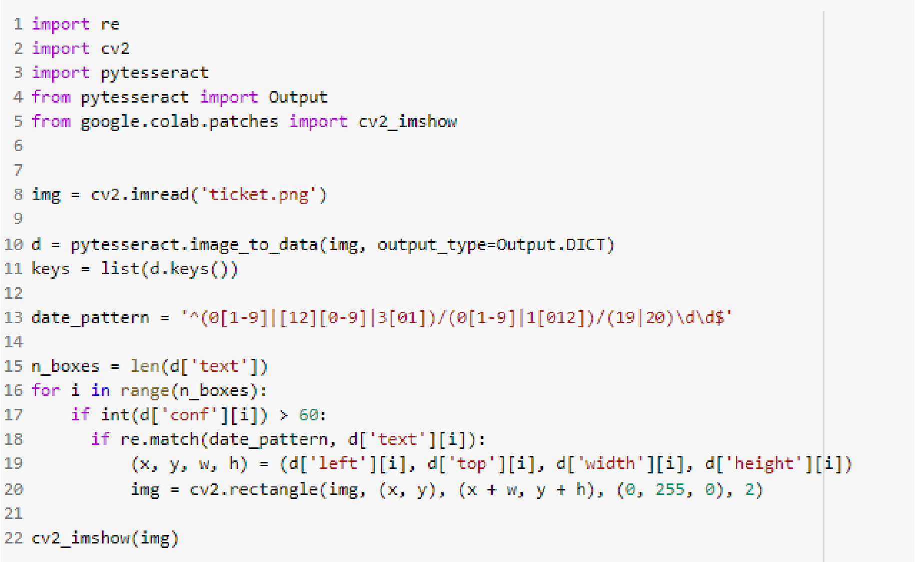


http://demo-env.eba-r8em2rmc.us-east-1.elasticbeanstalk.com/extractFeatures

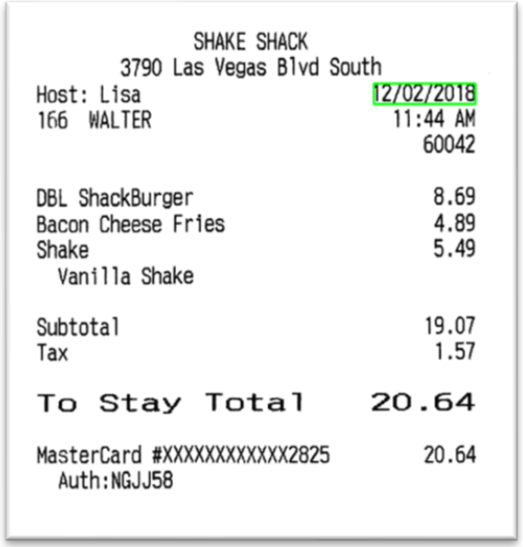
```
{
    "filename":"ticket.jpg",
    "image": "iVBORw0KGgoAAAANSUhEUgAAAmkAAARdCAIAAACFKEIEAADGM01EQVR4nOydr7fkSJbfVXnmHDOzN0uUScyqoYkySc0iMyM
        fR4QQQggZzOjWBSCEEEIeDGonIYQQEga1kxBCCAmD2kkIIYSEQe0khBBCwqB2EkIIIWFQOwkhhJAwqJ2EEEJIGNROQgghJAxqJyGE
        YSQgghYVA7CSGEkDConYQQQkgY1E5CCCEkDGonIYQQEga1kxBCCAmD2kkIIYSEQe0khBBCwqB2EkIIIWFQOwkhhJAwqJ2EEEJIGNR
        IRB7SSEEELCoHYSQgghYVA7CSGEkDConYQQQkgY1E5CCCEkDGonIYQQEga1kxBCCAmD2kkIIYSEQe0khBBCwqB2EkIIIWFQOwkhhJ
        EEIICYPaSQghhIRB7SSEEELCoHYSQgghYVA7CSGEkDConYQQQkgY1E5CCCEkDGonIYQQEga1kxBCCAmD2kkIIYSEQe0khBBCwqB2E
        MaichhBASBrWTEEIICYPaSQghhIRB7SSEEELCoHYSQgghYVA7CSGEkDConYQQQkgY1E5CCCEkDGonIYQQEga1kxBCCAmD2kkIIYSE
        BCSBjUTkIIISQMaichhBASBrWTEEIICYPaSQghhIRB7SSEEELCoHYSQgghYVA7CSGEkDConYQQQkgY1E5CCCEkDGonIYQQEga1kxB
        DsJIYSQMKidhBBCSBjUTkIIISQMaichhBASBrWTEEIICYPaSQghhIRB7SSEEELCoHYSQgghYVA7CSGEkDConYQQQkgY1E5CCCEkDG
        QsKgdhJCCCFhUDv/xrt3725dhLuA9QBYD4D1AFgPgPUA3r17926/39
        +6GIQQQsgjQbuTEEIICYPaSQghhIRB7SSEEELCoHYSQgghYVA7CSGEkDConYQQQkgY1E5CCCEkDGonIYQQEga1kxBCCAmD2kkIIYS
        BBCSBjUTkIIISQMaichhBASBrWTEEIICYPaSQghhIRB7SSEEELCoHYSQgghYVA7CSGEkDConYQQQkgY1E5CCCEkDGonIYQQEga1kx
        UDsJIYSQMKidhBBCSBjUTkIIISQMaichhBASBrWTEEIICYPaSQghhIRB7SSEEELCoHYSQgghYVA7CSGEkDConYQQQkgY1E5CCCEkD
```

```
{
    "date": "2018-12-02",
    "items": {
        "Bacon Cheese Fries ": "4.89",
        "DBL ShackBurger ": "8.69",
        "Shake Vanilla Shake ": "5.49"
    },
    "shop": "SHAKE SHACK"
}
```

