PRACTICAL 2

Name: Mihir Bhanderi

Roll : 19BCE023

Batch : A1

```
pip install tabula-py
```

```
Collecting tabula-py
  Downloading tabula_py-2.2.0-py3-none-any.whl (11.7 MB)
     |████████████████████████████████| 11.7 MB 225 kB/s
Collecting distro
  Downloading distro-1.6.0-py2.py3-none-any.whl (19 kB)
Requirement already satisfied: pandas>=0.25.3 in /usr/local/lib/python3.7/dist-packages (from tabula-py) (1.1.5)
Requirement already satisfied: numpy in /usr/local/lib/python3.7/dist-packages (from tabula-py) (1.19.5)
Requirement already satisfied: pytz>=2017.2 in /usr/local/lib/python3.7/dist-packages (from pandas>=0.25.3->tabula-py) (2018.9)
Requirement already satisfied: python-dateutil>=2.7.3 in /usr/local/lib/python3.7/dist-packages (from pandas>=0.25.3->tabula-py) (2.8.1)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/dist-packages (from python-dateutil>=2.7.3->pandas>=0.25.3->tabula-py) (1.15.0)
Installing collected packages: distro, tabula-py
Successfully installed distro-1.6.0 tabula-py-2.2.0
```

```
[ ] pip install camelot-py
```

```
Collecting camelot-py
  Downloading camelot_py-0.10.1-py3-none-any.whl (40 kB)
     |████████████████████████████████| 40 kB 21 kB/s
Collecting PyPDF2>=1.26.0
  Downloading PyPDF2-1.26.0.tar.gz (77 kB)
     |████████████████████████████████| 77 kB 3.2 MB/s
Requirement already satisfied: pandas>=0.23.4 in /usr/local/lib/python3.7/dist-packages (from camelot-py) (1.1.5)
Requirement already satisfied: numpy>=1.13.3 in /usr/local/lib/python3.7/dist-packages (from camelot-py) (1.19.5)
Requirement already satisfied: tabulate>=0.8.9 in /usr/local/lib/python3.7/dist-packages (from camelot-py) (0.8.9)
Requirement already satisfied: chardet>=3.0.4 in /usr/local/lib/python3.7/dist-packages (from camelot-py) (3.0.4)
Requirement already satisfied: click>=6.7 in /usr/local/lib/python3.7/dist-packages (from camelot-py) (7.1.2)
Requirement already satisfied: openpyxl>=2.5.8 in /usr/local/lib/python3.7/dist-packages (from camelot-py) (2.5.9)
Collecting pdfminer.six>=20200726
  Downloading pdfminer.six-20201018-py3-none-any.whl (5.6 MB)
     |████████████████████████████████| 5.6 MB 21.1 MB/s
Requirement already satisfied: et-xmlfile in /usr/local/lib/python3.7/dist-packages (from openpyxl>=2.5.8->camelot-py) (1.1.0)
Requirement already satisfied: jdcal in /usr/local/lib/python3.7/dist-packages (from openpyxl>=2.5.8->camelot-py) (1.4.1)
Requirement already satisfied: pytz>=2017.2 in /usr/local/lib/python3.7/dist-packages (from pandas>=0.23.4->camelot-py) (2018.9)
Requirement already satisfied: python-dateutil>=2.7.3 in /usr/local/lib/python3.7/dist-packages (from pandas>=0.23.4->camelot-py) (2.8.1)
Requirement already satisfied: sortedcontainers in /usr/local/lib/python3.7/dist-packages (from pdfminer.six>=20200726->camelot-py) (2.4.0)
Collecting cryptography
  Downloading cryptography-3.4.7-cp36-abi3-manylinux2014_x86_64.whl (3.2 MB)
     |████████████████████████████████| 3.2 MB 44.3 MB/s
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/dist-packages (from python-dateutil>=2.7.3->pandas>=0.23.4->camelot-py) (1.15.0)
Requirement already satisfied: cffi>=1.12 in /usr/local/lib/python3.7/dist-packages (from cryptography->pdfminer.six>=20200726->camelot-py) (1.14.6)
Requirement already satisfied: pycparser in /usr/local/lib/python3.7/dist-packages (from cffi>=1.12->cryptography->pdfminer.six>=20200726->camelot-py) (2.20)
Building wheels for collected packages: PyPDF2
  Building wheel for PyPDF2 (setup.py) ... done
  Created wheel for PyPDF2: filename=PyPDF2-1.26.0-py3-none-any.whl size=61100 sha256=d5777954c8dfcc7ee4b52237b92c63cd90788c41d52abf9af96dbc55f057e5b2
  Stored in directory: /root/.cache/pip/wheels/80/1a/24/648467ade3a77ed20f35cfd2badd32134e96dd25ca811e64b3
Successfully built PyPDF2
Installing collected packages: cryptography, PyPDF2, pdfminer.six, camelot-py
Successfully installed PyPDF2-1.26.0 camelot-py-0.10.1 cryptography-3.4.7 pdfminer.six-20201018
```

```
[ ] !pip install ghostscript
```

```
Collecting ghostscript
  Downloading ghostscript-0.7-py2.py3-none-any.whl (25 kB)
Requirement already satisfied: setuptools>=38.6.0 in /usr/local/lib/python3.7/dist-packages (from ghostscript) (57.2.0)
Installing collected packages: ghostscript
Successfully installed ghostscript-0.7
```

```
[ ] import tabula
```

```
[ ] data = tabula.read_pdf("Test1.pdf",lattice=True ,multiple_tables=True, pages='all')
    data
```

```
[  Name  Test1  Test2  Test3  Test4  Test5
 0  NaN    NaN    NaN    NaN    NaN    NaN
 1  Abc    9.0    8.0    7.0    9.0   10.0
 2  Def    8.0    7.0    8.0    8.0   10.0
 3  Ghi    8.0    8.0    9.0    6.0    5.0
 4  Jkh    4.0    5.0    8.0    7.0    9.0
 5  Lmn    7.0    8.0    4.0    9.0    8.0
 6  Opq    8.0    8.0    7.0    9.0   10.0]
```

```
[ ] data = tabula.read_pdf("No borders.pdf", pages='all')
    tabula.convert_into("No borders.pdf","Noborders.csv", pages="all")
    data
```

```
[  Name  Test1  Test2  Test3  Test4  Test5
 0  NaN    NaN    NaN    NaN    NaN    NaN
 1  Abc    9.0    8.0    7.0    9.0   10.0
 2  Def    8.0    7.0    8.0    8.0   10.0
 3  Ghi    8.0    8.0    9.0    6.0    5.0
 4  Jkh    4.0    5.0    8.0    7.0    9.0
 5  Lmn    7.0    8.0    4.0    9.0    8.0
 6  Opq    8.0    8.0    7.0    9.0   10.0]
```

```
[ ] import camelot
```

```
[ ] data = camelot.read_pdf('Test5.pdf')
    data[0].df
    data[0].to_csv("Test5-1.csv")
```

```
[ ] data[0].df
```

	0	1	2	3
0	1	2	3	4
1	5	6	7	8
2	9	10	11	12
3	13	14	15	16

Here, both the libraries camelot and tabula can read the data from pdf very effectively. But there is one difference we can notice in the output that tabula library considers the first row as header while in camelot it is considered as data itself.

We can use this libraries while dealing with datasets in pdf format. We can easily convert it into csv file.