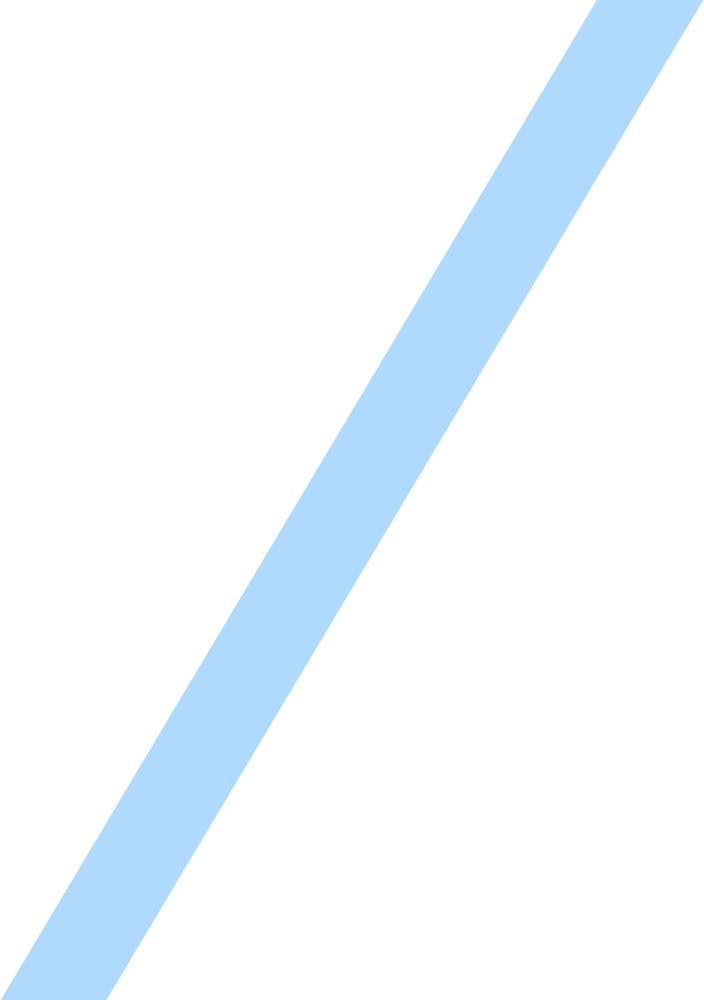
|  |
| --- |
| OCR Table Extraction |

|  |
| --- |
| VizuaMatix  Website: <https://www.vizuamatix.com/> |

|  |  |
| --- | --- |
| Implementation of Data Analyst Team  Author:- Pandula Pallewatta (DA Intern) |  |



Contents

[Introduction to Table Extraction Using OCR 2](#_Toc94618682)

[Importance of Table Recognition …………………………………………………………………………3](#_Toc94618683)

[Industrial use cases ……………………………………………………………………………………………..3](file:///D:\DC%20Universe\Ucsc\Third%20Year\ENH%203201%20Industrial%20Placements\ECL%20Project\ECL%20Demo\Analysis\OCR%20Work\OCR%20Identification.docx#_Toc94618684)

[Method 1: Camelot Library 4](#_Toc94618685)

[Prerequisites 6](#_Toc94618686)

[Method 2: Using Tabula Library 9](#_Toc94618687)

[Prerequisites 9](#_Toc94618688)

[References …………………………………………………………………………………………………………11](#_Toc94618689)

|  |
| --- |
|  |

|  |  |  |
| --- | --- | --- |
| Introduction to Table Extraction Using OCR To get started right away, just tap any placeholder text (such as this) and start typing to replace it with your own.  Some of the sample text in this document indicates the name of the style applied, so that you can easily apply the same formatting again. For example, this is the Content style. | | |
| person at a table writing in a notebook with people around | | |
| Tables are typically used to show facts in a tidy manner. We see them all the time, from organizing our work by organizing data across tables to holding massive corporate assets. Every day, many businesses must deal with millions of tables. We need to use speedier approaches to help with such time-consuming duties of completing everything manually. Let's look at a couple scenarios where extracting tables is critical in **Industrial Business Cases**: |  |  |

|  |
| --- |
| Importance of Table Recognition |

|  |  |
| --- | --- |
| [**” Table Extraction**](https://nanonets.com/table-extraction?&utm_source=nanonets.com%2Fblog%2F&utm_medium=blog&utm_content=Table%20OCR%20To%20Detect%20&%20Extract%20Table%20from%20Image)**is the task of detecting and decomposing table information in a document.”** | person sitting at table working on laptop |
| Industrial use cases **Quality Control:**  In the industrial sector, tables are used to write down daily checklists and notes to track the progress of manufacturing lines. All of this may be easily documented in one place using table extraction.  **Track Of Assets:**  Every manufactured item is assigned a unique number, and tables are used to keep track of the products produced and delivered each day. |

Method 1: Camelot Library

Introduction

Camelot only works with text-based PDFs and not scanned documents.