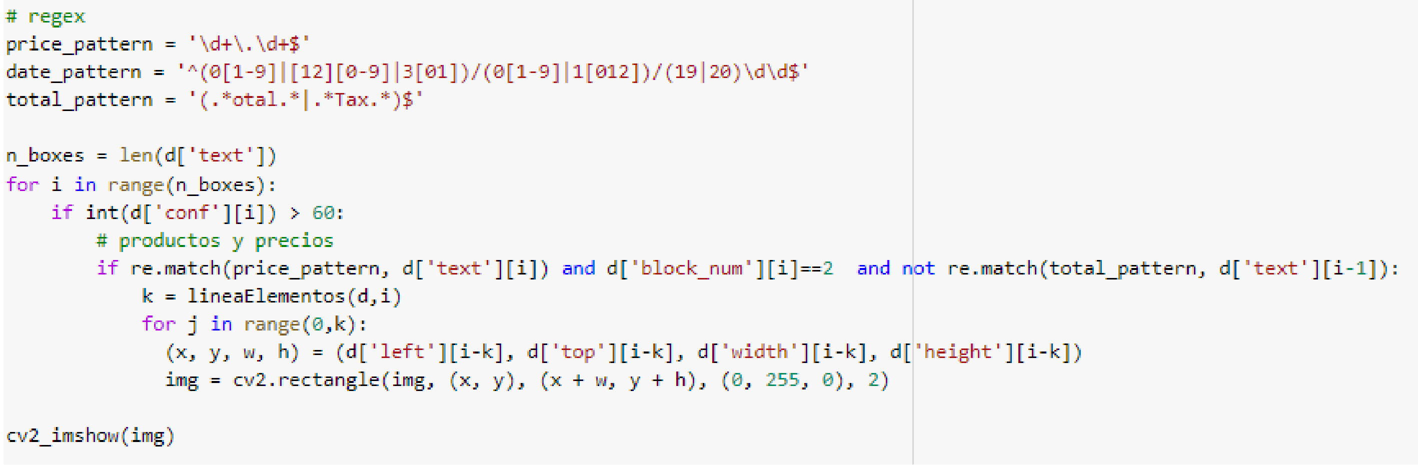
Primeros pasos: *pytesseract*



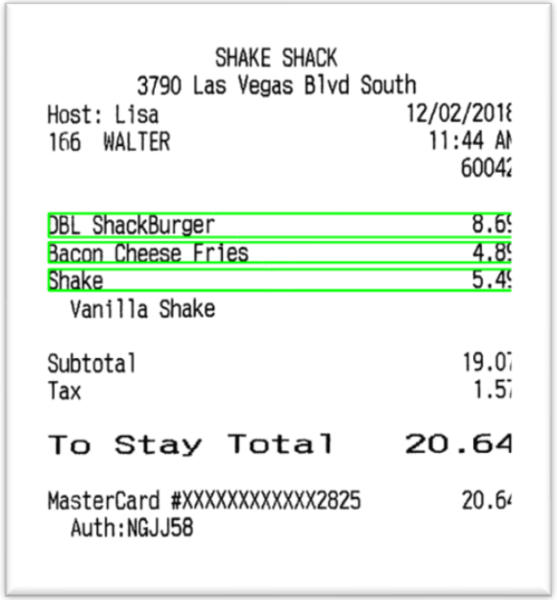
```
import re
import cv2
import pytesseract
from pytesseract import Output
from google.colab.patches import cv2_imshow


img = cv2.imread('ticket.png')

d = pytesseract.image_to_data(img, output_type=Output.DICT)
keys = list(d.keys())

date_pattern = '^(0[1-9]|[12][0-9]|3[01])/(0[1-9]|1[012])/(19|20)\d\d$'

n_boxes = len(d['text'])
for i in range(n_boxes):
    if int(d['conf'][i]) > 60:
      if re.match(date_pattern, d['text'][i]):
          (x, y, w, h) = (d['left'][i], d['top'][i], d['width'][i], d['height'][i])
          img = cv2.rectangle(img, (x, y), (x + w, y + h), (0, 255, 0), 2)

cv2_imshow(img)
```



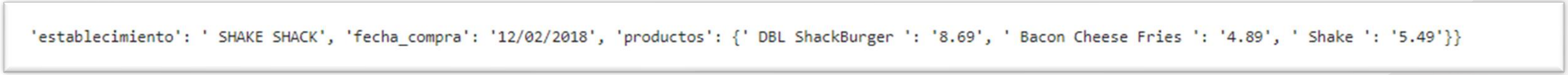
```
               SHAKE SHACK
          3790 Las Vegas Blvd South
Host: Lisa                    12/02/2018
166  WALTER                     11:44 AM
                                   60042

DBL ShackBurger                     8.69
Bacon Cheese Fries                  4.89
Shake                               5.49
  Vanilla Shake

Subtotal                           19.07
Tax                                 1.57

To Stay Total                      20.64

MasterCard #XXXXXXXXXXXX2825       20.64
  Auth:NGJJ58
```



```
# regex
price_pattern = '\d+\.\d+$'
date_pattern = '^(0[1-9]|[12][0-9]|3[01])/(0[1-9]|1[012])/(19|20)\d\d$'
total_pattern = '(.*otal.*|.*Tax.*)$'

n_boxes = len(d['text'])
for i in range(n_boxes):
    if int(d['conf'][i]) > 60:
        # productos y precios
        if re.match(price_pattern, d['text'][i]) and d['block_num'][i]==2  and not re.match(total_pattern, d['text'][i-1]):
            k = lineaElementos(d,i)
            for j in range(0,k):
              (x, y, w, h) = (d['left'][i-k], d['top'][i-k], d['width'][i-k], d['height'][i-k])
              img = cv2.rectangle(img, (x, y), (x + w, y + h), (0, 255, 0), 2)

cv2_imshow(img)
```



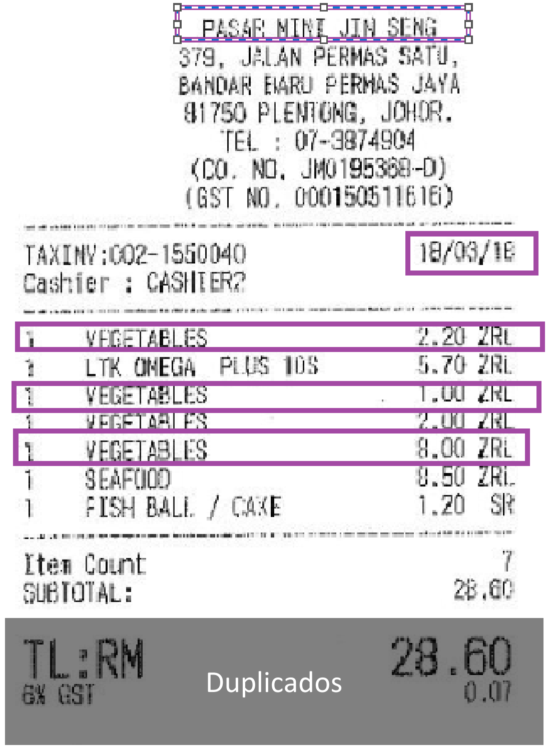
```
               SHAKE SHACK
          3790 Las Vegas Blvd South
Host: Lisa                     12/02/201
166  WALTER                      11:44 A
                                    6004

DBL ShackBurger                      8.6
Bacon Cheese Fries                   4.8
Shake                                5.4
  Vanilla Shake

Subtotal                            19.0
Tax                                  1.5

To Stay Total                      20.64

MasterCard #XXXXXXXXXXXX2825        20.6
  Auth:NGJJ58
```



```
'establecimiento': ' SHAKE SHACK', 'fecha_compra': '12/02/2018', 'productos': {' DBL ShackBurger ': '8.69', ' Bacon Cheese Fries ': '4.89', ' Shake ': '5.49'}}
```

DATASET

Recibos Compras (I)



PASAR MINI JIN SENG
379, JALAN PERMAS SATU,
BANDAR BARU PERMAS JAYA
81750 PLENTONG, JOHOR.
TEL : 07-3874904
(CO. NO. JM0195368-D)
(GST NO. 000150511616)

