Scraping Data From PDF

March 6, 2022

1 Jupiter notebook that retrieves data from Facture

1.1 Importing Libraries

[1]:
```
import pandas as pd
import pdfplumber as pdfp
with pdfp.open('Factures/11.02.2022.pdf') as pdf:
    global text
    for page in pdf.pages:
        text = page.extract_text()
        file_object = open('texte.txt', 'a')
# Append 'hello' at the end of file
        file_object.write(text)
# Close the file
        file_object.close()
        print(text)
```

```
Facture
Numéro de note Date Numéro doc.
8185864998 11.02.2022 40045168
Adresse de livraison: Adresse de facturation:
RDMC RDMC
1 RUE DE LA DESTINEE 1 RUE DE LA DESTINEE
95800 CERGY 95800 CERGY
Page  1/  2
N° de commande / Date S301571478 / 09.02.2022 N° de livraison / Date  /
11.02.2022
Votre compte chez nous 757769 Contact client STREAM_ONE
Numéro TVA du client FR37880905088 Pers. contact TD- Sales Digitally Led
VNo. bon d'achat / Date See_Line_Level_31/Jan/2022 / Pers. contact TD- Finance
Valérie Lair
09.02.2022
N° d'art. Description % VAT
Ligne N° d'art.fabricant Quantité Prix unitaire Montant EUR
Del. Grp numéro EAN N° d'art. client
4194212 Acronis Backup CloudÂ#Acronis Hosted/GB  20,00 %
000010 SPBAMSENSCOMIT1     1 336,50 EUR 336,50 EUR
```

Billing Period: 01/Jan/2022 - 31/Jan/2022 (31 days)
Original StreamOne Sales Order: S000615933 (SKU: SK11456)
Your Purchase Order: 2021-07-22_31/Jan/2022
End User: RDMC
Product:CPC_PG Acronis Hosted Storage Quantity:3365.04
Total H.T. 336,50 EUR
Bases T.V.A. 336,50 EUR
Taux - Base de calcul - montant T.V.A. 20,00 % 336,50 EUR 67,30 EUR
NET A PAYER 403,80 EUR
RPCP:
Notice officielle d.information sur la copie privée à :
http://www.copieprivee.culture.gouv.fr
Remboursement/exonération de la rémunération pour usage professionnel,
http://www.copiefrance.fr
Référence à mentionner lors du règlement ou pour toutreé clamation: 8185864998
Mode d'expédition: PA - Paris
Mode de règlement: 30 jours nets / Traite LCR
Date d'échéance: 13.03.2022
N° TVA Tech Data: FR07722065638
Ces biens sont contrôlés par les réglementations du gouvernement des États-Unis et de l'UE sur le contrôle des exportations. S'ils
sont exportés ou réexportés, ils sont autorisés à l'exportation uniquement vers le pays de destination finale, en vue d'être utilisés par
le destinataire final ou le(s) utilisateur(s) final(s) identifié(s) dans le présent document. Ils ne peuvent pas être revendus ou transférés
à une personne autre que le destinataire final ou le(s) utilisateur(s) final(s) autorisé(s) et ils ne peuvent pas être revendus, transférés
ou mis au rebut de quelque manière que ce soit dans un autre pays, que ce soit sous leur forme originale ou après avoir été
intégrés à d'autres biens, sans l'obtention préalable de l'approbation nécessaire du gouvernement des États-Unis, de l'UE ou du
gouvernement national compétent ou selon les modalités autorisées par la loi et les réglementations des États-Unis.
Facture en EUR, merci d'adresser votre règlement au RIB suiva n:t
SWIFT code: CITIFRPP IBAN: FR76 1168 9007 0000 6555 1301 926
Conformément aux Conditions Générales de Vente de Tech Data France, un escompte pour paiement anticipé calculé au taux
annualisé de deux virgule cinq pourcent (2,5%) sera consenti par TDF aux Clients bénéficiant d'une ligne de crédit en cas de
paiement avec réception des fonds dans les dix (10) jours suivant la date de la facture.
Cet escompte est calculé sur le montant net encaissé.
Tout paiement postérieur à la date d.échéance entraîne l.exigibilité de pénalités de retard dont le taux est égal à trois (3) fois le taux
d.intérêt légal en vigueur en France. Par ailleurs, conformément à l.article L. 441-10 du Code de commerce, tout retard de paiement
entraîne de plein droit et sans qu.un rappel ne soit nécessaire, outre les

