Smart OCR for Document Digitization

Introduction:

Overview:

1. We provide input through post method in web page.
2. Preprocessing of pdf occurs,within pytesseract. Text localistion,detectection,segmentation,recognition.
3. It returns content in text format along with the confidence, bouding boxes for what it has been recognized.
4. Take that output and saving it in a file.

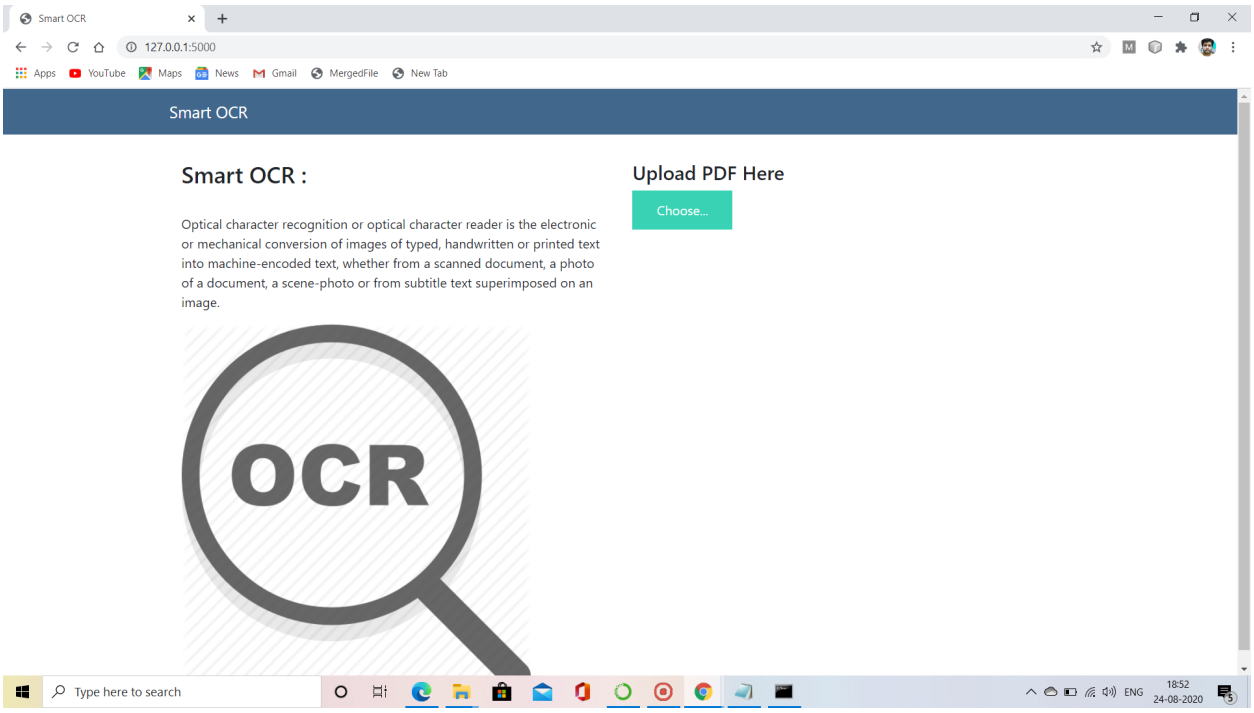
The whole process is dumped in Flask.