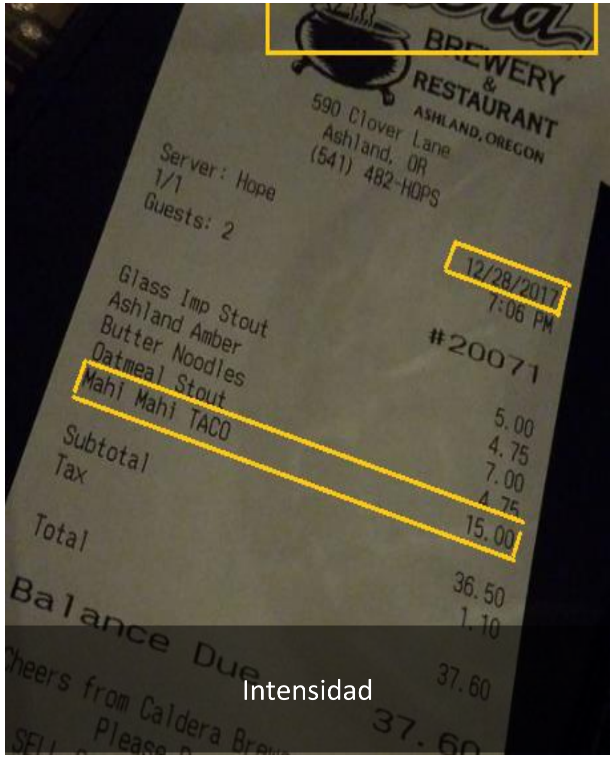
TAXINV:002-1550040 18/03/18
Cashier : CASHIER2

1	VEGETABLES	2.20 ZRL
1	LTK OMEGA PLUS 10S	5.70 ZRL
1	VEGETABLES	1.00 ZRL
1	VEGETABLES	2.00 ZRL
1	VEGETABLES	8.00 ZRL
1	SEAFOOD	8.50 ZRL
1	FISH BALL / CAKE	1.20 SR

Item Count 7
SUBTOTAL: 28.60

TL:RM 28.60
6% GST 0.07

Duplicados



BREWERY
&
RESTAURANT
ASHLAND, OREGON
590 Clover Lane
Ashland, OR
(541) 482-HOPS

Server: Hope
1/1
Guests: 2

12/28/2017
7:06 PM
#20071

Glass Imp Stout	5.00
Ashland Amber	4.75
Butter Noodles	7.00
Oatmeal Stout	4.75
Mahi Mahi TACO	15.00

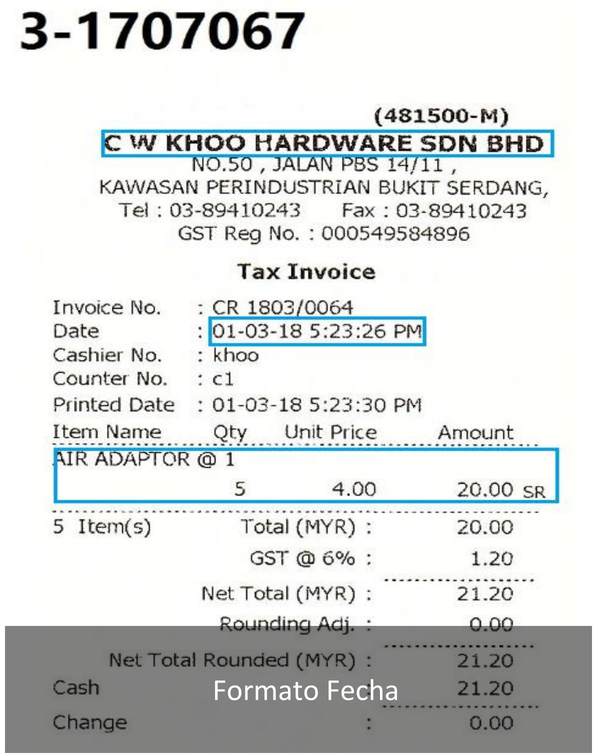
Subtotal 36.50
Tax 1.10

Total 37.60

Balance Due 37.60

Cheers from Caldera Brew
Please

Intensidad



3-1707067

(481500-M)
C W KHOO HARDWARE SDN BHD
NO.50 , JALAN PBS 14/11 ,
KAWASAN PERINDUSTRIAN BUKIT SERDANG,
Tel : 03-89410243 Fax : 03-89410243
GST Reg No. : 000549584896

Tax Invoice

Invoice No. : CR 1803/0064
Date : 01-03-18 5:23:26 PM
Cashier No. : khoo
Counter No. : c1
Printed Date : 01-03-18 5:23:30 PM

Item Name	Qty	Unit Price	Amount
AIR ADAPTOR @ 1			
	5	4.00	20.00 SR

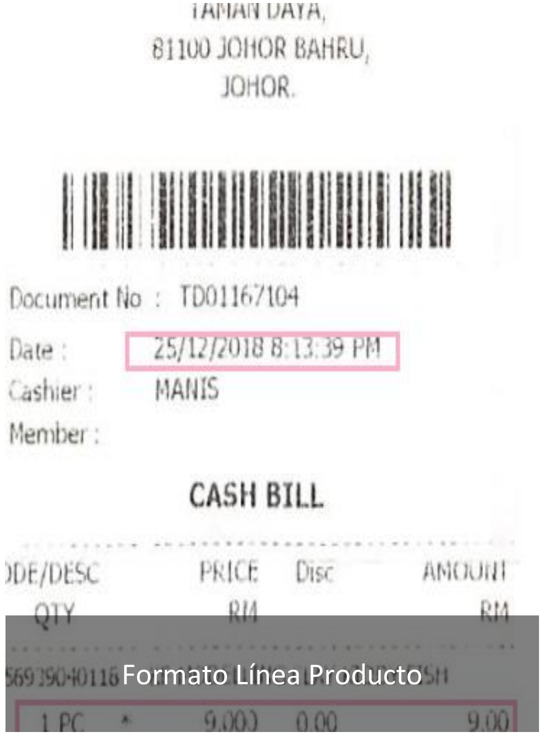
5 Item(s) Total (MYR) : 20.00
GST @ 6% : 1.20
Net Total (MYR) : 21.20
Rounding Adj. : 0.00
Net Total Rounded (MYR) : 21.20
Cash 21.20
Change : 0.00

Formato Fecha

SROIE 2019 *https://drive.google.com/drive/folders/1ShItNWXyiY1tFDM5W02bceHuJjyeeJl2*

DATASET

Recibos Compras (II)



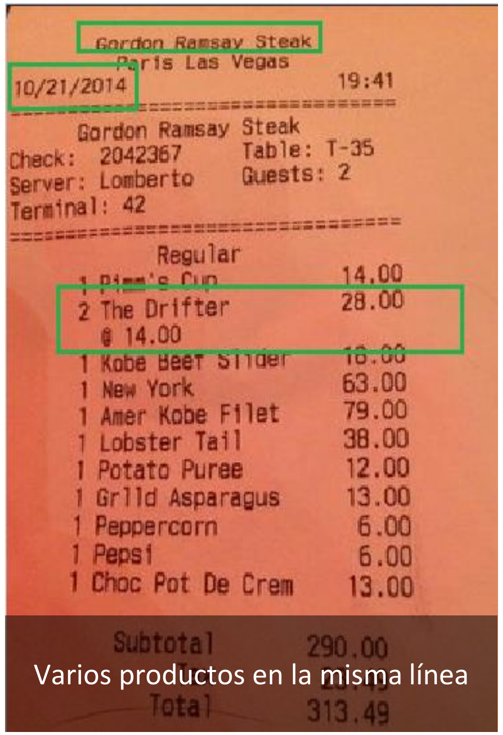
TAMAN DAYA,
81100 JOHOR BAHRU,
JOHOR.

