**Tabula-py**

**PROBLEM STATEMENT**

In this project our goal is to extract entities from pdf documents.

**DESCRIPTION OVERVIEW**

To read a pdf file along with the tables and create a CSV file for the same using the Tabula-py.

**TECHNOLOGY USE**

python

**INSTALLATION**

Installation of this project is pretty easy. Please do follow the following steps to create a virtual environment and then install the necessary packages in the following environment.

**In Pycharm it’s easy**

1. Create a new project.

2. Navigate to the directory of the project

3. Select the option to create a new new virtual environment using conda with python3.7

4. Finally create the project using used resources.

5. After the project has been created, install the necessary packages from requirements.txt file using the command pip install -r requirements.txt

**In Conda also it’s easy**

1. Create a new virtual environment using the command

conda create -n your\_env\_name python=3.7

2. Navigate to the project directory.

3. Install the necessary packages from requirements.txt file using the command

pip install -r requirements.txt

**PROCEDURE** :

#### 

#### Reading a PDF file

#### Reading a Table on a particular page of a PDF File.

#### If there are multiple tables on the same page of a PDF file

* + To read multiple tables we need to add extra parameter
  + multiple\_tables = True -> Read multiple tables as independent tables
  + multiple\_tables = False -> Read multiple tables as single table

#### Reads Multiple Tables as Independent tables

* Reading Multiple Table as a Single Table

#### Convert a PDF file directly to a CSV file

* + We can directly convert a file containing tabular data directly to a CSV file using the convert\_into() method in the tabula library.

###### Converting tables in 1 page to CSV

###### Converting all tables to CSV

**TESTING IN LOCAL**

in this project user needs to provide a pdf documents.

As an output you will get the tabular data which are used in the documents file.

**RESULT**

|  |  |  |
| --- | --- | --- |
| Data | Pdf File | Accuracy |
| Structured Data | * 45CANDIDATES.pdf * FoodCaloriesList.pdf * LOs\_29102020.pdf | * 95% * 98% * 95% |
| Semi-Structured Data | * 08122021\_113706.pdf | * 70% |
| Unstructured Data | * scanned\_example\_1.pdf | * 0% * 0% |

* The result can easily be converted into a CSV file, tsv, etc.
* Tabula has not been bound to particular any language. hence, we can extract data in any language.

**CONCLUSION**

We have successfully created a CSV file for structured and semi-structured data of a pdf.

Tabula is one of the useful packages which not only allows you to scrape tables from PDF files but also convert a PDF file directly into a CSV file.