Date: **18th August, 2021**

Roll No.: **19BCE245**

Name : **Aayush Shah**

Course Code and Name: **2CS501 MACHINE LEARNING**

Practical: **2**

Description :

Read the pdf's tables' data using camelot and tabula libraries and generate different file formats.

```
[ ]  1 pip install camelot-py[cv] tabula-py[cv]
```

```
Collecting camelot-py[cv]
  Downloading camelot_py-0.10.1-py3-none-any.whl (40 kB)
     |████████████████████████████████| 40 kB 11 kB/s
Collecting tabula-py[cv]
  Downloading tabula_py-2.3.0-py3-none-any.whl (12.0 MB)
     |████████████████████████████████| 12.0 MB 96 kB/s
Requirement already satisfied: tabulate>=0.8.9 in /usr/local/lib/python3.7/dist-packages (from camelot-py[cv]) (0.8.9)
Requirement already satisfied: chardet>=3.0.4 in /usr/local/lib/python3.7/dist-packages (from camelot-py[cv]) (3.0.4)
Requirement already satisfied: pandas>=0.23.4 in /usr/local/lib/python3.7/dist-packages (from camelot-py[cv]) (1.1.5)
Collecting PyPDF2>=1.26.0
  Downloading PyPDF2-1.26.0.tar.gz (77 kB)
     |████████████████████████████████| 77 kB 5.7 MB/s
Requirement already satisfied: click>=6.7 in /usr/local/lib/python3.7/dist-packages (from camelot-py[cv]) (7.1.2)
Requirement already satisfied: openpyxl>=2.5.8 in /usr/local/lib/python3.7/dist-packages (from camelot-py[cv]) (2.5.9)
Requirement already satisfied: numpy>=1.13.3 in /usr/local/lib/python3.7/dist-packages (from camelot-py[cv]) (1.19.5)
Collecting pdfminer.six>=20200726
  Downloading pdfminer.six-20201018-py3-none-any.whl (5.6 MB)
     |████████████████████████████████| 5.6 MB 64.4 MB/s
Requirement already satisfied: opencv-python>=3.4.2.17 in /usr/local/lib/python3.7/dist-packages (from camelot-py[cv]) (4.1.2.30)
Collecting ghostscript>=0.7
  Downloading ghostscript-0.7-py2.py3-none-any.whl (25 kB)
Collecting pdftopng>=0.2.3
  Downloading pdftopng-0.2.3-cp37-cp37m-manylinux2010_x86_64.whl (11.7 MB)
     |████████████████████████████████| 11.7 MB 22.3 MB/s
Requirement already satisfied: setuptools>=38.6.0 in /usr/local/lib/python3.7/dist-packages (from ghostscript>=0.7->camelot-py[cv]) (57.4.0)
Requirement already satisfied: et-xmlfile in /usr/local/lib/python3.7/dist-packages (from openpyxl>=2.5.8->camelot-py[cv]) (1.1.0)
Requirement already satisfied: jdcal in /usr/local/lib/python3.7/dist-packages (from openpyxl>=2.5.8->camelot-py[cv]) (1.4.1)
Requirement already satisfied: python-dateutil>=2.7.3 in /usr/local/lib/python3.7/dist-packages (from pandas>=0.23.4->camelot-py[cv]) (2.8.2)
Requirement already satisfied: pytz>=2017.2 in /usr/local/lib/python3.7/dist-packages (from pandas>=0.23.4->camelot-py[cv]) (2018.9)
Collecting cryptography
  Downloading cryptography-3.4.8-cp36-abi3-manylinux_2_24_x86_64.whl (3.0 MB)
     |████████████████████████████████| 3.0 MB 56.3 MB/s
Requirement already satisfied: sortedcontainers in /usr/local/lib/python3.7/dist-packages (from pdfminer.six>=20200726->camelot-py[cv]) (2.4.0)
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[cv]) (1.15.0)
Requirement already satisfied: cffi>=1.12 in /usr/local/lib/python3.7/dist-packages (from cryptography->pdfminer.six>=20200726->camelot-py[cv]) (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[cv]) (2.
WARNING: tabula-py 2.3.0 does not provide the extra 'cv'
Collecting distro
  Downloading distro-1.6.0-py2.py3-none-any.whl (19 kB)
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=61101 sha256=82a007899fb5208ea5939bc90e4b11be969265cd7cd3cb4f70158b6fab3400d4
  Stored in directory: /root/.cache/pip/wheels/80/1a/24/648467ade3a77ed20f35cfd2badd32134e96dd25ca811e64b3
Successfully built PyPDF2
