Project_python handbook

1. Extract data from the tables in pdf format document

Code in Jupyter Notebook (screenshot):

```
In [ ]: import camelot
        import pdfplumber
        import pandas as pd

        file_name='1-s2.0-S0047248414000864-main.pdf'  # name of your document
        tables = camelot.read_pdf(file_name,flavor='stream',pages='all')  # if use stream
        #tables = camelot.read_pdf(file_name,pages='all')       #if use lattice!!!!!!!!!!

        print(tables)
        #tables[0]
        print(tables[0].parsing_report)

        export_file_name=file_name+'.xlsx'
        tables.export(export_file_name, f='excel')
```

Code in text:

```
import camelot
import pdfplumber
import pandas as pd

file_name='1-s2.0-S0047248414000864-main.pdf'    # name of your document
tables = camelot.read_pdf(file_name,flavor='stream',pages='all')      # if use stream
#tables = camelot.read_pdf(file_name,pages='all')           #if use lattice!!!!!!!!!!

print(tables)
#tables[0]
print(tables[0].parsing_report)

export_file_name=file_name+'.xlsx'
tables.export(export_file_name, f='excel')
```

---

2. Extract DOI from paper

Code in Jupyter Notebook (screenshot):

```
In [ ]: # Get doi

In [ ]: import PyPDF2
        import re
        from urlextract import URLExtract
        import pandas as pd
        import numpy as np
```

```
In [ ]: # Open The File in the Command
name="paper1"
file_name=name+".pdf"
file = open(file_name, 'rb')
readPDF = PyPDF2.PdfReader(file)

print(file_name)

print(len(readPDF.pages))
```

```
In [ ]: extractor = URLExtract()
li = []


# Iterating over all the pages of File
for page_no in range(len(readPDF.pages)):
   page=readPDF.pages[page_no]
   #Extract the text from the page
   text = page.extract_text()
   text2= text.replace("\n", "")
   #print(text2)
   urls = extractor.find_urls(text2)
   for i in urls:
       li.append(i)

#for i in li:
    #print(i)
 #   for ii in i:
#print(li)

data = pd.DataFrame(data=li)

print(data)

file_export_name=name+" url.xls"

data.to_csv(file_export_name,index=False)

 # Print all URL
   #print(find_url(text2))
# CLost the file
file.close()
```

Code in text:

```
import PyPDF2
import re
from urlextract import URLExtract
import pandas as pd
import numpy as np

# Open The File in the Command
name="paper1"
file_name=name+".pdf"
file = open(file_name, 'rb')
readPDF = PyPDF2.PdfReader(file)

print(file_name)

print(len(readPDF.pages))

extractor = URLExtract()
li = []
```

```
# Iterating over all the pages of File
for page_no in range(len(readPDF.pages)):
    page=readPDF.pages[page_no]
    #Extract the text from the page
    text = page.extract_text()
    text2= text.replace("\n", "")
    #print(text2)
    urls = extractor.find_urls(text2)
    for i in urls:
        li.append(i)

#for i in li:
      #print(i)
 #     for ii in i:
#print(li)

data = pd.DataFrame(data=li)

print(data)

file_export_name=name+" url.xls"

data.to_csv(file_export_name,index=False)

 # Print all URL
    #print(find_url(text2))

# Clost the file
file.close()
```

---

3. Paper download

4. Extract table data

(1) Extract table data from a single file

Code in Jupyter Notebook (screenshot):

In [ ]:
```
file_name='1-s2.0-S0047248414000864-main.pdf'
tables = camelot.read_pdf(file_name,flavor='stream',pages='all')
#tables = camelot.read_pdf(file_name,pages='all')#if use lattice!!!!!!!!!!
print(tables)

##tables[0]
##print(tables[0].parsing_report)
#for i in range(10):
    #print(tables[i].df)
##tables[3].df
```

In [26]:
```
file_name=r'C:\Users\Sheng\Paper_failed\10.1046%j.1365-2699.2000.00431.x.pdf'
tables = camelot.read_pdf(file_name,flavor='stream',pages='3-12,14-16,19-20')
#tables = camelot.read_pdf(file_name,pages='all')#if use lattice!!!!!!!!!!
print(tables)

##tables[0]
##print(tables[0].parsing_report)
#for i in range(10):
    #print(tables[i].df)
##tables[3].df

export_file_name=file_name+'.xlsx'
tables.export(export_file_name, f='excel')
```

In [63]:
```
#import ctypes
#from ctypes.util import find_library
#print(find_library("".join(("gsdll", str(ctypes.sizeof(ctypes.c_voidp) * 8), ".dll"))))
import camelot
import pdfplumber
import pandas as pd
#import camelot.io as camelot
#import cv2
```

In [69]:
```
table_settings = {
   "vertical_strategy": "text",
    "horizontal_strategy": "text"
}

pdf = pdfplumber.open(r'C:\Users\Sheng\Paper_failed\10.1046%j.1365-2699.2000.00431.x.pdf')

df=pd.DataFrame()

i=0

# for page in pdf.pages:
#     i+=1
#     file_export_name=str(i)
#     table=page.extract_table(table_settings)
#    # print(table)
#     #print(table[1::])
#    # df2=pd.DataFrame(table[1::],columns=table[0])
#     df2=pd.DataFrame(table)
#     #print(df.append(df2))
#     with pd.ExcelWriter(file_export_name) as writer:
#         df.append(df2).to_excel(writer)


for page in pdf.pages:
    i+=1
    file_export_name=str(i)
    table=page.extract_table(table_settings)
   # print(table)
    #print(table[1::])
   # df2=pd.DataFrame(table[1::],columns=table[0])
    df2=pd.DataFrame(table)
    #print(df.append(df2))
    df3=df.append(df2)
    df3.to_csv(file_export_name,index=False)

#     with pd.ExcelWriter('output3.xlsx') as writer:
#         df.to_excel(writer, "sheet"+str(i))
    # #     df.to_excel("output.xlsx")
#     df.to_excel("output2.xlsx",sheet_name=i)
```

(2) Extract table from files in a dir

Code in Jupyter Notebook (screenshot):

In [20]:

```
import pandas as pd
import numpy as np
import os
from pathlib import Path
import time
import camelot
import pdfplumber
import pandas as pd


p = Path(r'C:\Users\Sheng\Paper_total')

# 所有以pdf结尾的文件
for file in p.rglob('*.pdf'):
    file_name=file.__str__()
    print(file_name)
    #print(type(file_name))
    # 直接遍历出文件绝对路径
    tables = camelot.read_pdf(file_name,flavor='stream',pages='all')
    #tables = camelot.read_pdf(file_name,pages='all')#if use lattice!!!!!!!!!!
    print(tables)
    export_file_name=file_name+'.xlsx'
    tables.export(export_file_name,f='excel')
    #tables.export(export_file_name,'C:/Users/Sheng/Paper_test',f='excel')
```