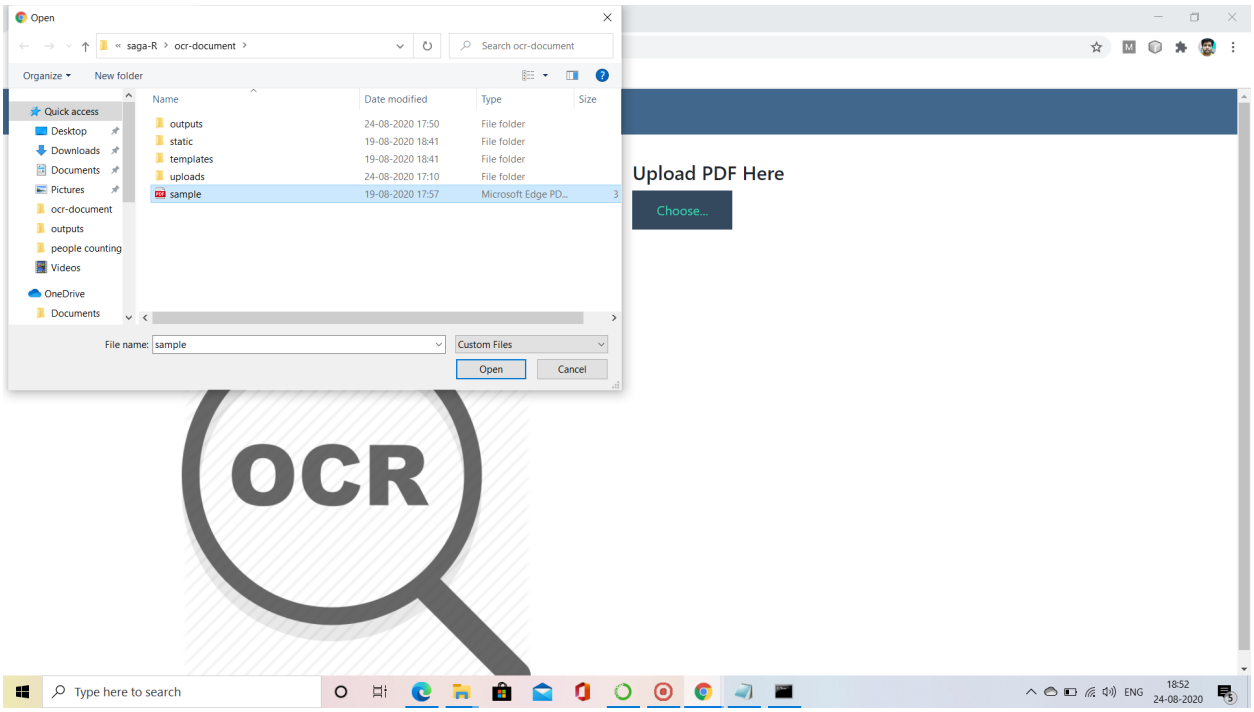
Purpose:

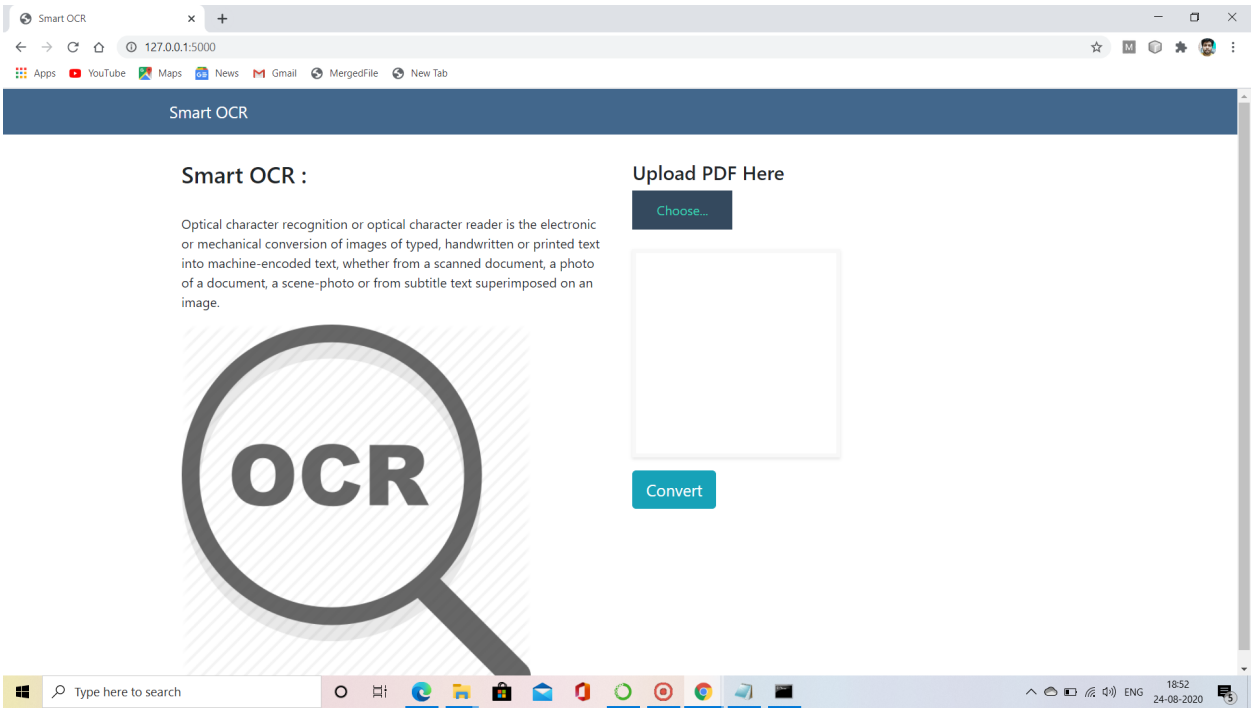
With the advent of OCR techniques, much time has been saved by automatically extracting the text out of a digital image of any invoice or a document.

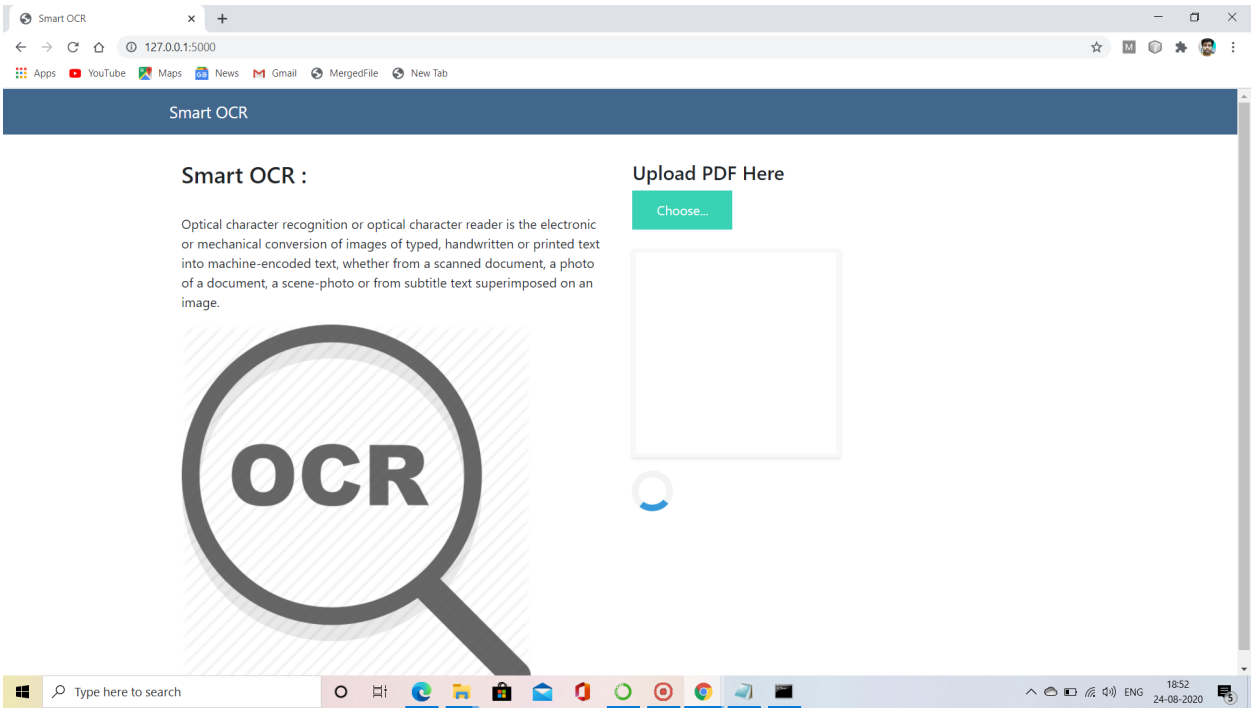
Result:

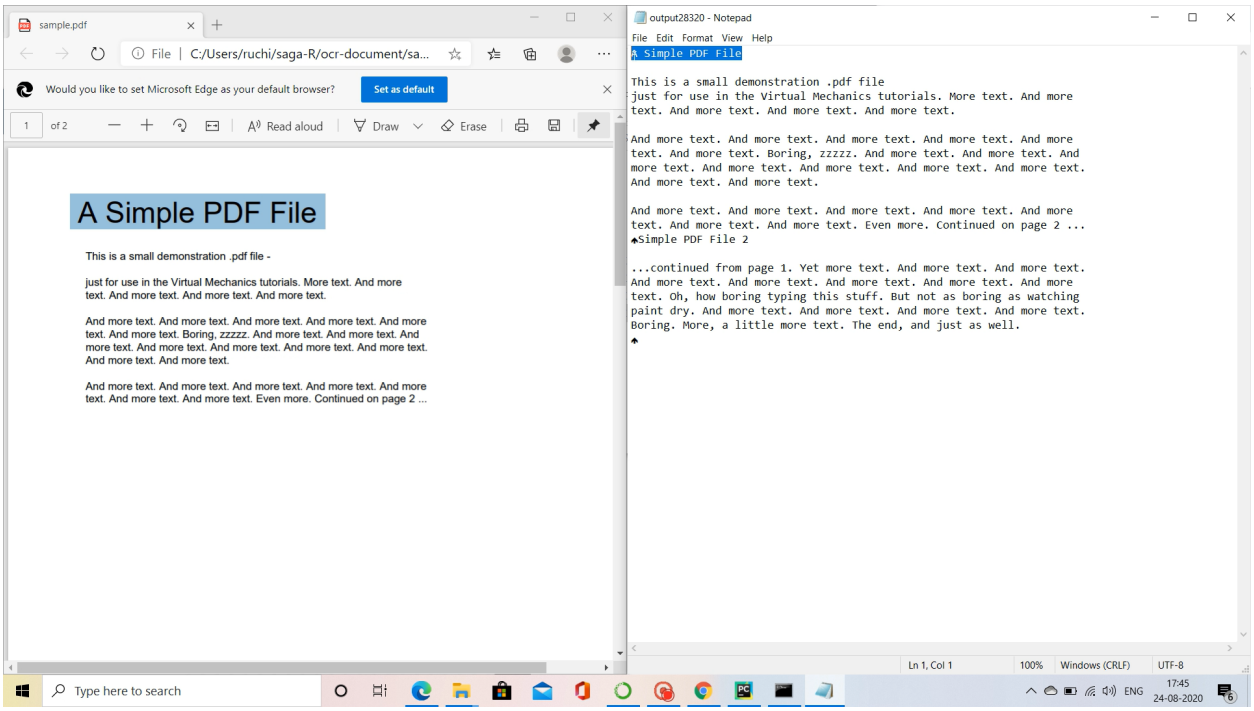
Screenshots:

