***Task 2:Computer Vision(OCR/ICR)****:*

*Go to* [*https://content.sbigeneral.in/uploads/e1904ff17d084f6582d5cc43bb6e059e.pdf*](https://content.sbigeneral.in/uploads/e1904ff17d084f6582d5cc43bb6e059e.pdf)*, print the form, fill it (with fake details), scan and perform OCR.*

***The process of extracting information from a digital copy of invoice can be a tricky task. There are various tools that are available in the market that can be used to perform this task. However, there are many factors due to which most of us want to solve this problem using Open Source Libraries.***

**OCR :**Optical Character Recognition.

OCR systems transform a two-dimensional image of text, that could contain machine printed or handwritten text from its image representation into machine-readable text. OCR as a process generally consists of several sub-processes to perform as accurately as possible.

Firstly, we need to convert the pages of the PDF to images and then, use OCR (Optical Character Recognition) to read the content from the image and store it in a text file.

***There are two parts to the program.***

**Part-1: D**eals with converting the PDF into image files. Each page of the PDF is stored as an image file.

**Part-2:** D**eals** with recognizing text from the image files and storing it into a text file. Here, we process the images and convert it into text. Once we have the text as a string variable, we can do any processing on the text.

***Required Installations:***

* pip install pdf2image
* pip install pytesseract
* pip install opencv-python
* pip install Pillow

***The libraries that I used for developing this solution*** *:*

* **pdf2image** is a python library which converts PDF to a sequence of PIL Image objects using pdftoppm library.   
  Note: pdf2image uses **Poppler** which is a PDF rendering library based on the xpdf-3.0 code base and will not work without it.
* For this OCR project, I am using the [Python-Tesseract](https://pypi.org/project/pytesseract/), or simply ***PyTesseract*,** library which is a wrapper for [Google's Tesseract-OCR Engine](https://github.com/tesseract-ocr/tesseract). I chose this because it is completely open-source and being developed and maintained by the giant that is Google.
* Besides those, I am using the[***Pillow***](https://github.com/python-pillow/Pillow) library which is a fork of the Python Imaging Library (PIL) to handle the opening and manipulation of images in many formats in Python.
* **OpenCV** for Image pre-processing.

***Conclusion:***

* The pages of the PDF were converted to images. Then the images were read, and the content was written into a text file.
* Avoiding text-based conversion because of encoding scheme resulting in loss of data.
* Handwritten content in PDF can be recognized using OCR.
* Recognizing only particular pages of the PDF is also possible.
* Using OCR cannot guarantee 100% accuracy.
* Handwritten PDFs are still recognized, but the accuracy depends on various factors like handwriting, page color, etc.