intérêts de retard susmentionnés, une obligation pour le
Client de payer une indemnité forfaitaire pour frais de recouvrement de
quarante euros hors taxes (40 euros H.T.) par facture.
La vente faisant l'objet des présentes est, de l'accord formel des parties,
régie par les conditions générales de vente TECH DATA
France que l'acquéreur déclare expressément accepter,
notamment la clause de réserve de propriété des biens vendusa,u profit de Tech
Data, jusqu'au paiement intégral du prix.
TECH DATA France S.A.S TVA: FR 07 722065638 Veuillez addresser votre
règlement
au capital de 77.995.212 Euros 722 065 638 RCS Meaux à TECH DATA
France
APE 4651Z Comptabililé Clients Collection Colombes
Siège Social & Bureaux Agence Colombes
5, avenue de I.Europe Bâtiment les corvettes web: www.techdata.fr
Bussy Saint-Georges 142, Avenue de Stalingrad Comptabililé Clients Collection
Bussy
77611 Marne-la-Vallée Cedex 3 92714 Colombes Cedex
Facture
Numéro de note Date Numéro doc.
8185864998 11.02.2022 40045168
Page 2/ 2
Pour visualiser vos factures, les CGV Tech Data France, ou déclarer vos litiges
(rubrique « service client online)
rendez-vous sur www.techdata.fr
Tech Data France S.A.S. (" nous") et Tech Data UK Resources Limited (" TD UK")
vous notifient par la présente qu'en tant que
propriétaire légal et bénéficiaire de cette facture, avant la date d'échéance
de cette facture, nous céderons dans chaque cas
complètement tous nos droits, titres, intérêts et bénéfices liés à la créance
résultant de cette facture (ou, le cas échéant, desdites
factures antérieures) (les " Créances") à TD UK.
Veuillez noter que (i) vous pouvez continuer à traiter avec nous en ce qui
concerne les Créances, (ii) vous devez continuer à
effectuer les paiements à notre bénéfice jusqu'à ce que vous receviez une
notification contraire de la part de TD UK, (iii) nous vous
donnons irrévocablement et inconditionnellement instruction et autorisation
(cette autorisation ne pouvant être ni révoquée ni
modifiée, malgré toute instruction précédente contraire), à la réception d'une
notification de TD UK conformément à la précédente
clause (ii) ci-dessus, de payer toute somme que vous nous devez au titre des
Créances à TD UK (et non nous) ou tout autre compte
que TD UK spécifie, et (iv) vous êtes autorisés à communiquer toute information
relative aux Créances à TD UK, à la demande de
ce dernier. TD UK et nous-mêmes restons à tout moment seuls responsables envers
vous de l'exécution de toutes les obligations
relatives aux Créances.
TECH DATA France S.A.S TVA: FR 07 722065638 Veuillez addresser votre

```
règlement
au capital de 77.995.212 Euros 722 065 638 RCS Meaux              à TECH DATA
France
APE 4651Z Comptabililé Clients Collection Colombes
Siège Social & Bureaux Agence Colombes
5, avenue de I.Europe Bâtiment les corvettes web: www.techdata.fr
Bussy Saint-Georges 142, Avenue de Stalingrad Comptabililé Clients Collection
Bussy
77611 Marne-la-Vallée Cedex 3 92714 Colombes Cedex
```

[2]:
```
# READ FILE
df = open("texte.txt")
# read file
read = df.read()
df.seek(0)
#read
```

[2]:
```
0
```

[3]:
```
arr = []
# count number of
# lines in the file
line = 1
for word in text:
        if word == '\n':
                line += 1
print("Number of lines in file is: ", line)
```

```
Number of lines in file is:  25
```

