# Text Extraction
# Project Report

INFOBHAN

SYSTEMS & SERVICES

**Ahmed Mustafa**

**Department Of Software Development,**

**Infobhan Systems & Services, Doha**

## Abstract:

This project aims towards creating an optical character recognition system to be used within the organization. The primary objective is to create a backend system that identifies text within pdf and image files.

## Requirement Specification:

### Hardware Requirements:

| Hardware Component | Requirement |
|---|---|
| RAM | 4 GB |
| Storage/Hard Disk/SSD | 100 GB |
| Processor | Intel core i5 |

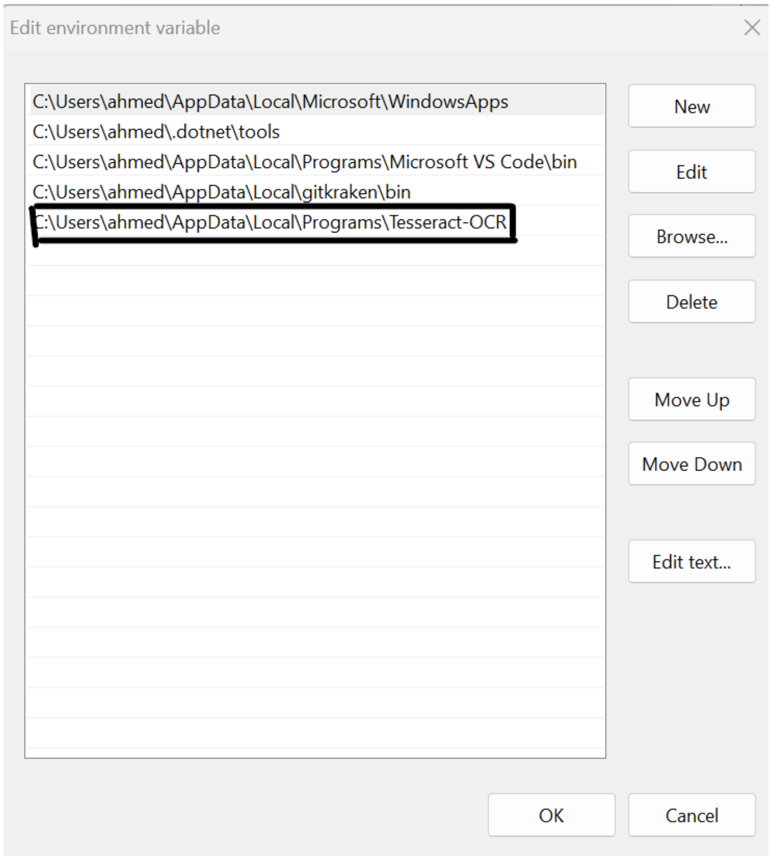
### Software Requirements (libraries & dependencies):

- Python IDLE (development environment) 3.9 or above.
- PyPDF2 library (Python module that allows you to work with PDF file).
- tesseract-ocr engine version 5.3.1.2 (open-source OCR engine developed by Google. It is designed to recognize text in images).
- pytesseract library (allows you to easily integrate Tesseract OCR functionality into your Python applications).
- Pillow library (fast access to data stored in a few basic pixel formats)

## Approach:

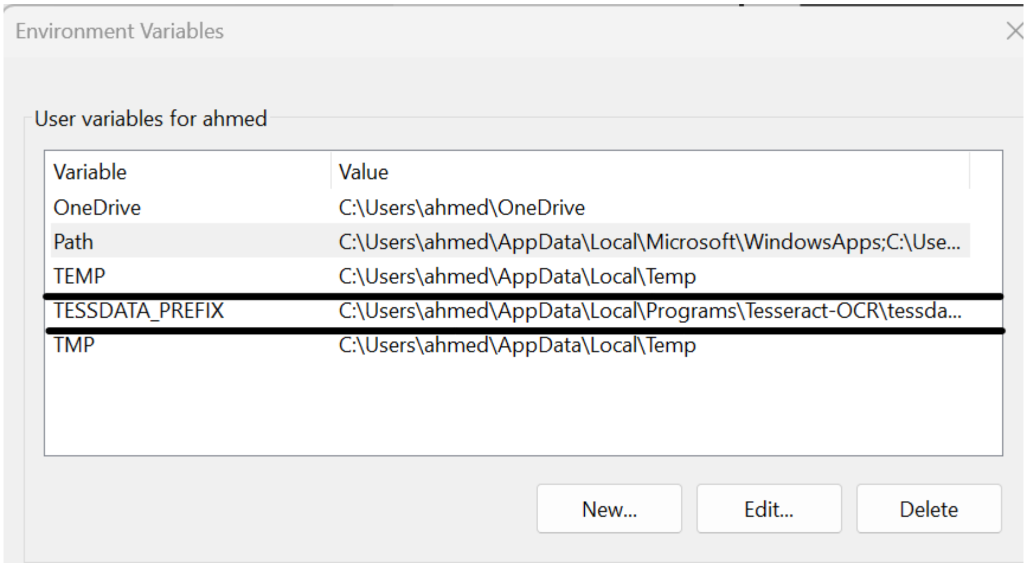
### Installation:

Assuming we have a python IDLE present, we begin with the installation of the tesseract-ocr engine by downloading the executable file for windows. After installation process is complete, we are required to add the path of the tesseract engine to the environment variables as shown in image below.



```
C:\Users\ahmed\AppData\Local\Microsoft\WindowsApps
C:\Users\ahmed\.dotnet\tools
C:\Users\ahmed\AppData\Local\Programs\Microsoft VS Code\bin
C:\Users\ahmed\AppData\Local\gitkraken\bin
C:\Users\ahmed\AppData\Local\Programs\Tesseract-OCR
```

Additionally, we also add a new variable in the user variable with title of TESSDATA_PREFIX and add path to the folder titled tessdata within the Tesseract-OCR folder as its value.



User variables for ahmed

| Variable | Value |
|---|---|
| OneDrive | C:\Users\ahmed\OneDrive |
| Path | C:\Users\ahmed\AppData\Local\Microsoft\WindowsApps;C:\Use... |
| TEMP | C:\Users\ahmed\AppData\Local\Temp |
| TESSDATA_PREFIX | C:\Users\ahmed\AppData\Local\Programs\Tesseract-OCR\tessda... |
| TMP | C:\Users\ahmed\AppData\Local\Temp |

We also install the python libraries required using package installer python (pip).

**Detecting text in PDF file:**

- Firstly, we import the required library which is PyPDF2.
- Create a pdf file object.
- Create a pdf reader object.
- Create a page object.
- Specify page number (first page starting with 0) and extract text.
- Print or perform other operations with the extracted text.

