POC to extract tabular information from scanned PDF :

We have used TABULA and CAMELOT to extract tabular informations from PDF.Below are the steps involved for both.

The dataset used in both the cases is a Basic Planetary Dataset.

TABULA

1) In the first step we import all the required libraries and packages.

2) We do pip install tabula-py and tabulate-py

3) In the next step,we read the input file using tabula.read_pdf

4) We display the data frame after extracting the tables from PDF in the form of dataframe and display the head of same

5) Using try and error block ,we extract the tables from PDF but not as data frame(It is just another way of displaying the extracted tabular data from pdf)

6)The last step involves,converting and saving the

extracted table into the required output format as “TabulaPlanetaryDataNEWTEST1.csv”

CAMELOT

1) In the first step we import all the required libraries and packages

2) We do pip install camelot-py all and tabulate-py

3) In the next step,we read the input file using camelot.read_pdf

4) We display the data frame after extracting the tables from PDF in the form of dataframe and display the head of same

5) Using try and error block ,we extract the tables from PDF but not as data frame(It is just another way of displaying the extracted tabular data from pdf)

6) The last step involves,converting and saving the

extracted table into the required output format as “CamelotPlanetaryDataNEWTEST1.csv”