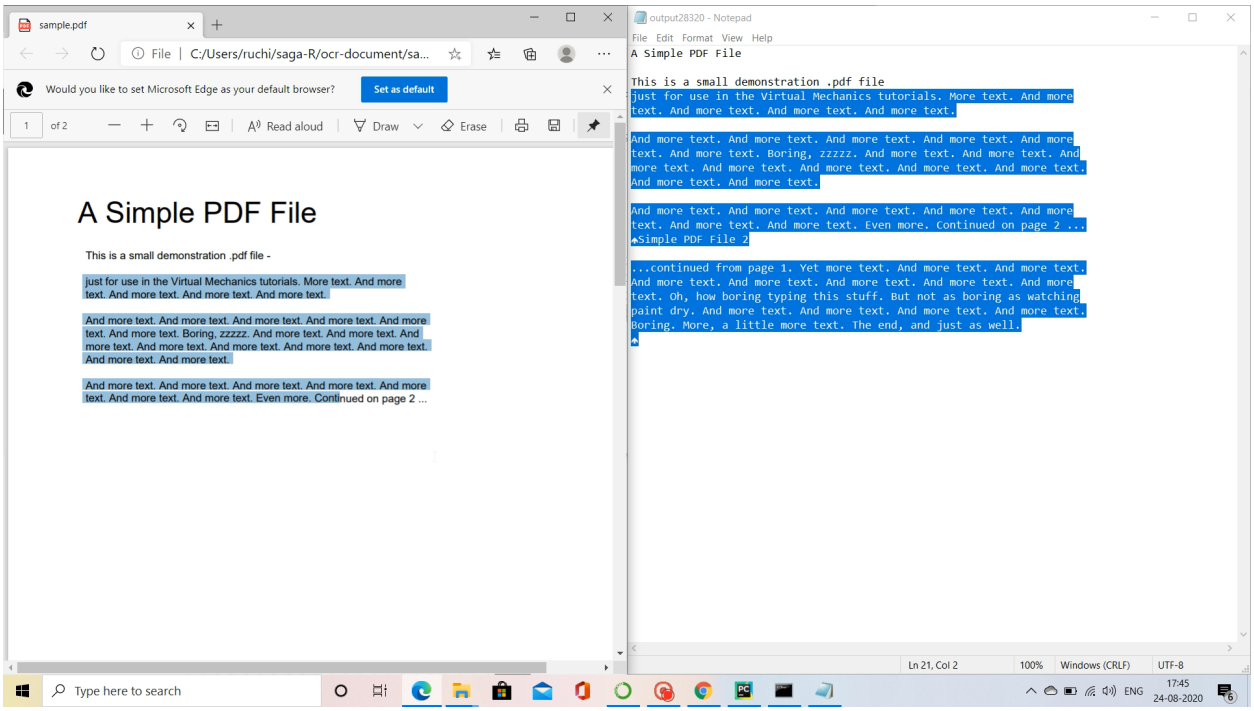

Smart OCR
127.0.0.1:5000
Apps
YouTube
Maps
News
Gmail
MergedFile
New Tab
Smart OCR
Smart OCR :
Optical character recognition or optical character reader is the electronic or mechanical conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo or from subtitle text superimposed on an image.
Upload PDF Here
Choose...
OCR
Type here to search
ENG
18:52
24-08-2020


Open
saga-R > ocr-document
Search ocr-document
Organize
New folder
Quick access
Desktop
Downloads
Documents
Pictures
ocr-document
outputs
people counting
Videos
OneDrive
Documents
Name
Date modified
Type
Size
outputs
24-08-2020 17:50
File folder
static
19-08-2020 18:41
File folder
templates
19-08-2020 18:41
File folder
uploads
24-08-2020 17:10
File folder
sample
19-08-2020 17:57
Microsoft Edge PD...
File name: sample
Custom Files
Open
Cancel
Upload PDF Here
Choose...
OCR
Type here to search
ENG
18:52
24-08-2020


Smart OCR
127.0.0.1:5000
Apps YouTube Maps News Gmail MergedFile New Tab
Smart OCR
Smart OCR :
Optical character recognition or optical character reader is the electronic or mechanical conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo or from subtitle text superimposed on an image.
OCR
Upload PDF Here
Choose...
Convert
Type here to search
ENG
18:52
24-08-2020