Installing collected packages: cryptography, PyPDF2, pdfminer.six, distro, tabula-py, pdftopng, ghostscript, camelot-py
Successfully installed PyPDF2-1.26.0 camelot-py-0.10.1 cryptography-3.4.8 distro-1.6.0 ghostscript-0.7 pdfminer.six-20201018 pdftopng-0.2.3 tabula-py-2.3.0
```

```
[ ]  1 !sudo apt install ghostscript
```

```
Get:11 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 libcupsfilters1 amd64 1.20.2-0ubuntu3.1 [108 kB]
Fetched 14.1 MB in 1s (22.5 MB/s)
debconf: unable to initialize frontend: Dialog
debconf: (No usable dialog-like program is installed, so the dialog based frontend cannot be used. at /usr/share/perl5/Debconf/FrontEnd/Dialog.pm line 76, <> 
debconf: falling back to frontend: Readline
debconf: unable to initialize frontend: Readline
debconf: (This frontend requires a controlling tty.)
debconf: falling back to frontend: Teletype
dpkg-preconfigure: unable to re-open stdin:
Selecting previously unselected package fonts-droid-fallback.
(Reading database ... 148486 files and directories currently installed.)
Preparing to unpack .../00-fonts-droid-fallback_1%3a6.0.1r16-1.1_all.deb ...
Unpacking fonts-droid-fallback (1:6.0.1r16-1.1) ...
Selecting previously unselected package poppler-data.
Preparing to unpack .../01-poppler-data_0.4.8-2_all.deb ...
Unpacking poppler-data (0.4.8-2) ...
Selecting previously unselected package fonts-noto-mono.
Preparing to unpack .../02-fonts-noto-mono_20171026-2_all.deb ...
Unpacking fonts-noto-mono (20171026-2) ...
Selecting previously unselected package libcupsimage2:amd64.
Preparing to unpack .../03-libcupsimage2_2.2.7-1ubuntu2.8_amd64.deb ...
Unpacking libcupsimage2:amd64 (2.2.7-1ubuntu2.8) ...
Selecting previously unselected package libijs-0.35:amd64.
Preparing to unpack .../04-libijs-0.35_0.35-13_amd64.deb ...
Unpacking libijs-0.35:amd64 (0.35-13) ...
Selecting previously unselected package libjbig2dec0:amd64.
Preparing to unpack .../05-libjbig2dec0_0.13-6_amd64.deb ...
Unpacking libjbig2dec0:amd64 (0.13-6) ...
Selecting previously unselected package libgs9-common.
Preparing to unpack .../06-libgs9-common_9.26~dfsg+0-0ubuntu0.18.04.14_all.deb ...
Unpacking libgs9-common (9.26~dfsg+0-0ubuntu0.18.04.14) ...
Selecting previously unselected package libgs9:amd64.
Preparing to unpack .../07-libgs9_9.26~dfsg+0-0ubuntu0.18.04.14_amd64.deb ...
Unpacking libgs9:amd64 (9.26~dfsg+0-0ubuntu0.18.04.14) ...
Selecting previously unselected package ghostscript.
```

```
Preparing to unpack .../08-ghostscript_9.26~dfsg+0-0ubuntu0.18.04.14_amd64.deb ...
Unpacking ghostscript (9.26~dfsg+0-0ubuntu0.18.04.14) ...
Selecting previously unselected package gsfonts.
Preparing to unpack .../09-gsfonts_1%3a8.11+urwcyr1.0.7~pre44-4.4_all.deb ...
Unpacking gsfonts (1:8.11+urwcyr1.0.7~pre44-4.4) ...
Selecting previously unselected package libcupsfilters1:amd64.
Preparing to unpack .../10-libcupsfilters1_1.20.2-0ubuntu3.1_amd64.deb ...
Unpacking libcupsfilters1:amd64 (1.20.2-0ubuntu3.1) ...
Setting up libgs9-common (9.26~dfsg+0-0ubuntu0.18.04.14) ...
Setting up fonts-droid-fallback (1:6.0.1r16-1.1) ...
Setting up gsfonts (1:8.11+urwcyr1.0.7~pre44-4.4) ...
Setting up poppler-data (0.4.8-2) ...
Setting up fonts-noto-mono (20171026-2) ...
Setting up libcupsfilters1:amd64 (1.20.2-0ubuntu3.1) ...
Setting up libcupsimage2:amd64 (2.2.7-1ubuntu2.8) ...
Setting up libjbig2dec0:amd64 (0.13-6) ...
Setting up libijs-0.35:amd64 (0.35-13) ...
Setting up libgs9:amd64 (9.26~dfsg+0-0ubuntu0.18.04.14) ...
Setting up ghostscript (9.26~dfsg+0-0ubuntu0.18.04.14) ...
Processing triggers for man-db (2.8.3-2ubuntu0.1) ...
Processing triggers for fontconfig (2.12.6-0ubuntu2) ...
Processing triggers for libc-bin (2.27-3ubuntu1.2) ...
