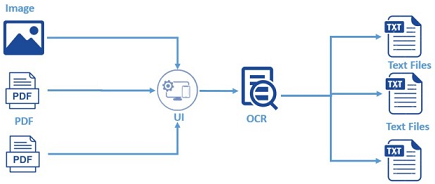
Smart OCR For Document Digitization-

In this project we aim to make and OCR scanner so that it takes images and pdf and then extracts the text and stores it in txt file.

The project aims at creating an application form where the user can upload a pdf document/Image containing text, the document is analysed by an Optical character recognition (OCR) to extract text from it. The extracted text is again saved in a text document in the local drive.



**Project Flow:**

1. Upload a pdf document
2. Convert PDF document to image
3. Extract the text from the image
4. Store the extracted text in the text document

**Learning Objective:**

* Use Python, tesseract modules  for text extraction
* Flask Web Framework

**Prerequisite:**

* **Python IDE**- For programming
* **pytesseract** -OCR package in python
* **pdf2image**- Converting PDF to Image
* **tesseract-ocr execution file** -Backend used for pytesseract
* **poppler**-Supporting file for pdf2image package
* **Flask**-Build web application

Commands –

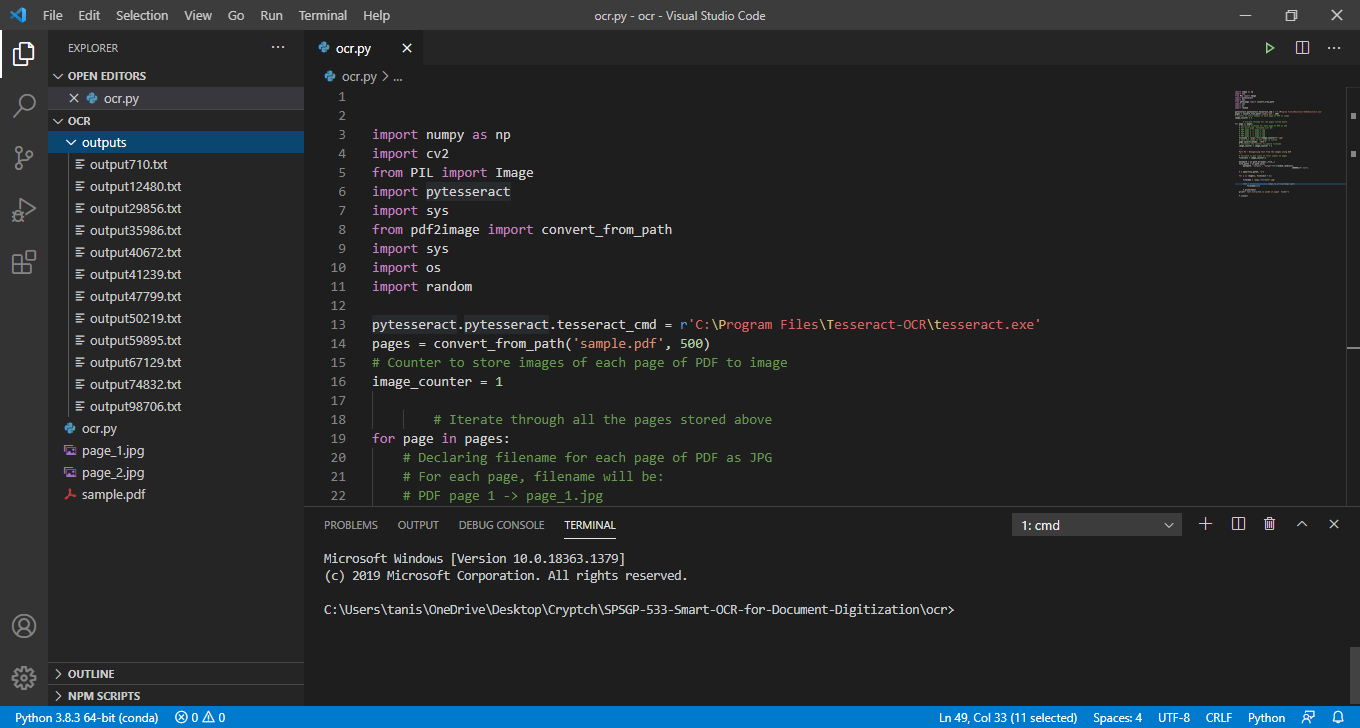
pip install opencv

pip install pdf2image

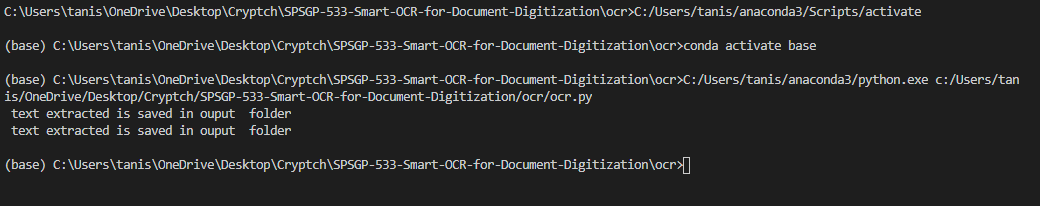
for further installation <https://www.youtube.com/watch?v=Rb93uLXiTwA>

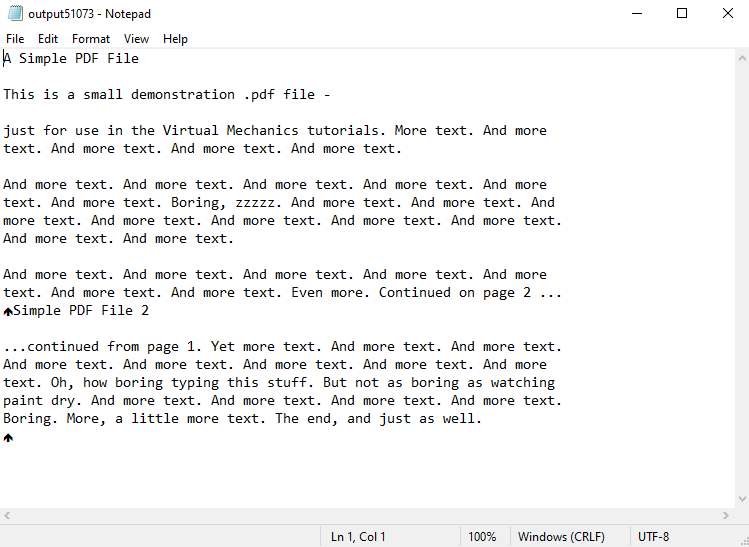
Now we will write the code it is available in Github-

<https://github.com/smartinternz02/SPSGP-533-Smart-OCR-for-Document-Digitization/tree/master/ocr>