* **Configurability:** Camelot gives you control over the table extraction process with [tweakable settings](https://camelot-py.readthedocs.io/en/master/user/advanced.html#advanced).
* **Metrics:** You can discard bad tables based on metrics like accuracy and whitespace, without having to manually look at each table.
* **Output:** Each table is extracted into a **pandas DataFrame,** which seamlessly integrates into [ETL and data analysis workflows](https://gist.github.com/vinayak-mehta/e5949f7c2410a0e12f25d3682dc9e873). You can also export tables to multiple formats, which **include CSV, JSON, Excel, HTML, Markdown, and Sqlite.**

Documentation: - [Click Me](https://camelot-py.readthedocs.io/en/master/)

Installation: -

**PIP**

To install Camelot from PyPI using pip, please include the extra cv requirement as shown:

A picture containing text

Description automatically generated

**Conda**

[conda](https://camelot-py.readthedocs.io/en/master/user/install.html#conda) is a package manager and environment management system for the [Anaconda](https://anaconda.org/) distribution. It can be used to install Camelot from the conda-forge channel:

A picture containing chart

Description automatically generated

Graphical user interface, text, application

Description automatically generated**Example Source Code: -**

Text

Description automatically generatedAccuracy of each Table extractions

### Prerequisites

1. Packages to be Installed

install ghostscript (This Package is needed to be installed with additional DLL file)

* First got this site and install Ghostscript DDL (I Windows/Linux).

Link:- [Ghostscript](https://www.ghostscript.com/releases/gsdnld.html)

Graphical user interface, text, application, email

Description automatically generated

Select necessary package for your OS.

* Then install ghostscript according to your python environment.

camelot-py (Use the Pip installation in Page 4 to install Table Extraction model in Python environment)

1. Working with Camelot

* Specify page numbers

By default, Camelot only uses the first page of the PDF to extract tables. To specify multiple pages, you can use the pages keyword argument:



The pages keyword argument accepts pages as comma-separated string of page numbers. You can also specify page ranges — for example, pages=1,4-10,20-30 or pages=1,4-10,20-end.

* Reading encrypted PDFs

To extract tables from encrypted PDF files you must provide a password when calling read\_pdf().

Chart

Description automatically generated with low confidence

Currently Camelot only supports PDFs encrypted with ASCII passwords and algorithm [code 1 or 2](https://github.com/mstamy2/PyPDF2/issues/378). An exception is thrown if the PDF cannot be read. This may be due to no password being provided, an incorrect password, or an unsupported encryption algorithm.

1. Exporting Extracted Tables

export the table as a CSV file using its [to\_csv()](https://camelot-py.readthedocs.io/en/master/api.html" \l "camelot.core.Table.to_csv" \o "camelot.core.Table.to_csv) method. Alternatively you can use [to\_json()](https://camelot-py.readthedocs.io/en/master/api.html#camelot.core.Table.to_json), [to\_excel()](https://camelot-py.readthedocs.io/en/master/api.html#camelot.core.Table.to_excel) [to\_html()](https://camelot-py.readthedocs.io/en/master/api.html#camelot.core.Table.to_html) [to\_markdown()](https://camelot-py.readthedocs.io/en/master/api.html#camelot.core.Table.to_markdown) or [to\_sqlite()](https://camelot-py.readthedocs.io/en/master/api.html#camelot.core.Table.to_sqlite) methods to export the table as JSON, Excel, HTML files or a sqlite database respectively.

Example Code

Graphical user interface, text

Description automatically generated

Graphical user interface

Description automatically generated with medium confidenceAlso, we can find the Structure of the Extracted Table using this Code.

### Method 2: Using Tabula Library

Introduction

Tabula-Py

tabula-py is a simple Python wrapper of [tabula-java](https://github.com/tabulapdf/tabula-java), which can read tables in a PDF. You can read tables from a PDF and convert them into a pandas DataFrame. tabula-py also enables you to convert a PDF file into a CSV, a TSV or a JSON file.

### Prerequisites

1. Required Environments

* Java 8+
* Python 3.6+

Installation

A picture containing text

Description automatically generatedEnsure you have a Java runtime and set the PATH for it.

Example Code:

tabula-py enables you to extract tables from a PDF into a DataFrame, or a JSON. It can also extract tables from a PDF and save the file as a CSV, a TSV, or a JSON.

A picture containing text

Description automatically generated

Converting Specific Pdf’s

Graphical user interface

Description automatically generatedA picture containing graphical user interface

Description automatically generated

## References

1. Analytics Vidhya. 2022. *Camelot - An Amazing Python Library to Extract Tabular Data from PDFs*. [online] Available at: <https://www.analyticsvidhya.com/blog/2020/08/how-to-extract-tabular-data-from-pdf-document-using-camelot-in-python> [Accessed 1 February 2022].
2. Table Extraction with OCR (Demo)· <https://github.com/Pandula1234/Flask-Projects-Dcau/tree/main/Table%20Extraction%20with%20OCR> [Accessed 29 January 2022].
3. F. Du, “Extract tabular data from PDF with Camelot Using Python,” 13-Jan-2019. [Online]. Available: <https://www.youtube.com/watch?v=LoiHI-IB3lY>. [Accessed 31 January 2022].