Document No : TD01167104
Date : 25/12/2018 8:13:39 PM
Cashier : MANIS
Member :

CASH BILL

DE/DESC	PRICE	Disc	AMOUNT
QTY	RM		RM
569390401116			SH
1 PC *	9.00	0.00	9.00

Formato Línea Producto



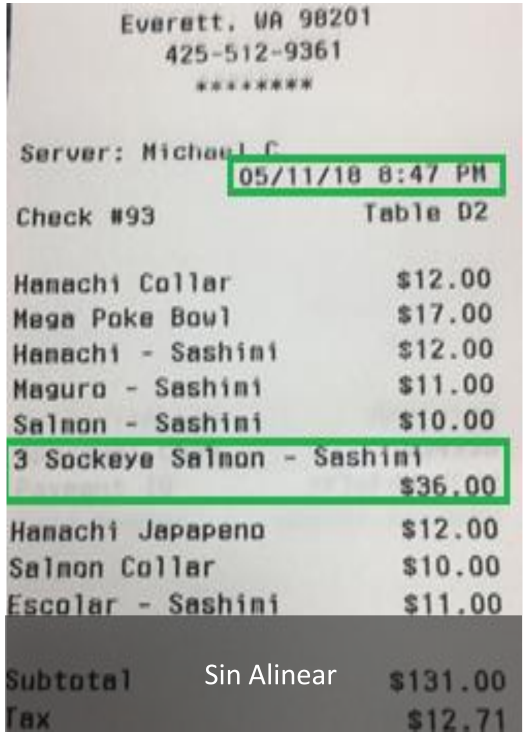
Gordon Ramsay Steak
Paris Las Vegas
10/21/2014 19:41

Gordon Ramsay Steak
Check: 2042367 Table: T-35
Server: Lomberto Guests: 2
Terminal: 42

Regular

1 Pimm's Cup	14.00
2 The Drifter	28.00
@ 14.00	
1 Kobe Beef Slider	18.00
1 New York	63.00
1 Amer Kobe Filet	79.00
1 Lobster Tail	38.00
1 Potato Puree	12.00
1 Grlld Asparagus	13.00
1 Peppercorn	6.00
1 Pepsi	6.00
1 Choc Pot De Crem	13.00
Subtotal	290.00
Total	313.49

Varios productos en la misma línea

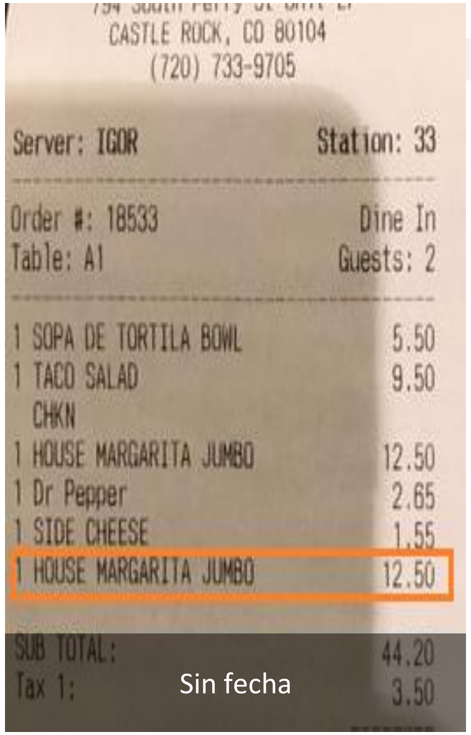


Everett, WA 98201
425-512-9361
****

Server: Michael C
05/11/18 8:47 PM
Check #93 Table D2

Hamachi Collar	$12.00
Mega Poke Bowl	$17.00
Hamachi - Sashimi	$12.00
Maguro - Sashimi	$11.00
Salmon - Sashimi	$10.00
3 Sockeye Salmon - Sashimi	
	$36.00
Hamachi Japapeno	$12.00
Salmon Collar	$10.00
Escolar - Sashimi	$11.00
Subtotal	$131.00
Tax	$12.71

Sin Alinear



CASTLE ROCK, CO 80104
(720) 733-9705

Server: IGOR Station: 33
Order #: 18533 Dine In
Table: A1 Guests: 2

1 SOPA DE TORTILA BOWL	5.50
1 TACO SALAD	9.50
CHKN	
1 HOUSE MARGARITA JUMBO	12.50
1 Dr Pepper	2.65
1 SIDE CHEESE	1.55
1 HOUSE MARGARITA JUMBO	12.50
SUB TOTAL:	44.20
Tax 1:	3.50

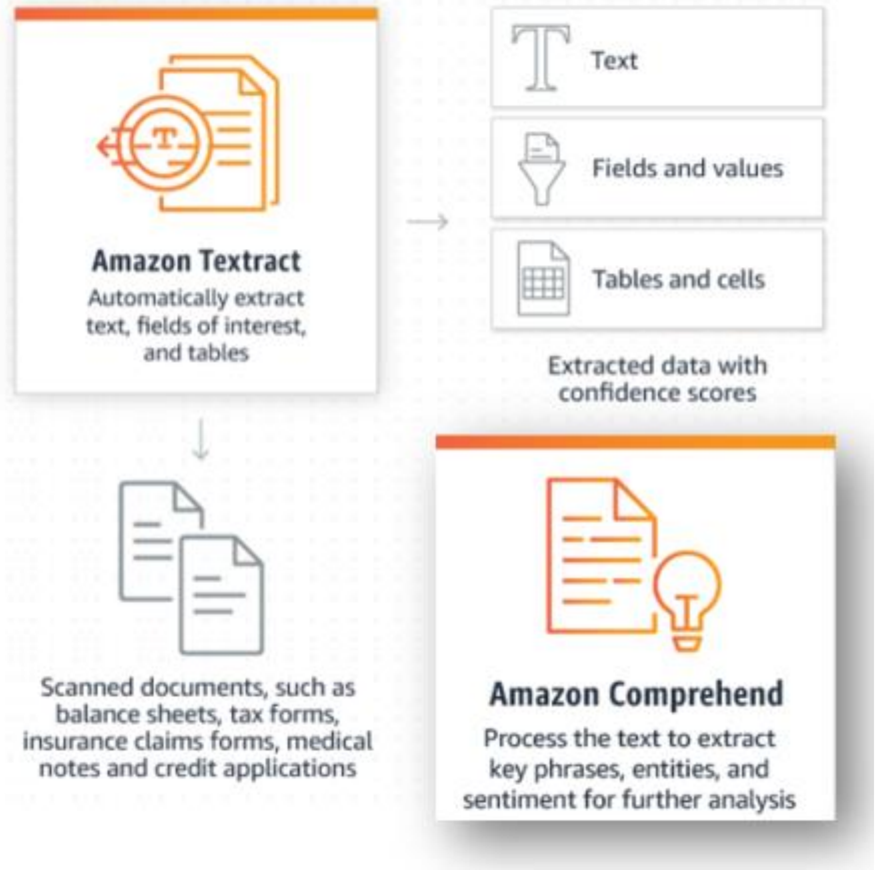
Sin fecha

SROIE 2019 *https://drive.google.com/drive/folders/1ShItNWXyiY1tFDM5W02bceHuJjyeeJl2*