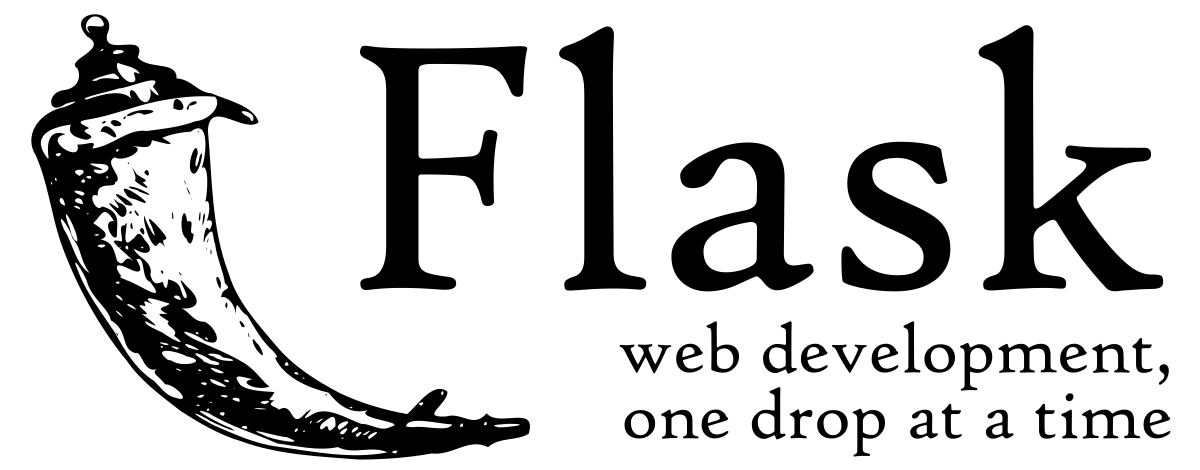
When we run the code, we get





# Building flask web application

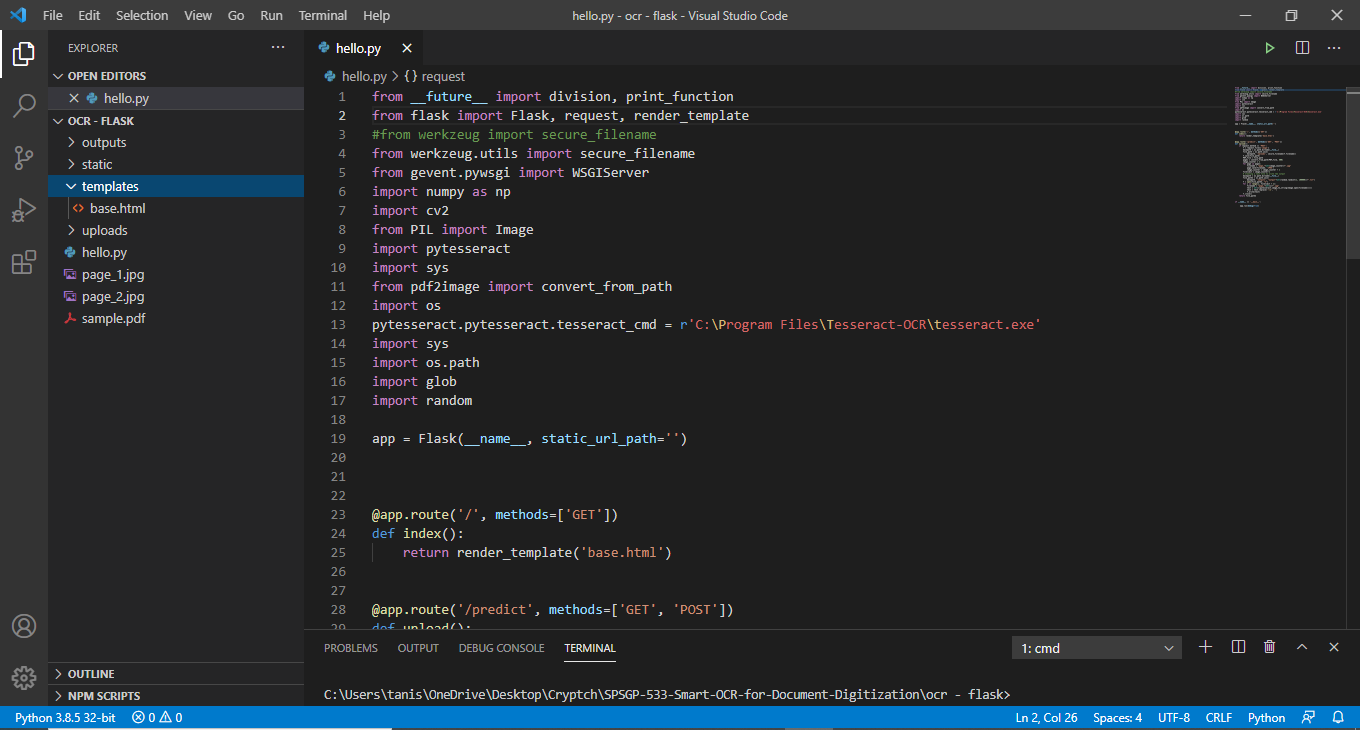
Flask is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions.