/sbin/ldconfig.real: /usr/local/lib/python3.7/dist-packages/ideep4py/lib/libmkldnn.so.0 is not a symbolic link
```

```
[ ]  1 import camelot
     2 tables = camelot.read_pdf('apple_report_q1.pdf')
```

```
[ ]  1 tables
```

```
<TableList n=1>
```

```
[ ]  1 tables[0]
```

```
<Table shape=(28, 3)>
```

```
[ ]  1 tables.export('web_report.html',f='html')
```

```
[ ]  1 tables[0].to_html('web_report_1.html')
     2 tables[0].to_json('json_report.json')
     3 tables[0].to_excel('excel_report.xlsx')
     4 tables[0].to_sqlite('sql_report.sqlite')
```

```
[ ]  1 import tabula
     2 tables_2 = tabula.read_pdf("apple_report_q1.pdf", pages='all')
```

```
[ ]  1 tables_2
```

```
[                                        Apple Inc.
 0   CONDENSED CONSOLIDATED STATEMENTS OF OPERATION...
 1   (In millions, except number of shares which ar...
 2                                  Three months ended
 3                 December 26, 2020 December 28, 2019
 4                                          Net sales
 5                              Products 95.678 79.104
 6                              Services 15,761 12,715
 7                     Total net sales 1,11,439 91,819
 8                      Operating Income 33,534 25,569
 9                    Other income/ expense, net 45 349
 10  Income before provision for income taxes 33,57...
 11            Provision for income taxes 4,824 3,682
 12                         Net income 28,755 22,236
 13                               Earnings per share
 14                                  Basic 1.70 1.26
 15                                Diluted 1.68 1.25
 16        Shares used in computing earnings per share
 17                       Basic 1,69,35,119 1,76,60,160
 18                     Diluted 1,71,13,688 1,78,18,417
 19                  Net sales by reportable segment :
 20                              Americas 46,310 41,367
 21                                Europe 27,306 23,273
 22                           Greater China 21,313 13,578
 23                                   Japan 8,225 7,378
 24                   Rest of Asia pacific 8,225 7,378
 25                    Total net sales 1,11,439 91,819]
```

Explanation :

- **Camelot**

1. Imported the camelot library and further installed ghostscript as an addon for working of camelot library.
2. Read the pdf using camelot's readpdf function.
3. Created different file formats lik html, excel, json, sql using camelot's different functions.

- **Tabula**

1. Read the pdf using tabula's read pdf function and displayed the table as an output.

Difference between tabula and camelot :

Tabula does not produce all of the header text, whereas Camelot correctly places all of the headers in the appropriate cells. Tabula moves part of the data points to the left, whereas Camelot keeps the table in its original position. Tabula will not be able to generate efficient output if the table is rotated, however Camelot would. Camelot produces superior output than Tabula in terms of Column span and Row span. Tabula transfers some headers from the top-right to the left, whereas Camelot does not.

```
# Conclusion :

From this practical, I learned about Camelot and tabula libraries which are used to extract the information from the pdf and can convert it into another forms like sql, html and json without much effort. from this we can analysis easily.
```

Conclusion :

From this practical, I learned about Camelot and tabula libraries which are used to extract the information from the pdf and can convert it into another forms like sql, html and json without much effort. from this we can analysis easily.