Introduction

Please read the file carefully to understand working of the pdf_to_excel tool/scraper.
This tool works with OCR (Optical Character Recognition) technology where scanned images or PDF is unreadable to system and can't copy text. This technology converts scanned document into readable or text document. Same tech is used in our code because all the pdf's we are working are scanned documents.
Our tool goes to the folder named as all_pdf and convert them as OCR_document and put it in the folder OCR_Doucments. Then it takes the pdfs in the OCR_Document folder and read them and extract them one by one.
We were provided two types of scanned document. One was little bit easier and other were harder to get extracted. Our tool uses regular expression library to extract the text.
This library is used to extract the required text from the given text, it uses expressions for example getting CFN: was easy we just gave a expression to just extract the whole text just after CFN: till line breaks, and it appended all the CFN number with book and page in our CFN Column. This library is useful but can be harder too because we tried hard to get the Owner Name from the text and from one type of pdf it was easier because we gave expression to get the result after Name of Violator: but on other type of file, it is very hard because there were other names too and we can't figure out it. So, you will find some name spaces empty because of this issue.
let me describe you all the columns one by one.

CFN:

this column named as CFN is extracting good from all the files. No issue with this col.

Parcel Id:

This field is also extracting the data from all the files with out any issue it was also easier.

Property Address:

This field took our time but we figure out how to extract it and this one is also working good and extracting good with both types of files with the help of tag Property Address: OR Subject Property:, the address in front of this is the Property Address.

Property City/State:

We figured out that all the damages or violations are done in the city Miami and this is the reason Miami gov is taking actions so Property is from Miami city and just wrote Miami in this col and same goes for the column Property State and Miami is the city of Florida state and FL is written in Property State.

Property Zip code:

This field was mandatory but we searched for all the documents and fount nothing against this column and this column is remained empty. We are sorry for this, if you can mention it on the file we can definitely integrate it.

Mailing Address:

This column took the most of our time in coding and this is because there were two same addresses in the file and figuring out how to handle was tough, but we did it and still there are one or rows in which it didn't appended correctly, that can be ignored.

Mailing City/Mailing State/Mailing Zip code:

These are working well and getting the data as needed.

Company Name/Owner First Name/Owner Last Name:

The name is taken and decided if there is LLC or INC or LTD or Corporation in the name then it goes in Company name, if these are not then split into two and goes in First name and last name.

These columns have still some missing rows and that is because we tried hard to get the Owner Name from the text and from one type of pdf it was easier because we gave expression to get the result after Name of Violator: but on other type of file, it is very hard because there were other names too and we can't figure out it. So, you will find some name spaces empty because of this issue.

Deceased/Notes:

These columns are empty because we didn't find anything regarding this in any pdf and it was also cleared in the meeting dated 04-Oct-2024

Property Tag:

This we found that there are few files in which Unsafe Structure is mentioned and so if this same word found in any document it will append it under that.

This project taught us a lot of things from the prospective of coding. We would be glad to work with you again.
please contact developer Safeer Abbas WhatsApp for any other information +923312378492. Thanks.