Herramientas

AWS + Regex + libs

- *Amazon Textract*
- *Amazon Comprehend*
- *Elastic Beanstalk*
- Dateparser
- Spacy
- Expresiones regulares



Amazon Textract

Automatically extract text, fields of interest, and tables

Text

Fields and values

Tables and cells

Extracted data with confidence scores

Scanned documents, such as balance sheets, tax forms, insurance claims forms, medical notes and credit applications

Amazon Comprehend

Process the text to extract key phrases, entities, and sentiment for further analysis

https://www.python.org/

https://dateparser.readthedocs.io/

https://spacy.io/

Lógica del servicio

- *1. Decodificar la imagen en directorio /tmp*

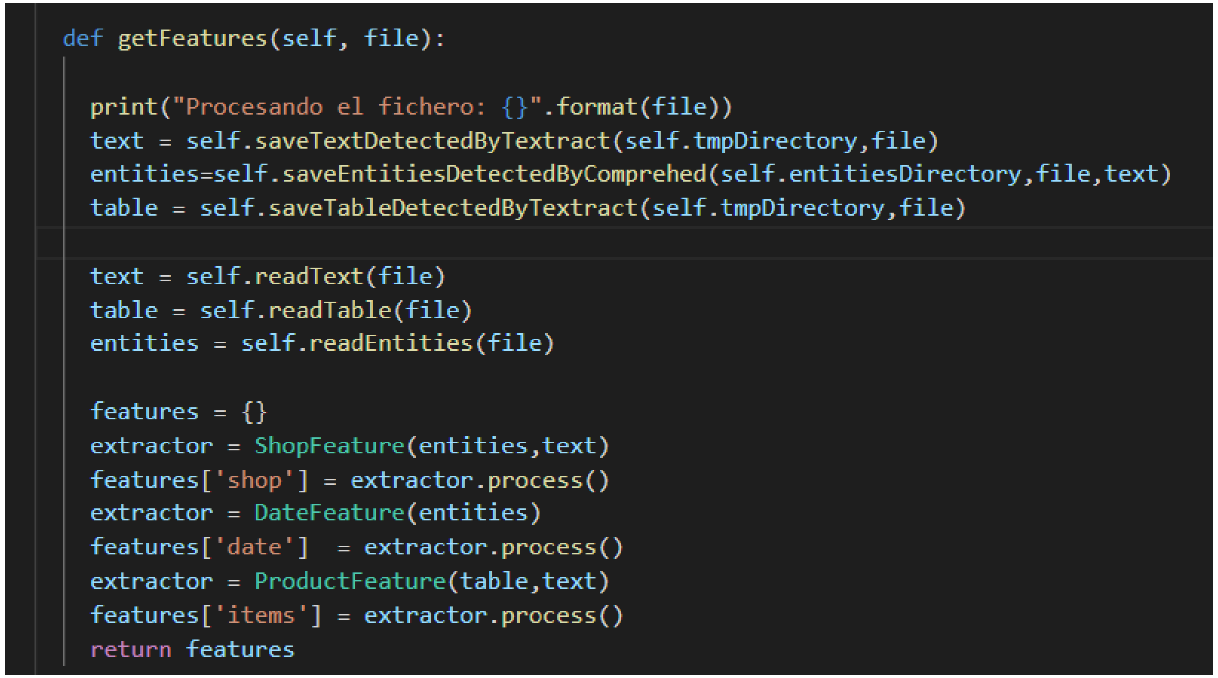
- *2. Procesar imagen con 2 servicios de Textract y 1 servicio de Comprehed. Salvar resultados en directorios*

  /text /tables /entidades

- *3. Procesar los ficheros anteriores y extraer Features*
  - *Extraer Establecimiento (ShopFeature)*
  - *Extraer Listado de productos y precio (ProductFeature)*
  - *Extraer Fecha (DateFeature)*

  (Doble validación con expresiones regulares)

- *4. Eliminar ficheros e imagen de los directorios temporales*

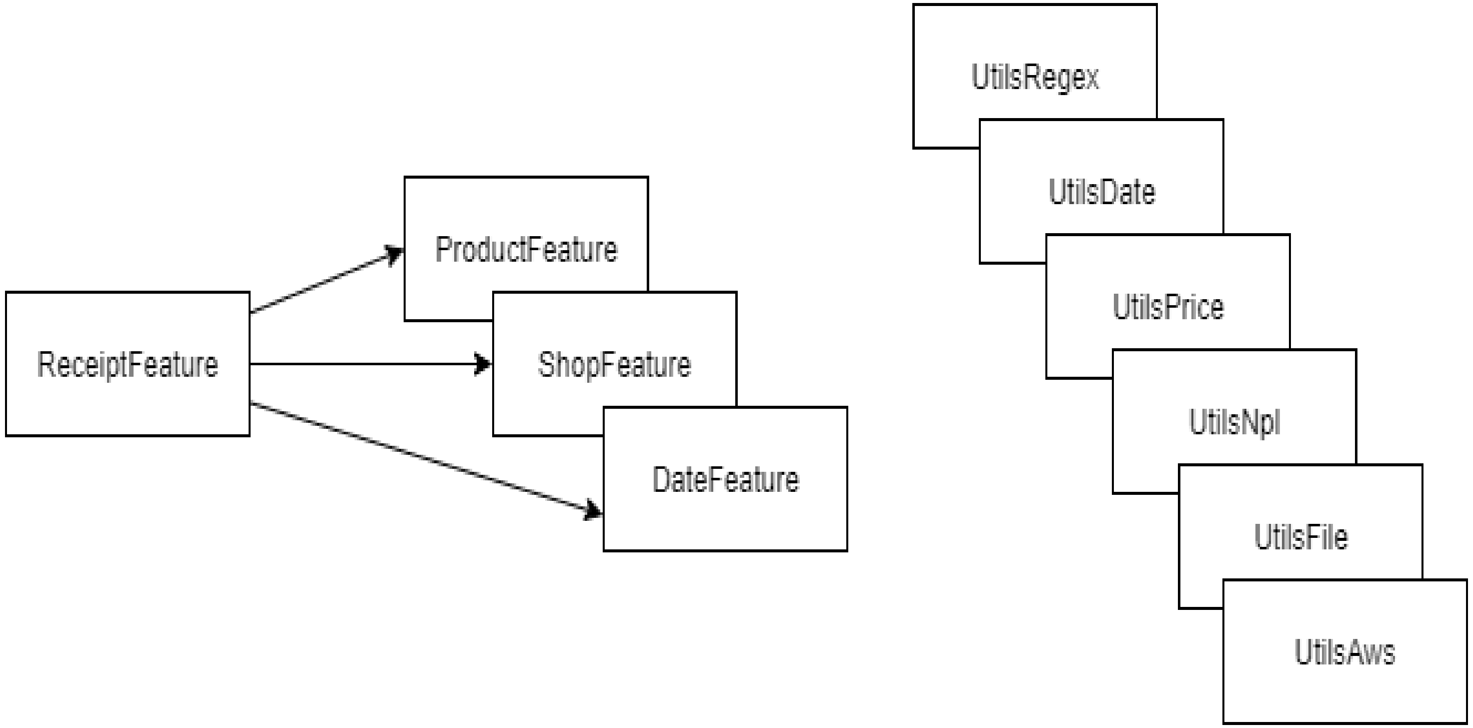


```python
def getFeatures(self, file):

    print("Procesando el fichero: {}".format(file))
    text = self.saveTextDetectedByTextract(self.tmpDirectory,file)
    entities=self.saveEntitiesDetectedByComprehed(self.entitiesDirectory,file,text)
    table = self.saveTableDetectedByTextract(self.tmpDirectory,file)

    text = self.readText(file)
    table = self.readTable(file)
    entities = self.readEntities(file)

    features = {}
    extractor = ShopFeature(entities,text)
    features['shop'] = extractor.process()
    extractor = DateFeature(entities)
    features['date']  = extractor.process()
    extractor = ProductFeature(table,text)
    features['items'] = extractor.process()
    return features
```