Smart OCR
127.0.0.1:5000
Apps YouTube Maps News Gmail MergedFile New Tab
Smart OCR
Smart OCR :
Optical character recognition or optical character reader is the electronic or mechanical conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo or from subtitle text superimposed on an image.
OCR
Upload PDF Here
Choose...
Type here to search
ENG
18:52
24-08-2020


sample.pdf
File | C:/Users/ruchi/saga-R/ocr-document/sa...
Would you like to set Microsoft Edge as your default browser?
Set as default
1 of 2
Read aloud
Draw
Erase
A Simple PDF File
This is a small demonstration .pdf file -
just for use in the Virtual Mechanics tutorials. More text. And more text. And more text. And more text.
And more text. And more text. And more text. And more text. And more text. Boring, zzzzz. And more text. And more text. And more text. And more text. And more text. And more text. And more text. And more text.
And more text. And more text. And more text. And more text. And more text. Even more. Continued on page 2 ...
output28320 - Notepad
File Edit Format View Help
A Simple PDF File
This is a small demonstration .pdf file
just for use in the Virtual Mechanics tutorials. More text. And more text. And more text. And more text. And more text.
And more text. And more text. And more text. And more text. And more text. And more text. Boring, zzzzz. And more text. And more text. And more text. And more text. And more text. And more text. And more text. And more text.
And more text. And more text. And more text. And more text. And more text. Even more. Continued on page 2 ...
Simple PDF File 2
...continued from page 1. Yet more text. And more text. And more text. And more text. And more text. And more text. And more text. And more text. And more text. Oh, how boring typing this stuff. But not as boring as watching paint dry. And more text. And more text. And more text. And more text. Boring. More, a little more text. The end, and just as well.
Ln 1, Col 1
100%
Windows (CRLF)
UTF-8
Type here to search
ENG
17:45
24-08-2020


sample.pdf
File | C:/Users/ruchi/saga-R/ocr-document/sa...
Would you like to set Microsoft Edge as your default browser?
Set as default
1 of 2
Read aloud
Draw
Erase
A Simple PDF File
This is a small demonstration .pdf file -
just for use in the Virtual Mechanics tutorials. More text. And more text. And more text. And more text.
And more text. And more text. And more text. And more text. And more text. Boring, zzzzz. And more text. And more text. And more text. And more text. And more text. And more text. And more text. And more text.
And more text. And more text. And more text. And more text. And more text. Even more. Continued on page 2 ...
output28320 - Notepad
File Edit Format View Help
A Simple PDF File
This is a small demonstration .pdf file
just for use in the Virtual Mechanics tutorials. More text. And more text. And more text. And more text. And more text.
And more text. And more text. And more text. And more text. And more text. And more text. Boring, zzzzz. And more text. And more text. And more text. And more text. And more text. And more text. And more text. And more text.
And more text. And more text. And more text. And more text. And more text. Even more. Continued on page 2 ...
Simple PDF File 2
...continued from page 1. Yet more text. And more text. And more text. And more text. And more text. And more text. And more text. And more text. And more text. Oh, how boring typing this stuff. But not as boring as watching paint dry. And more text. And more text. And more text. And more text. Boring. More, a little more text. The end, and just as well.
Ln 21, Col 2
100%
Windows (CRLF)
UTF-8
Type here to search
ENG
17:45
24-08-2020

Applications:

1. Upload a pdf document, the document is analyzed by Optical character recognition package to extract text from it. The extracted text is again saved in a text document in the local drive.

2. Can be deployed in self driving cars,to read what is mentioned on road side saftey board to get precaution while driving .

Conclusion:

Run python file on web,since it is deployed in flask,in the address you shown in command prompt.Upload a document file,ou will get a text file as output after processing through tesseract.