Camelot Python

## **Introduction:**

## This is one of the python libraries that helps us extract tables from PDFs and we can convert those tables into CSV,XLSX,TSV and JSON files.

## **Steps :**

1. We need to read the file. While reading the file we assign the reading to an object, so that we can access the table using index , like we can check out the shape and accuracy, and whitespace.
2. If the accuracy is more and whitespace is less that means the table is extracted correctly.
3. We can access the table as a pandas dataframe .
4. We can export the table to a csv file by using export() method.

The above steps are for only one page pdf that contains tables for multiple pages containing multiple tables. We need to mention the pages while reading the file , we can also specify “all” for if the pages and tables were more.

To extract tables from an encrypted pdf we need to specify a password while reading the file.

## **Results of implementation:**

* For structured data I can extract the tables easily and also the accuracy is good.
* For unstructured data we cannot extract any kind of data; it fails to create a dataset.
* For Semi-structured data I am able to create a dataset and the accuracy is really good.
* The results of structured and semi-structured data after applying are converted into a csv file.

## **Some of the advantages of Camelot library are :**

* Camelot provides a lot of flexibility when extracting the tables via the number of parameters.
* Bad tables can be discarded based on metrics like accuracy and whitespace.
* Each table is a pandas dataframe.
* We can export tables to multiple formats.

## **Limitations of Camelot library:**

* Tables aren’t automatically detected; treating the whole page as a single data gives bad output when there are multiple tables on the same page .
* Works only on text based pdfs not on the scanned documents.

I pushed the codes to the github

You can find the link here

“ <https://github.com/mohana-sai/projectcode> ”

For structured data i used 45CANDIDATES pdf and got the accuracy : 100 and whitespace : 0.0

For Unstructured data i used “08122021\_141745” pdf from the drive link. As you know we cannot extract data which is not in tabular format.

For semi structured data I used “08122021\_113706” pdf from the drive link. And got an accuracy of 100 and white space of 10.71.

## **All in one :**

|  | Accuracy | Whitespace | Name of the PDF |
| --- | --- | --- | --- |
| Structured data | 100 | 0.0 | 45CANDIDATES |
| Unstructured data |  |  | 08122021\_141745 |
| Semi-structured | 100 | 10.71 | 08122021\_113706 |