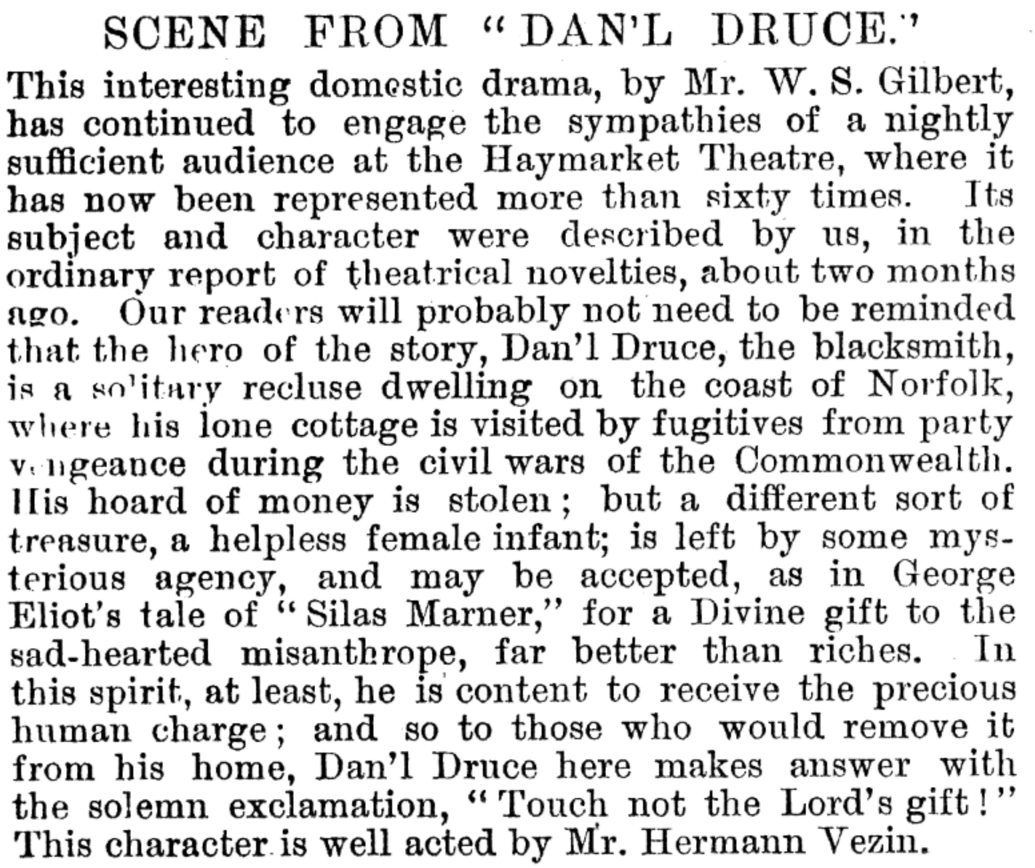
**Detecting text in an image:**

- Import Required libraries which are pytesseract and pillow.
- Create an image object using pillow library.
- Pass the object to the image_to_string function in pytesseract.
- Print or perform other operations with extracted text.

## Results:

### Detecting text in image:

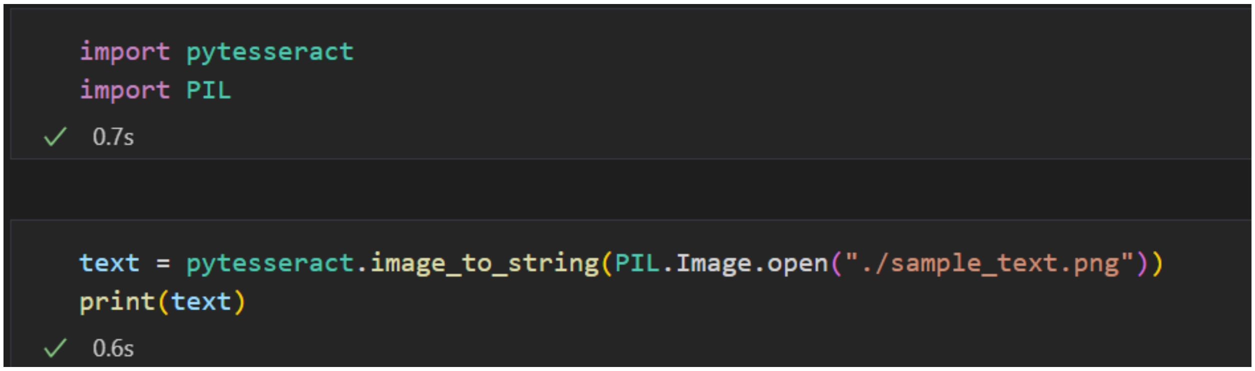
Input:



## SCENE FROM "DAN'L DRUCE."

This interesting domestic drama, by Mr. W. S. Gilbert, has continued to engage the sympathies of a nightly sufficient audience at the Haymarket Theatre, where it has now been represented more than sixty times. Its subject and character were described by us, in the ordinary report of theatrical novelties, about two months ago. Our readers will probably not need to be reminded that the hero of the story, Dan'l Druce, the blacksmith, is a solitary recluse dwelling on the coast of Norfolk, where his lone cottage is visited by fugitives from party vengeance during the civil wars of the Commonwealth. His hoard of money is stolen; but a different sort of treasure, a helpless female infant; is left by some mysterious agency, and may be accepted, as in George Eliot's tale of "Silas Marner," for a Divine gift to the sad-hearted misanthrope, far better than riches. In this spirit, at least, he is content to receive the precious human charge; and so to those who would remove it from his home, Dan'l Druce here makes answer with the solemn exclamation, "Touch not the Lord's gift!" This character is well acted by Mr. Hermann Vezin.