[4]:
```
labels = ["Numéro de note","Numéro doc",
          "Adresse de livraison","Adresse de facturation",
          "N° de commande","Votre compte chez nous","Numéro TVA du client",
          "VNo. bon d'achat","Contact  client","Pers.  contact  TD-  Sales",
          "Pers.  contact  TD-  Finance","Total H.T.","Taux - Base de␣
 ↪calcul","NET A PAYER","Mode d'expédition:",
          "Mode de règlement:","Date d'échéance","N° TVA Tech Data:"]
#outputs = []
list = []
for word in labels:
    for i in range(line):
        arr.append(str(df.readline()))
#global findline
    def findline(word):
        #list = [] ##############################
        for i in range(len(arr)):
            #list = []
```

```
            if word in arr[i] and 0 < i < 59:
                #for i in range(line):
                  ######################
                print( word, i+1)  ########////////////////// i+1, end=" "
                list.append(i+1)
        return list

    words = findline(word)
#print (a)  ########################################
#words = outputs.append(findline)

print("Words List Line :", words)
```

```
Numéro de note 2
Numéro doc 2
Adresse de livraison 4
Adresse de facturation 4
N° de commande 9
Votre compte chez nous 10
Numéro TVA du client 11
VNo. bon d'achat 12
Total H.T. 25
Taux - Base de calcul 27
NET A PAYER 28
Mode d'expédition: 33
Mode de règlement: 34
Date d'échéance 35
N° TVA Tech Data: 36
Words List Line : [2, 2, 4, 4, 9, 10, 11, 12, 25, 27, 28, 33, 34, 35, 36]
```

1.1.1 Numéro_de_note

[5]:
```
a = read.split('\n')[words[0]]
a
```

[5]:
```
'8185864998 11.02.2022 40045168'
```

[6]:
```
Numéro_de_note = a[0:10]
print(Numéro_de_note)
```

```
8185864998
```

1.1.2 Date

[7]:
```
Date = a[11:21]
print(Date)
```

```
11.02.2022
```

1.1.3 Numéro_doc

[8]:
```
Numéro_doc = a[22:30]
print(Numéro_doc)
```

```
40045168
```

1.1.4 Adresse_de_livraison

[9]:
```
Adresse_de_livraison = read.split('\n')[words[3]+1]
print(Adresse_de_livraison)
```

```
1 RUE DE LA DESTINEE 1 RUE DE LA DESTINEE
```

1.1.5 Adresse_de_facturation

[10]:
```
Adresse_de_facturation=read.split('\n')[words[3]+2]
print(Adresse_de_facturation)
```

```
95800 CERGY 95800 CERGY
```

1.1.6 Votre_compte_chez_nous

[11]:
```
b=read.split('\n')[words[4]]
b
```

[11]:
```
'Votre compte chez nous 757769 Contact client STREAM_ONE'
```

[12]:
```
Votre_compte_chez_nous=b[23:29]
print(Votre_compte_chez_nous)
```

```
757769
```

1.1.7 Contact_client

[13]:
```
Contact_client=b[45:68]
print(Contact_client)
```

```
STREAM_ONE
```

1.1.8 num_tva_client

[14]:
```
read.split('\n')[words[5]]
```

[14]:
```
'Numéro TVA du client FR37880905088 Pers. contact TD- Sales Digitally Led'
```

[15]:
```
c = read.split('\n')[words[5]]
```

[16]:
```
num_tva_client=c[21:34]
print(num_tva_client)
```

```
FR37880905088
```

1.1.9 contact_td_sales

[17]:
```
c[59:72]
```

[17]:
```
'Digitally Led'
```

[18]:
```
contact_td_sales=c[59:72]
print(contact_td_sales)
```

```
Digitally Led
```

1.1.10 num_voucher

[19]:
```
read.split('\n')[words[6]]
```

[19]:
```
"VNo. bon d'achat / Date See_Line_Level_31/Jan/2022 / Pers. contact TD- Finance
Valérie Lair"
```

[20]:
```
read.split('\n')[words[6]].replace('\t', '/').split("/ ")[1]
```

[20]:
```
'Date See_Line_Level_31/Jan/2022 '
```

[21]:
```
d= read.split('\n')[words[6]].replace('\t', '/').split("/ ")[1]
num_voucher=d[5:]
print(num_voucher)
```

See_Line_Level_31/Jan/2022

[22]:
```
num_voucher=d[5:]
print(num_voucher)
```

See_Line_Level_31/Jan/2022

1.1.11 contact_td_finance

[23]:
```
contact_td_finance=read.split('\n')[words[6]].replace('\t', '/').split("Finance␣
↪")[1]
print(contact_td_finance)
```

