In [59]:

```
!pip3 install pyPDF4
#extraction for pdf
!pip3 install pytesseract
#extraction from images
!pip3 install tesseract
```

executed in 26.6s, finished 17:53:58 2021-09-15

. . .

In [1]:

```
import os
os.getcwd()
```

executed in 35ms, finished 22:16:48 2021-09-17

Out[1]:

```
'D:\\data science\\Machine Learning(Applied AI Course)\\Mastersindia_Skill_T
est'
```

In [181]:

```
# os.chdir(".")
print(os.listdir("./ML Assignment/"))
```

executed in 10ms, finished 00:16:27 2021-09-18

```
['1291908241.pdf', '19019155.pdf', 'Decathlon Invoice.jpg', 'Decathlon Tax I
nvoice.jpg', 'IGST NO.333.pdf', 'IGST NO.334.pdf', 'Invoice  - 10001.pdf',
'Invoice - 20004.pdf', 'Sales_Invoice_1EOU192001343.pdf', 'Sales_Invoice_1EO
U192001344.pdf']
```

In [23]:

```
import re
import PyPDF4
import cv2
import numpy as np
import pytesseract
from PIL import Image
from pytesseract import image_to_string
EOF_MARKER = b'%%EOF'

pytesseract.pytesseract.tesseract_cmd = 'C:\Program Files\Tesseract-OCR\\tesseract.exe'


data = {}
files= os.listdir("./ML Assignment/")
for i in range(len(files)):
    ext = files[i].split('.')[-1]

    if ext in 'pdf':
        print(files[i])
        FILE_PATH = os.getcwd()+"//ML Assignment//"+files[i]
        #         print(FILE_PATH)
        with open(FILE_PATH, mode='rb') as f:
            reader = PyPDF4.PdfFileReader(f)
            page = reader.getPage(0)
            data[files[i]] = page.extractText()


    if ext in 'jpg':
        print(files[i])
        FILE_PATH = os.getcwd()+"//ML Assignment//"+files[i]
        #         print(FILE_PATH)
        with Image.open(FILE_PATH) as f:
            data[files[i]] = pytesseract.image_to_string(f)
print("*"*10, "Data Successfully stored in data dict","*"*10)
```

executed in 2.19s, finished 22:35:04 2021-09-17

. . .

In [250]:

```
def image_invoice_no(string):
    pattern_1 = re.compile('\d{4,5}[.-:]+\d{3,4}-\d{2,3}',re.IGNORECASE)
    pattern_2 = re.compile('\d{10,12}',re.IGNORECASE)

    if re.findall(pattern_1, string)!=[]:
        return re.findall(pattern_1, string)[0]
    elif re.findall(pattern_2, string)!=[]:
        return re.findall(pattern_2, string)[0]
    else:
        return "No Match"

def image_invoice_date(string):
    pattern= re.compile(r'\d{1,2}[-\/\.]+[0-9a-z]{1,3}[-\/\.]+\d{2,4}',re.IGNORECASE)
    if re.findall(pattern, string)!=[]:
        return re.findall(pattern, string)[0]
    else:
        return "No Match"


def line_items(string):
    pattern_1 = re.compile(r'\n.*[Description]{11,12}[\s\S]+(?=\n.*?total|$)', re.IGNOREC
    pattern_2 = re.compile(r'\n.*(?=\n.*?good)[\s\S]+(?=\n.*?total|$)', re.IGNORECASE)
    found_1 = re.findall(pattern_1, string)
    found_2 = re.findall(pattern_2, string)
    if found_1!=[]:
        for i in found_1[0].split("\n"):
            print(i)
        return ' '
    elif found_2!=[]:
        for i in found_2[0].split("\n"):
            print(i)
        return ' '
    else:
        return "No Match"

def file_data(file_path, file_name):
    ext = file_name.split('.')[-1]

    if ext in 'pdf':
        print(file_name)
        FILE_PATH = file_path+"/"+str(file_name)

        f= open(FILE_PATH, mode='rb')
        reader = PyPDF4.PdfFileReader(f)
        page = reader.getPage(0)
        string = page.extractText()
        print("Invoice Date: ",image_invoice_date(string))
        print("Invoice No.: ",image_invoice_no(string))
        print("Line Items: ",line_items(string))

    if ext in 'jpg':
        print(file_name)
        FILE_PATH = FILE_PATH = file_path+"/"+str(file_name)

        f = Image.open(FILE_PATH)
        string = pytesseract.image_to_string(f)
        print('*'*50,"Invoice Date: ",image_invoice_date(string),"\n")
        print('*'*50,"Invoice No.: ",image_invoice_no(string),"\n")
```

```
        print('*'*50,"Line Items: ",line_items(string),"\n")
```

executed in 45ms, finished 00:57:36 2021-09-18

In [251]:

```
file_path= input("Enter the file path")
file_name = input("Enter the name of file to be converted.")
file_data(file_path, file_name)
```

executed in 12.6s, finished 00:57:51 2021-09-18

```
Enter the file pathD:\data science\Machine Learning(Applied AI Course)\Maste
rsindia_Skill_Test\ML Assignment
Enter the name of file to be converted.1291908241.pdf
1291908241.pdf
Invoice Date:  21/01/2020
Invoice No.:  1291908241

Sl#Description of Goods
HSN Code
(GST)
QtyUOM
Rate
(INR)
AmountDiscount
Taxable
Value
CGST
SGST
IGST
%Amount
%
%
Amount
Amount
UP
Charges
Total Value
PCS
READYMADE GARMENTS (100% POLYESTER MENS
JACKETS WITH LINING AND PADDING )
4806
2473167.60
514.60
1
2473167.602.50
2.50
0.00
61829.19
61829.19
0.00
0.00
62033300
0.00
2596825.98
Net Net Weight:0.000
For   SHAHI EXPORTS PVT LTD
Signature and Date
Declaration :
1. We declare that this invoice shows the actual price of goods described an
d correct.
Values in Words (INR) : Twenty  Five Lakh Ninety  Six Thousand Eight Hundred
Twenty  Five   And   Ninety  Eight  Paise Only.
TOTAL #
0.00
```

```
Total Invoice Value :
61829.19
61829.19
0.00
2473167.60
2473167.
2596825.98
0.00
2596825.98
2.Reverse charge applicable : NA
Our Bank details for  payment :
YES, BANK, GR. FLOOR & SECOND FLOOR, SCO-4, SECTOR -16, FARIDABAD -121002,
SWIFT CODE: YESBINBBDEL, IFSC CODE- YESB0000020, SHAHI EXPORTS PVT. LTD,
Account  Number- 002081400000088

Line Items:
```