**Brazil HS/Tarrif Scraper Code Documentation**

This Python script extracts tables from the PDF file containing Brazilian Tarrif using the **tabula-py** library, combines all the extracted tables into a single DataFrame, and saves the data to a CSV file. The script is designed to handle PDFs with multiple tables, scraping them in sequence.

Requirements:

* Python 3.x
* Libraries: **tabula-py, pandas, requests**

Installation:

* Install the required libraries using pip, run the code below in pip package-managed python env:

***pip install tabula-py pandas requests***

The Code Explanation:

1. Import Libraries:

* tabula: Used to extract tables from the PDF.
* pandas: Used to handle and combine the extracted tables into a DataFrame.
* requests: Used to download the PDF file from the provided URL.

2. PDF URL:

* The variable pdf\_url stores the URL of the PDF file to be processed.

3. Download the PDF:

* The script uses the requests library to download the PDF file from the provided URL.
* The downloaded PDF is saved locally as tipi.pdf.

4. Extract Tables:

* The tabula.read\_pdf function is used to extract all tables from the PDF.
* Parameters:
* pdf\_path: The path to the downloaded PDF file.
* pages="all": Processes all pages in the PDF.
* multiple\_tables=True: Ensures that all tables on each page are detected.
* The extracted tables are stored in a list called tables.

5. Combine Tables:

* The pd.concat function is used to combine all the extracted tables into a single DataFrame.
* The ignore\_index=True parameter ensures that the index is reset in the combined DataFrame.

6. Save to CSV:

* The combined DataFrame is saved to a CSV file named **brazil\_tarrif.csv** using the to\_csv method.
* The index=False parameter ensures that the DataFrame index is not included in the CSV file as this would introduce unnecessary field.

7. Output:

* The script prints a confirmation message once the tables are extracted and saved to the CSV file.

Usage:

1. Save the script to a Python file, e.g., **brazil\_tarrif\_scraper.py**.
2. Run the script using the command:

python **brazil\_tarrif\_scraper.py**

1. The extracted tables will be saved in a file named extracted\_tables.csv in the same directory as the script.

Notes:

* If the PDF contains complex formatting, additional parameters such as **lattice=True** or **stream=True** may need to be passed to the tabula.read\_pdf function for better table detection and encrypted data handling. Was not necessary here.
* **The extracted csv file containing the data requires further data restructuring, reformatting and cleaning, which I did on MS Excel.**