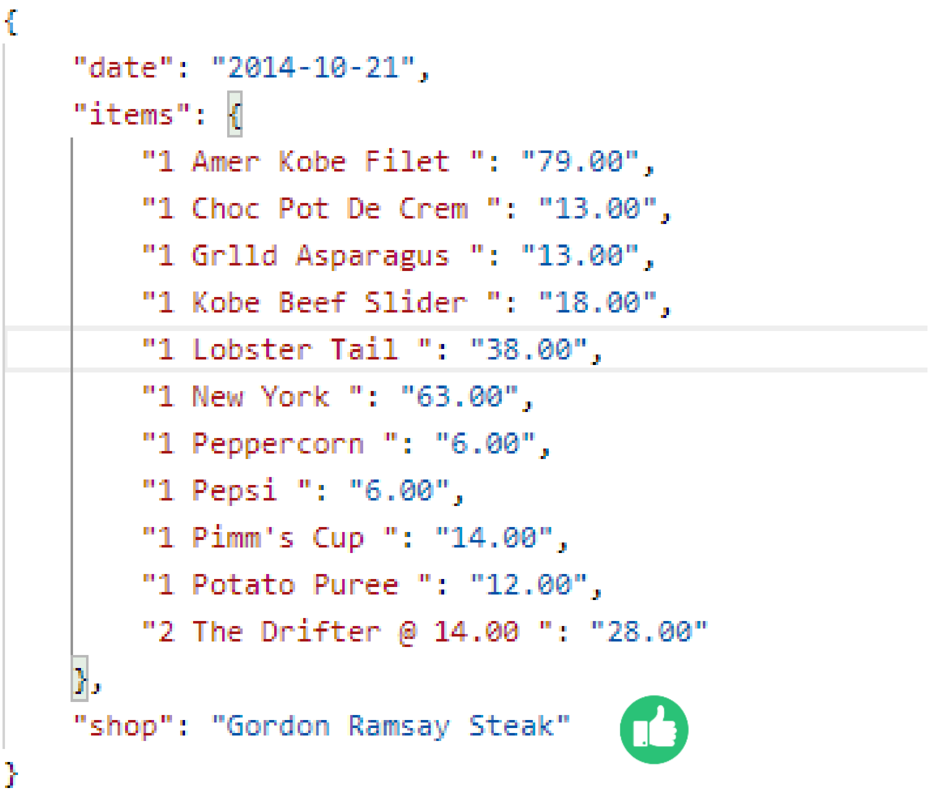
https://github.com/pilarcode/demo

Componentes

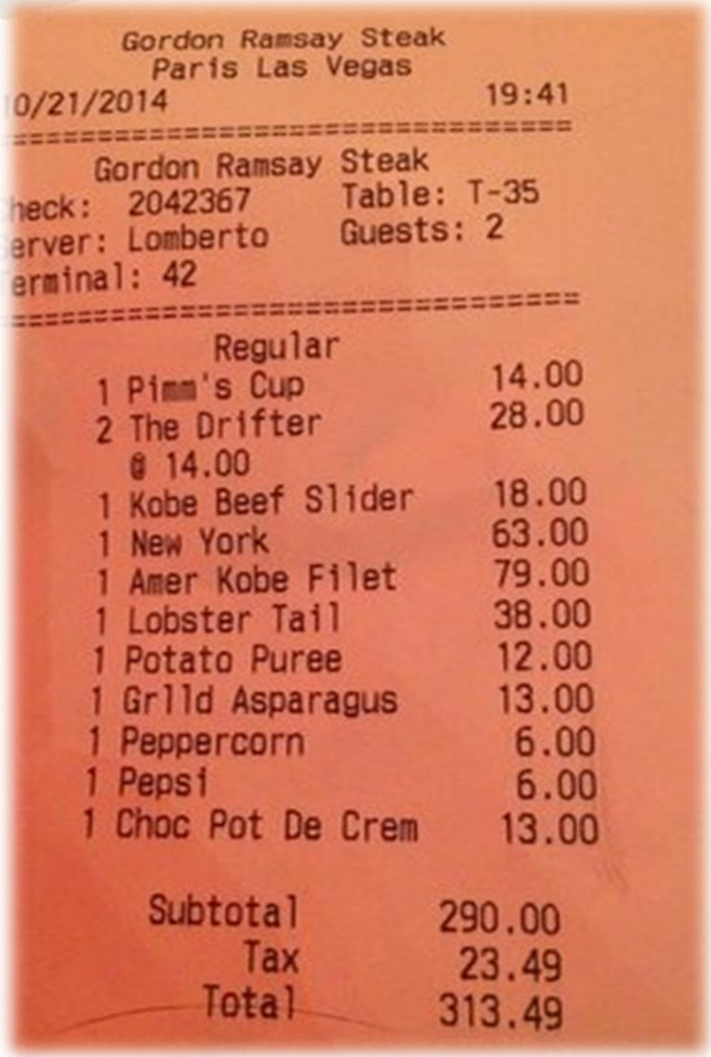
- *Application.py*
- *Utils.py*
- *ReceiptFeature.py*
- *ProductFeature.py*
- *ShopFeature.py*
- *DateFeature.py*



Tests (I)