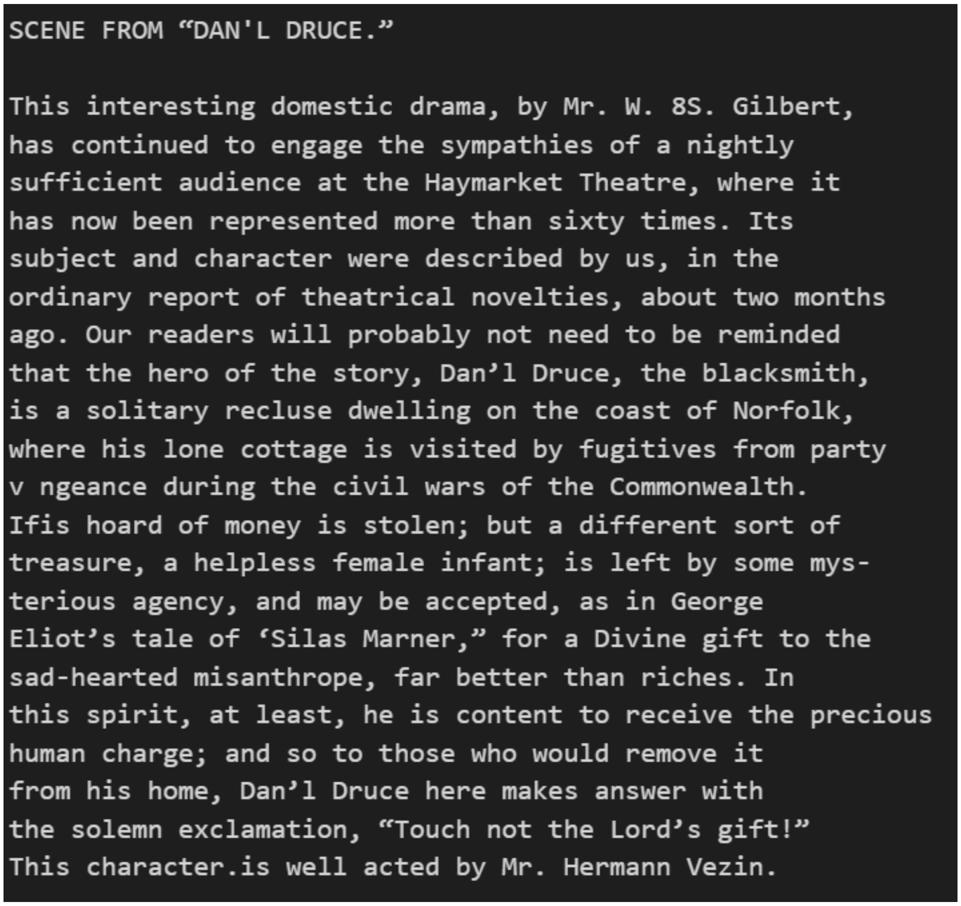
Code snippet:



```python
import pytesseract
import PIL
```

```python
text = pytesseract.image_to_string(PIL.Image.open("./sample_text.png"))
print(text)
```

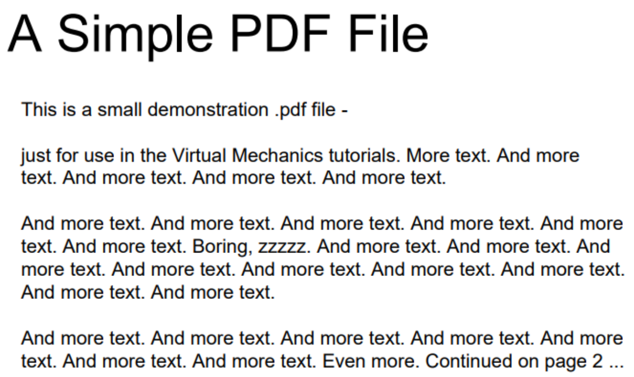
Result:



```
SCENE FROM "DAN'L DRUCE."

This interesting domestic drama, by Mr. W. 8S. Gilbert,
has continued to engage the sympathies of a nightly
sufficient audience at the Haymarket Theatre, where it
has now been represented more than sixty times. Its
subject and character were described by us, in the
ordinary report of theatrical novelties, about two months
ago. Our readers will probably not need to be reminded
that the hero of the story, Dan'l Druce, the blacksmith,
is a solitary recluse dwelling on the coast of Norfolk,
where his lone cottage is visited by fugitives from party
v ngeance during the civil wars of the Commonwealth.
Ifis hoard of money is stolen; but a different sort of
treasure, a helpless female infant; is left by some mys-
terious agency, and may be accepted, as in George
Eliot's tale of 'Silas Marner," for a Divine gift to the
sad-hearted misanthrope, far better than riches. In
this spirit, at least, he is content to receive the precious
human charge; and so to those who would remove it
from his home, Dan'l Druce here makes answer with
the solemn exclamation, "Touch not the Lord's gift!"
This character.is well acted by Mr. Hermann Vezin.
```

## Detecting text in PDF file:

Input:



# A Simple PDF File

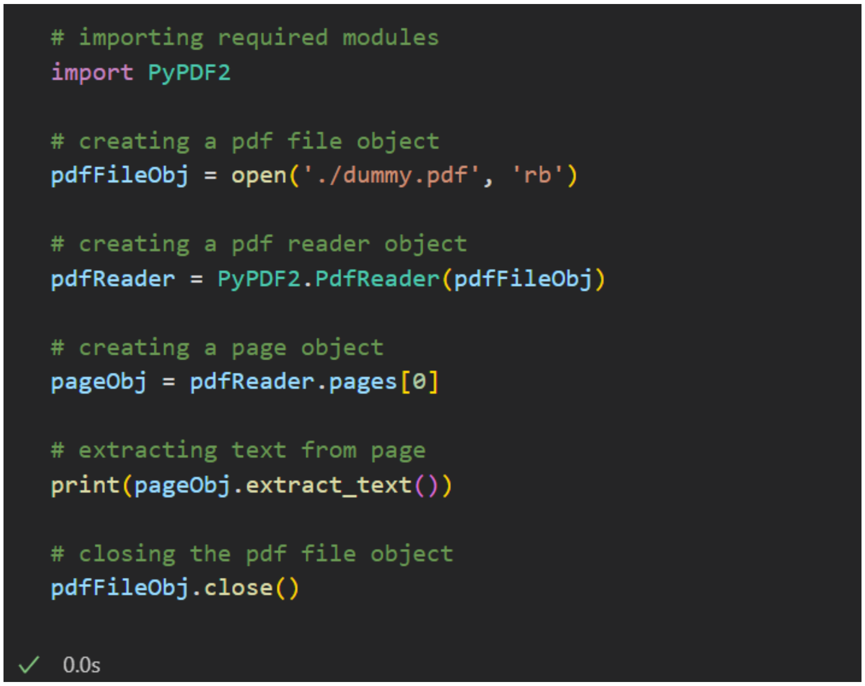
This is a small demonstration .pdf file -

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.

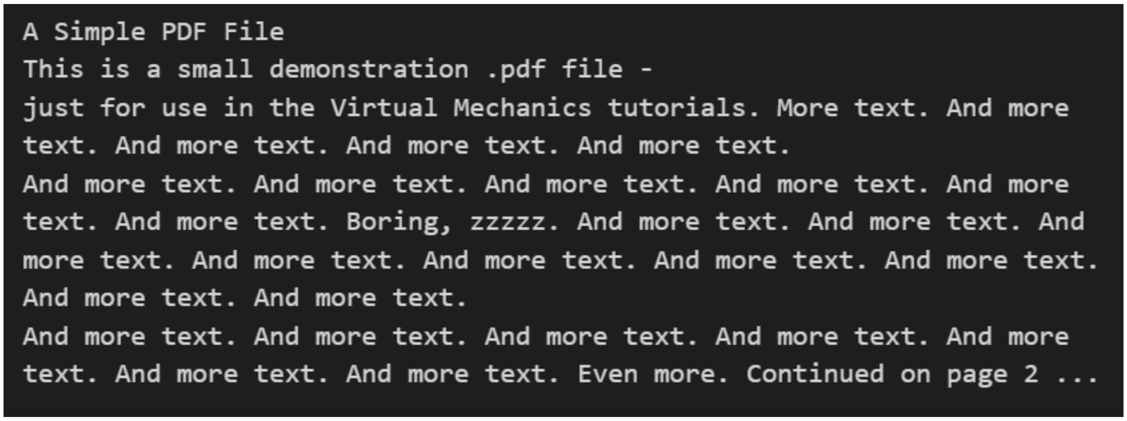
And more text. And more text. And more text. And more text. And more text. And more text. And more text. Even more. Continued on page 2 ...

Code snippet:



```python
# importing required modules
import PyPDF2

# creating a pdf file object
pdfFileObj = open('./dummy.pdf', 'rb')

# creating a pdf reader object
pdfReader = PyPDF2.PdfReader(pdfFileObj)

# creating a page object
pageObj = pdfReader.pages[0]

# extracting text from page
print(pageObj.extract_text())

# closing the pdf file object
pdfFileObj.close()
```

Result:



```
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. And more
text. And more text. And more text. Even more. Continued on page 2 ...
```

## Conclusion:

The project demonstrates a proof of concept for development of a system that can extract text from pdf documents and image files.