Valérie Lair

1.1.12 num_commande

[24]:
```
read.split('\n')[words[4]-1]
```

[24]: 'N° de commande / Date S301571478 / 09.02.2022 N° de livraison / Date / 11.02.2022'

[25]:
```
e=read.split('\n')[words[4]-1].replace('\t', '/ ').split("/ ")[1]
num_commande=e[5:]
print(num_commande)
```

S301571478

[26]:
```
num_commande=e[5:]
print(num_commande)
```

S301571478

1.1.13 date_commande

[27]:
```
f=read.split('\n')[words[4]-1].replace('\t', '/ Date').split("N° de␣
↪livraison")[0]
date_commande=f[35:]
print(date_commande)
```

09.02.2022

1.1.14 Total H.T

[28]:
```
read.split('\n')[words[8]-1]
```

[28]:
```
'Total H.T. 336,50 EUR'
```

[29]:
```
Total_HT=read.split('\n')[words[8]-1].replace('\t', '').split("Total H.T. ")[1]
print(Total_HT)
```

```
336,50 EUR
```

1.1.15 Bases_TVA

[30]:
```
read.split('\n')[words[8]]
```

[30]:
```
'Bases T.V.A. 336,50 EUR'
```

[31]:
```
Bases_TVA=read.split('\n')[words[8]].replace('\t', '').split("Bases T.V.A. ")[1]
print(Bases_TVA)
```

```
336,50 EUR
```

1.1.16 Taux

[32]:
```
j=read.split('\n')[words[8]+1].replace('\t', '').split("Taux - Base de calcul -␣
↪montant T.V.A. ")[1]
Taux=j[:7]
print(Taux)
```

```
20,00 %
```

1.1.17 Base_de_calcul

[33]:
```
k=read.split('\n')[words[8]+1].replace('\t', '').split(" % ")[1]
Base_de_calcul=k[:10]
print(Base_de_calcul)
```

```
336,50 EUR
```

1.1.18 Montant_TVA

[34]:
```
Montant_TVA=read.split('\n')[words[8]+1].replace('\t', '').split("EUR ")[1]
print(Montant_TVA)
```

```
67,30 EUR
```

1.1.19 NET_A_PAYER

[35]:
```
NET_A_PAYER=read.split('\n')[words[9]].replace('\t', '').split("NET A PAYER␣
↪")[1]
print(NET_A_PAYER)
```

```
403,80 EUR
```

1.1.20 Mode_d_expédition

[36]:
```
Mode_d_expédition=read.split('\n')[words[11]-1].replace('\t', '').split("Mode␣
↪d'expédition: ")[1]
print(Mode_d_expédition)
```

```
PA - Paris
```

1.1.21 Mode_de_règlement

[37]:
```
read.split('\n')[words[11]]
```

[37]:
```
'Mode de règlement: 30 jours nets   / Traite LCR'
```

[38]:
```
Mode_de_règlement=read.split('\n')[words[11]].replace('\t', '').split("Mode de␣
↪règlement: ")[1]
print(Mode_de_règlement)
```

```
30 jours nets   / Traite LCR
```

1.1.22 Date_d_échéance

[39]:
```
Date_d_échéance=read.split('\n')[words[12]].replace('\t', '').split("Date␣
↪d'échéance: ")[1]
print(Date_d_échéance)
```

```
13.03.2022
```

1.1.23 num_tva_tech_data

[40]:
```
num_tva_tech_data=read.split('\n')[words[13]].replace('\t', '').split("N° TVA␣
 ↪Tech Data: ")[1]
print(num_tva_tech_data)
```

