Approach

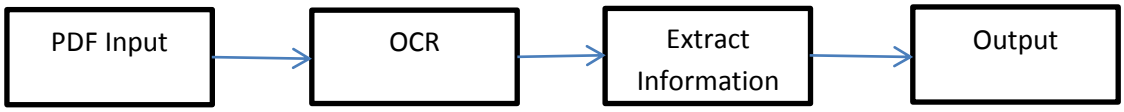

Fig-Overview to extract information

Given input data in pdf format and first needs convert into text format to process further

OCR is need to be done to get input pdf into text format

- Python supported libraries (Tesseract) don't provide that much good accuracy over paid services. Data loss may happen
- For OCR paid services/tools can be used (ABBYY FineReader, AWS) It provides bounding box values and text within that.

As expected output is specific to particular document type and after analysing document I think of 4 approaches to extract information.

1) OCR provides data in .json format with Bounding Box and text within it.

For given pdf input required output can be obtained by observing bounding box values for top, bottom, left, right co-ordinates

Top left corner (co-ordinates with left and top value is zero),

Top right corner (co-ordinates with right and top is zero),

Middle of page (middle of topmost +bottom point)

Using this information Text tag provide text data within that values

Paid tools also provides whole single bounding box data which can be further tuned to get expected output

2) Python Regex library (re module)can be used to get expected information as output. OCR gives .txt file which can be used for this approach

Below is detailed way to get information using regex

a) Top 3 lines from text belongs to expected block we can use regex to get anything till some line numbers.
   Dictionary for customer's name/address data can be used to find names in input data.
   Update dictionaries if new names find in scheduled manner.
b) First Top right block have policy number for this regex can be used to get keyword policy no followed by no with particular pattern and within that span other expected data
c) Second Top right block have .com for this regex can be used to and data after .com pattern
d) Box contains insurance and summary as keyword regex can used to extract data having such keywords
e) Middle data have date as parameter along with tags like premium changes. This tag used to identify required text data

3) Covert pdf to image and perform image processing to get highlighted box as separates images and OCR for those images can be done to get desired output.

4) Gate tool can be used to write rules for expected data.