P.2. Text Extraction and Processing for Images, PDFs, and Speech Data

In text extraction applications, the core is the **Optical Character Recognition (OCR)** technology.

- Its primary function is to extract texts from images.
- Using advanced AI algorithms and **machine learning**, the OCR can identify and convert **image texts** into **audio files**, for easy listening.

There are some powerful text extraction software (having accuracy of 98+%), which are not freely/conveniently available. Thus we will implement two text extraction algorithms, one for image data and the other for speech data.

Project Objectives: The project will develop two separate Python programs: `pdfim2text` and `speech2text`.

1. **PDF-Image to Text** (pdfim2text)
   - **Input**: an image or a pdf file
     - A PDF may include images.
     - A PDF may be generated by scanning, in which each page is an image.
   - **Core Task**: Extract all texts; play texts from an image
2. **Speech to Text+Speech** (speech2text)
   - **Input**: speech from **microphone** or a wave file
   - **Core Task**: Extract texts; play the extracted texts.

An Example

pdfim2text

```
#!/usr/bin/python

import pytesseract
from pdf2image import convert_from_path
from PIL import Image
from gtts import gTTS
from playsound import playsound
import os, pathlib, glob

def takeInput():
    pmode = 0;
    IN = input("Enter a pdf or an image: ")
    if os.path.isfile(IN):
        path_stem = pathlib.Path(IN).stem
        path_ext  = pathlib.Path(IN).suffix
        if path_ext.lower() == '.pdf': pmode=1
    else:
        exit()
    return IN, path_stem, pmode

def pdf2txt(IN):
    # you have to complete the function appropriately
    return 'For pdf2txt, you may save the text here without return.'

def im2txt(IN):
    # you have to complete the function appropriately
    return 'For im2txt, try to return the text to play'

if __name__ == '__main__':
    IN, path_stem, pmode = takeInput()   #pmode=0:image; pmode=1:pdf
    if pmode:
        txt = pdf2txt(IN)
    else:
        txt = im2txt(IN)

    audio = gTTS(text=txt, lang="en", slow=False);
    WAV = '0000_tmp.wav'; audio.save(WAV);
    playsound(WAV); os.remove(WAV)
```

What to Do

First download Image-Speech-Text-Processing.PY.tar.
Untar it to see the file `pdfim2text` and some example codes in the directory `example-code`.

1. Complete `pdfim2text` appropriately.
   - You may find clues from `example-code/pdf2txt.py`
2. Implement `speech2text` from the scratch.
   - You may get hints from `speech_mic2wave.py` and `image2text.py` in the directory `example-code`.

Try to put all functions into a single file for each command, which enhances portability of the commands.

Report

- Work in a directory, of which the name begins with your last name.
- Use the three-page project document as an example for `pdfim2text`.
- zip or tar your work directory and submit via email.
- Write a report to explain what you have done, including images and wave files; upload it to Canvas.

PDF-Image to Texts

As a part of the project,
you will develop a Python program
that can extract texts from
PDF files and images:
and generally,
from PDF files including images.

An example PDF is
the one you are reading now.

(This portion is an image, by text2image.py.)