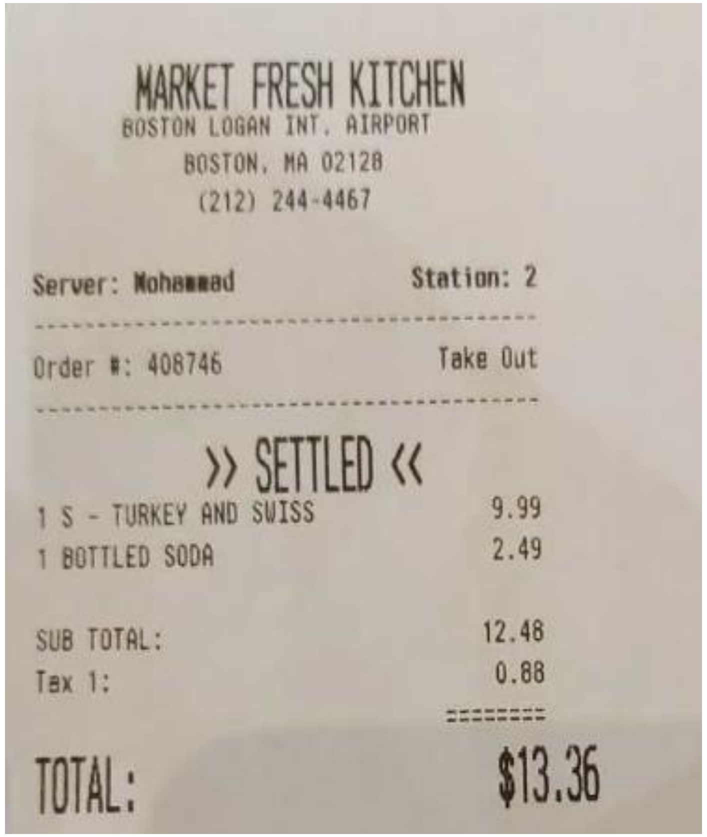


```
{
    "date": "2014-10-21",
    "items": {
        "1 Amer Kobe Filet ": "79.00",
        "1 Choc Pot De Crem ": "13.00",
        "1 Grlld Asparagus ": "13.00",
        "1 Kobe Beef Slider ": "18.00",
        "1 Lobster Tail ": "38.00",
        "1 New York ": "63.00",
        "1 Peppercorn ": "6.00",
        "1 Pepsi ": "6.00",
        "1 Pimm's Cup ": "14.00",
        "1 Potato Puree ": "12.00",
        "2 The Drifter @ 14.00 ": "28.00"
    },
    "shop": "Gordon Ramsay Steak"
}
```



```
         Gordon Ramsay Steak
          Paris Las Vegas
0/21/2014                      19:41
====================================
        Gordon Ramsay Steak
heck:   2042367         Table: T-35
erver:  Lomberto        Guests: 2
erminal: 42
====================================
              Regular
   1 Pimm's Cup                14.00
   2 The Drifter               28.00
     @ 14.00
   1 Kobe Beef Slider          18.00
   1 New York                  63.00
   1 Amer Kobe Filet           79.00
   1 Lobster Tail              38.00
   1 Potato Puree              12.00
   1 Grlld Asparagus           13.00
   1 Peppercorn                 6.00
   1 Pepsi                      6.00
   1 Choc Pot De Crem          13.00

        Subtotal              290.00
             Tax               23.49
           Total              313.49
```

Tests *(II)*



MARKET FRESH KITCHEN
BOSTON LOGAN INT. AIRPORT
BOSTON, MA 02128
(212) 244-4467

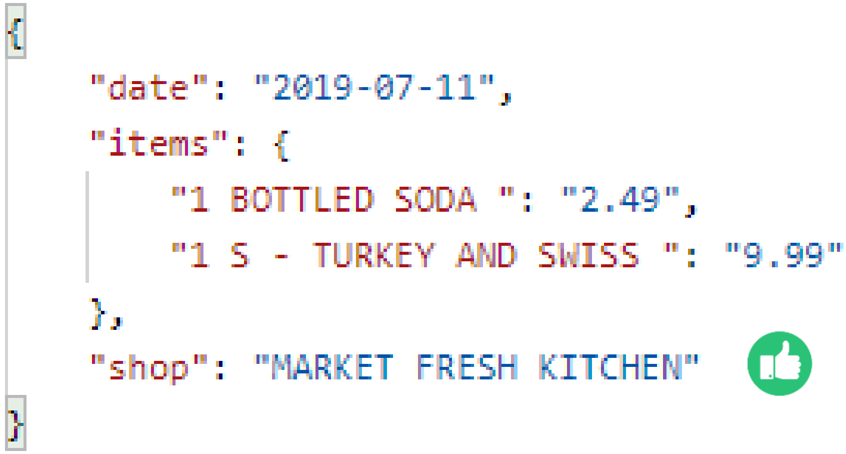
Server: Mohammad   Station: 2

Order #: 408746   Take Out

>> SETTLED <<

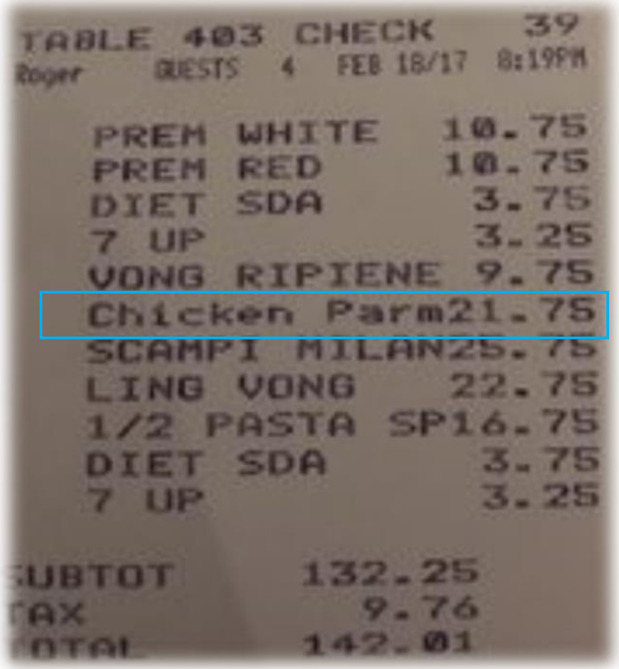
1 S - TURKEY AND SWISS	9.99
1 BOTTLED SODA	2.49
SUB TOTAL:	12.48
Tax 1:	0.88
	========
TOTAL:	$13.36

Se extrae la información aunque el ticket esta borroso.

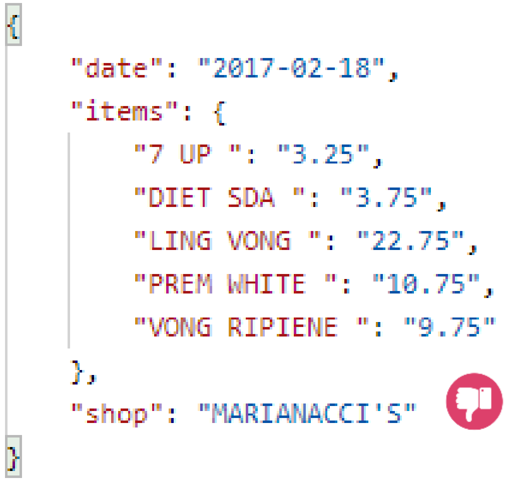


```
{
    "date": "2019-07-11",
    "items": {
        "1 BOTTLED SODA ": "2.49",
        "1 S - TURKEY AND SWISS ": "9.99"
    },
    "shop": "MARKET FRESH KITCHEN"
}
```

