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')
```

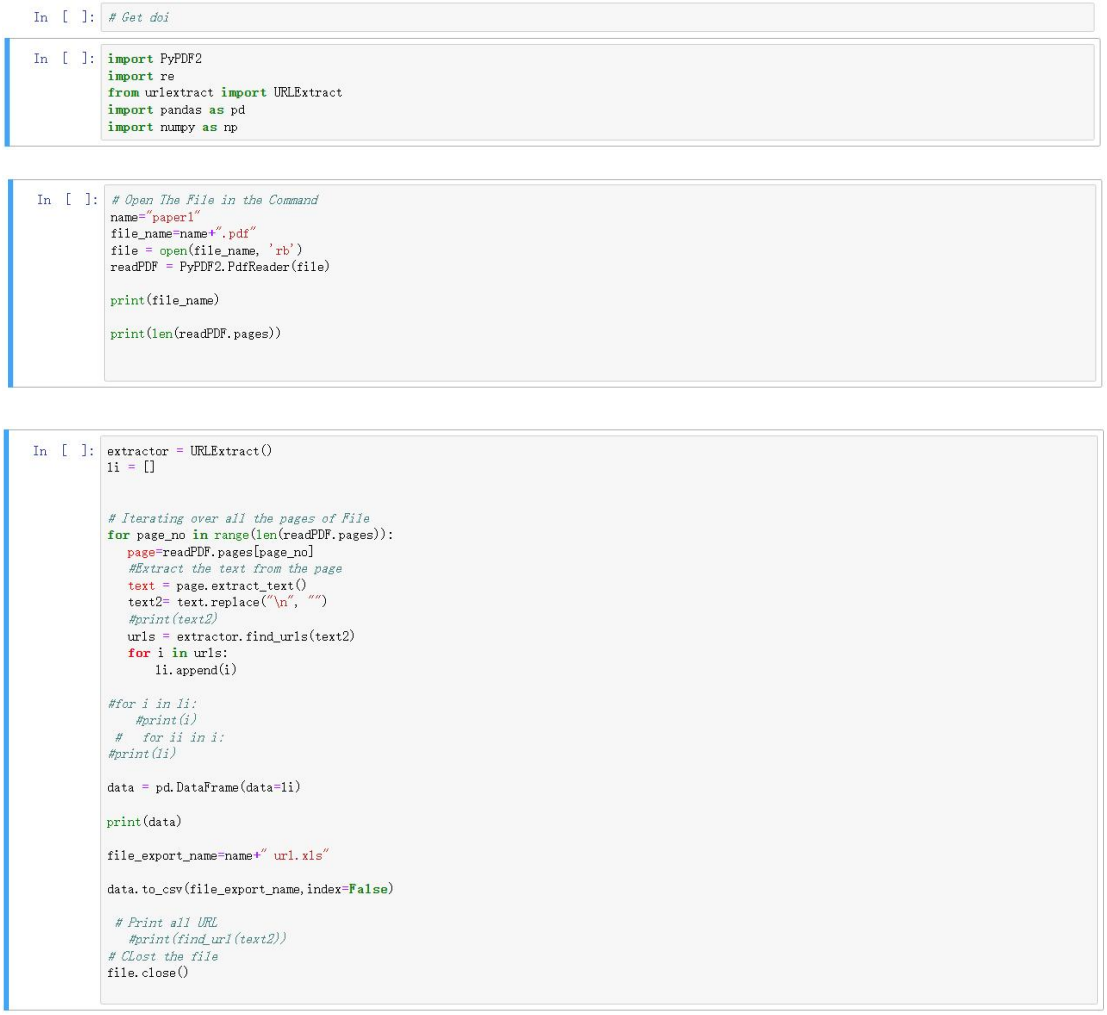
Code in text:

```
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()
```