### Build HTML Code

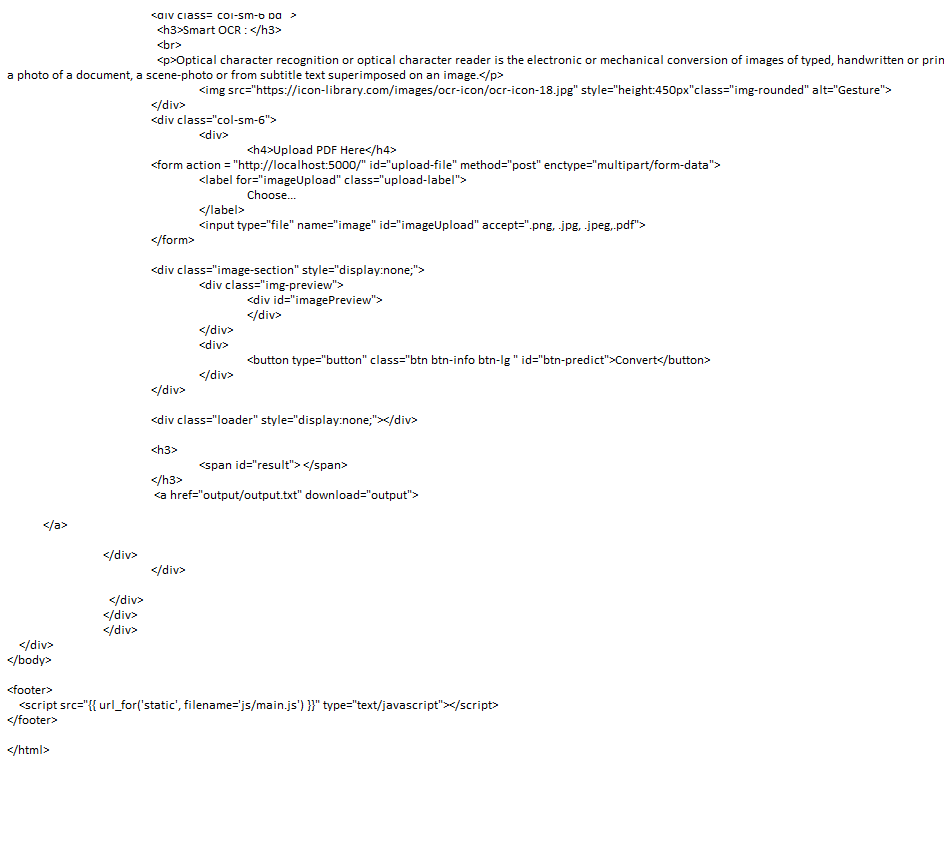
The basic HTML page for our Project. Here we are creating two buttons one used to browse PDF documents from the local drive and another button is used to send this pdf for text analyses.

**Project Structure:**

****

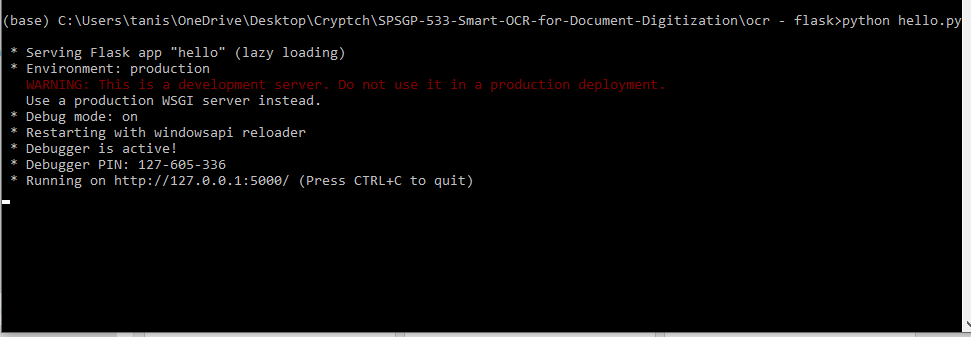
It’s code is also available in <https://github.com/smartinternz02/SPSGP-533-Smart-OCR-for-Document-Digitization/tree/master/ocr%20-%20flask>

The basic HTML page for our Project. Here we are creating two buttons one used to browse PDF documents from the local drive and another button is used to send this pdf for text analyses.



**Run the App**

* Open anaconda prompt/command from the start menu
* Navigate to the folder where your app.py resides
* Now type “python hello.py” command
* It will show the local host where your app is running.
* Navigate to the localhost where you can view your web page.



Here is the final output

