Problem:extract images from PDF

How to tackle the problem

Simple analogy of the solution to the problem

We will extract the images from PDF files and save them using PyMuPDF library.

Step to step way of tackling a problem.

Import necessary libraries

- •Specify the path of the file from which you want to extract images and open it
- •Iterate through all the pages of PDF and get all images objects present on every page
- •Use getImageList() method to get all image objects as a list of tuples
- •To get the image in bytes and along with the additional information about the image, use extractImage()

NB : PyMuPDF is used to access PDF files.

`page.getImageList()` return a list of all images present on the single page.

`my_pdf_file.extractImage(xref_value)` method returns all the information about the image, including its byte code and image extension.

`io.BytesIO(image_bytes)` changes the image bytes-like object to proper bytes object. method opens the image byte object.

Code:

```
import fitz # PyMuPDF
import io
from PIL import Image

#filename
filename = "Product-Management-Interview-Questions-Book.pdf"

# open file
with fitz.open(filename) as my_pdf_file:

    #loop through every page
    for page_number in range(1, len(my_pdf_file)+1):

        # acess individual page
        page = my_pdf_file[page_number-1]

        # accesses all images of the page
        images = page.getImageList()

        # check if images are there
        if images:
            print(f"There are {len(images)} image/s on page number {page_number}[+]")
        else:
            print(f"There are No image/s on page number {page_number}[!]")

        # loop through all images present in the page
        for image_number, image in enumerate(page.getImageList(), start=1):
```

```
            #access image xerf
            xref_value = image[0]

            #extract image information
            base_image = my_pdf_file.extractImage(xref_value)

            # access the image itself
            image_bytes = base_image["image"]

            #get image extension
            ext = base_image["ext"]

            #load image
            image = Image.open(io.BytesIO(image_bytes))

            #save image locally

image.save(open(f"Page{page_number}Image{image_number}.
{ext}", "wb"))
```

By karimi christine ML expert