👍

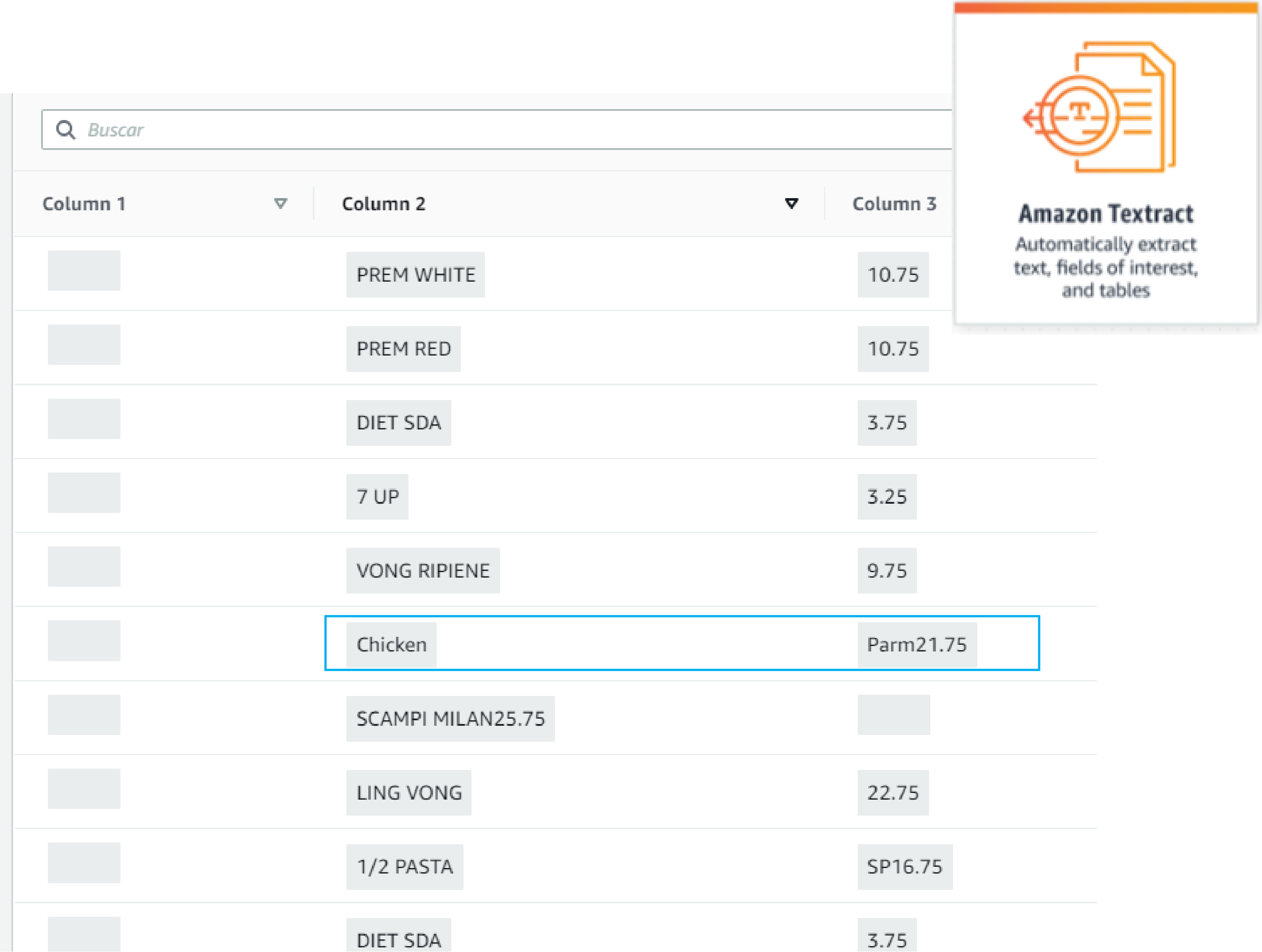


```
TABLE 403 CHECK          39
Roger   GUESTS  4  FEB 18/17  8:19PM

   PREM WHITE        10.75
   PREM RED          10.75
   DIET SDA           3.75
   7 UP               3.25
   VONG RIPIENE       9.75
   Chicken Parm21.75
   SCAMPI MILAN25.75
   LING VONG         22.75
   1/2 PASTA SP16.75
   DIET SDA           3.75
   7 UP               3.25

SUBTOT       132.25
TAX            9.76
TOTAL        142.01
```



```
{
    "date": "2017-02-18",
    "items": {
        "7 UP ": "3.25",
        "DIET SDA ": "3.75",
        "LING VONG ": "22.75",
        "PREM WHITE ": "10.75",
        "VONG RIPIENE ": "9.75"
    },
    "shop": "MARIANACCI'S"
}
```

Faltan algunos productos…

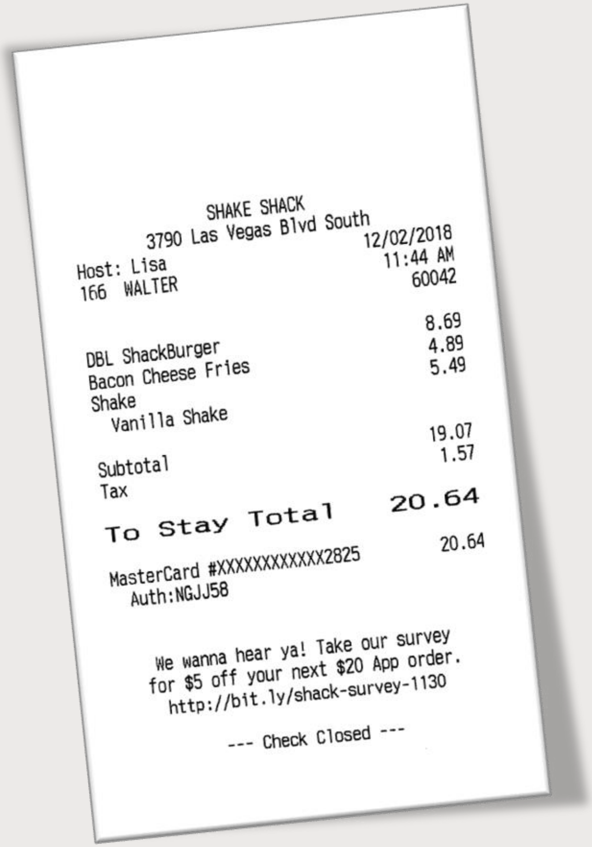


Column 1	Column 2	Column 3
	PREM WHITE	10.75
	PREM RED	10.75
	DIET SDA	3.75
	7 UP	3.25
	VONG RIPIENE	9.75
	Chicken	Parm21.75
	SCAMPI MILAN25.75	
	LING VONG	22.75
	1/2 PASTA	SP16.75
	DIET SDA	3.75

Amazon Textract
Automatically extract text, fields of interest, and tables

Algunos casos son difíciles hasta para el propio servicio de AWS.

EXTRACCIÓN DE DATOS EN *RECIBOS*



SHAKE SHACK
3790 Las Vegas Blvd South

Host: Lisa
166 WALTER

12/02/2018
11:44 AM
60042

DBL ShackBurger	8.69
Bacon Cheese Fries	4.89
Shake	5.49
Vanilla Shake	
Subtotal	19.07
Tax	1.57
To Stay Total	20.64
MasterCard #XXXXXXXXXXXX2825	20.64
Auth:NGJJ58	

We wanna hear ya! Take our survey
for $5 off your next $20 App order.
http://bit.ly/shack-survey-1130

--- Check Closed ---

¡Gracias!

¿Alguna pregunta?

- ***Pilar Madariaga Lasala***
- Estudiante Máster Inteligencia Artificial en Unir
- Pasantia en Grupo Bancolombia