FR07722065638

1.1.24 DataFrame Creation

[41]:
```
import pandas as pd
lst = [[
Date,
Numéro_de_note,
Numéro_doc,
Adresse_de_livraison,
Adresse_de_facturation,
Votre_compte_chez_nous,
Contact_client,
num_tva_client,
contact_td_sales,
num_voucher,
contact_td_finance,
num_commande,
date_commande,
Total_HT,
Bases_TVA,
Base_de_calcul,
Montant_TVA,
NET_A_PAYER,
Mode_d_expédition,
Mode_de_règlement,
Date_d_échéance,
num_tva_tech_data]]

df = pd.DataFrame(lst, columns =['Date',
                                 'Numéro_de_note ',
                                 'Numéro_doc.',
                                 'Adresse_de_livraison',
                                 'Adresse_de_facturation',
                                 'Votre_compte_chez_nous',
                                 'Contact_client',
                                 'num_tva_client',
                                 'contact_td_sales',
                                 'num_voucher',
                                 'contact_td_finance',
```

```
                                    'num_commande',
                                    'date_commande',
                                    'Total_HT',
                                    'Bases_TVA',
                                    'Base_de_calcul',
                                    'Montant_TVA',
                                    'NET_A_PAYER',
                                    'Mode_d_expédition',
                                    'Mode_de_règlement',
                                    'Date_d_échéance',
                                    'num_tva_tech_data'], dtype = int)
df.head()
```

[41]:
```
         Date Numéro_de_note  Numéro_doc.  \
0  11.02.2022      8185864998    40045168

                        Adresse_de_livraison   Adresse_de_facturation  \
0  1 RUE DE LA DESTINEE 1 RUE DE LA DESTINEE  95800 CERGY 95800 CERGY

  Votre_compte_chez_nous Contact_client num_tva_client contact_td_sales  \
0                 757769     STREAM_ONE  FR37880905088    Digitally Led

                   num_voucher  … date_commande    Total_HT   Bases_TVA  \
0  See_Line_Level_31/Jan/2022   …   09.02.2022   336,50 EUR  336,50 EUR

  Base_de_calcul Montant_TVA NET_A_PAYER Mode_d_expédition  \
0     336,50 EUR   67,30 EUR  403,80 EUR        PA - Paris

              Mode_de_règlement Date_d_échéance num_tva_tech_data
0  30 jours nets   / Traite LCR      13.03.2022     FR07722065638

[1 rows x 22 columns]
```

1.1.25 DataFrame To CSV File

[42]:
```
df.to_csv("data.csv",index=False)
data=pd.read_csv("data.csv")
data.head()
```

[42]:
```
         Date  Numéro_de_note   Numéro_doc.  \
0  11.02.2022       8185864998     40045168

                        Adresse_de_livraison   Adresse_de_facturation  \
0  1 RUE DE LA DESTINEE 1 RUE DE LA DESTINEE  95800 CERGY 95800 CERGY

   Votre_compte_chez_nous Contact_client num_tva_client contact_td_sales  \
```

```
0                  757769     STREAM_ONE  FR37880905088    Digitally Led

                   num_voucher  … date_commande    Total_HT   Bases_TVA  \
0  See_Line_Level_31/Jan/2022   …   09.02.2022   336,50 EUR  336,50 EUR

  Base_de_calcul Montant_TVA NET_A_PAYER Mode_d_expédition  \
0     336,50 EUR   67,30 EUR  403,80 EUR        PA - Paris

              Mode_de_règlement Date_d_échéance num_tva_tech_data
0  30 jours nets   / Traite LCR      13.03.2022     FR07722065638

[1 rows x 22 columns]
```

1.1.26 DataFrame To JSON File

[43]:
```
import json
df.to_json("data.json")
data_json=pd.read_json("data.json")
data_json.head()
```

[43]:
```
        Date  Numéro_de_note   Numéro_doc.  \
0 2022-11-02       8185864998     40045168

                        Adresse_de_livraison   Adresse_de_facturation  \
0  1 RUE DE LA DESTINEE 1 RUE DE LA DESTINEE  95800 CERGY 95800 CERGY

   Votre_compte_chez_nous Contact_client num_tva_client contact_td_sales  \
0                  757769     STREAM_ONE  FR37880905088    Digitally Led

                   num_voucher  … date_commande    Total_HT   Bases_TVA  \
0  See_Line_Level_31/Jan/2022   …   09.02.2022   336,50 EUR  336,50 EUR

  Base_de_calcul Montant_TVA NET_A_PAYER Mode_d_expédition  \
0     336,50 EUR   67,30 EUR  403,80 EUR        PA - Paris

              Mode_de_règlement Date_d_échéance num_tva_tech_data
0  30 jours nets   / Traite LCR      13.03.2022     FR07722065638

[1 rows x 22 columns]
```

[ ]: