Anjuman College of Engineering and Technology

“Optical Character Recognition Based Webapp”

Under the guidance of
Prof. Kamlesh Kelwade
(Assiociate Prof,CSE)

Submitted by
Akshay Gharde (28)
Asit Damke (29)
Pratik Sahare(07)
Sumit Tonge(32)
Vipin Suryawanshi(33)

INTRODUCTION

- We are living in a data driven society where the generation and consumption of data is witnessing an exponential growth.

- The Physical form of data sources such as textbooks,invoices,precriptions , reports , bulletins , etc occupy a significant portion and are expected to grow in coming times.

- The physical sources of data do possess inherent challenges of storage and safeguarding and are vulnerable to fading, damage or being misplaced, even the cloning requires a lot of manual effort.

- The data present in physical form needs to be transformed into digital format to facilitate efficient storage, retrieval, sharing and back up of the information

- These days there is a huge demand in “storing the information available in these paper documents in to a computer storage disk.

- One simple way to store information in these paper documents in to computer system is to first scan the documents and then store them as IMAGES.

- But to reuse this information it is very difficult to read the individual contents and searching the contents form these documents line-by-line and word-by-word

Aim and Objective

- The project aims to obtain an image containing text from the user through django Interface
- Then the user have to select the language of text in the the image
- The image is then processed using pytesseract library and the text inside the image will be extracted
- Then the extracted text will be displayed on the app and will be available to download in text or pdf format

Proposed Work

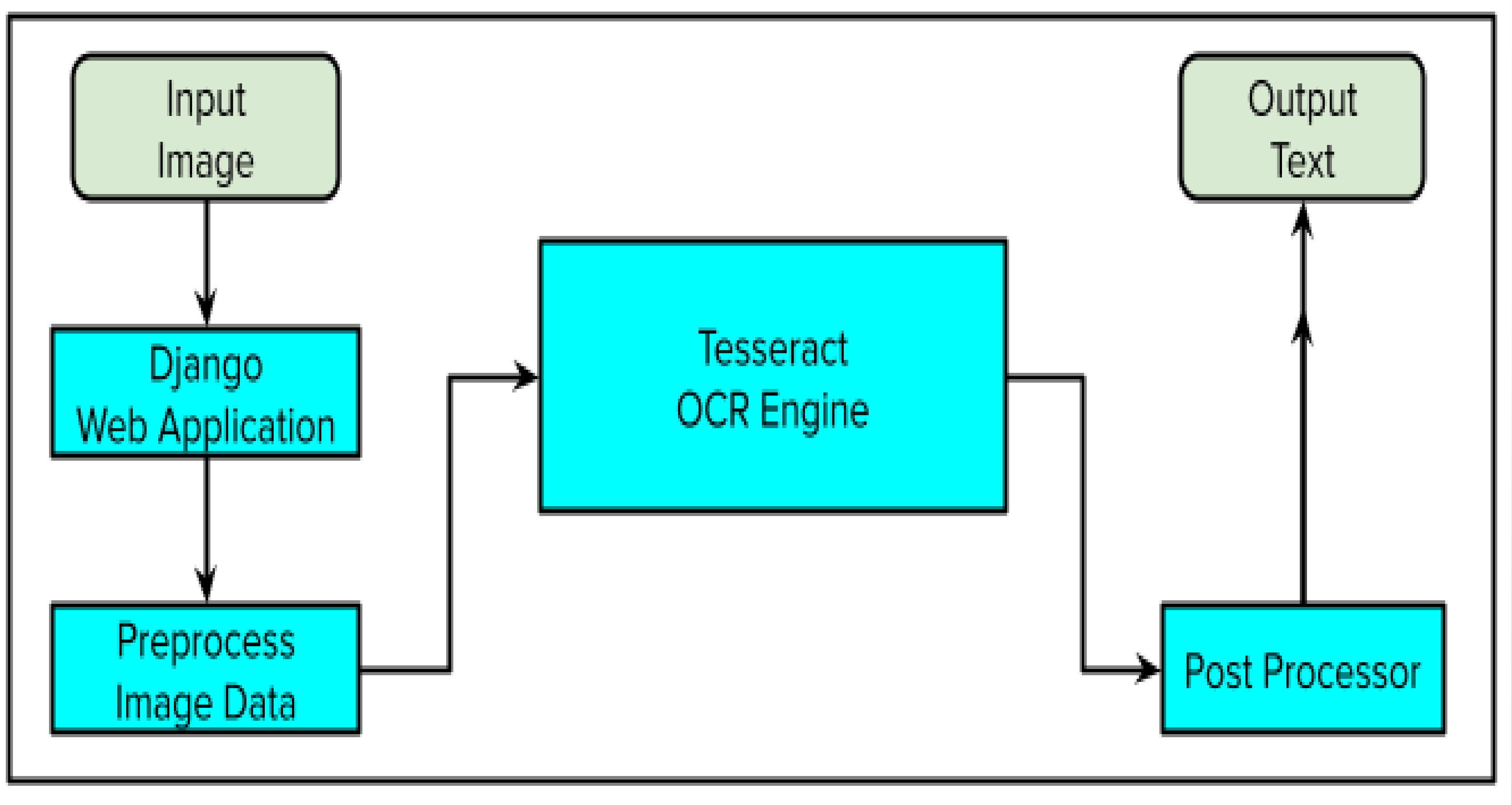
- Our proposed system is Django (Python Framework) based Web application

- Once the user click extract button , it triggers our ocr pipeline which consists of image pre-processing operations such as noise removal and localization

- The processed image is fed into the pyTesseract OCR engine where textual characters are indentified and categorized
- And then the OCR engine outputs the extracted which can be downloaded in pdf or txt format

- We have sqlite3 as our database.

Architecture of the proposed system



Main Technology used

- Python
- Django
- Bootstrap
- Html & CSS
- Javascript
- Sqlite3
- pytesseract
- Visual Code Studio

Steps envolved to use the app

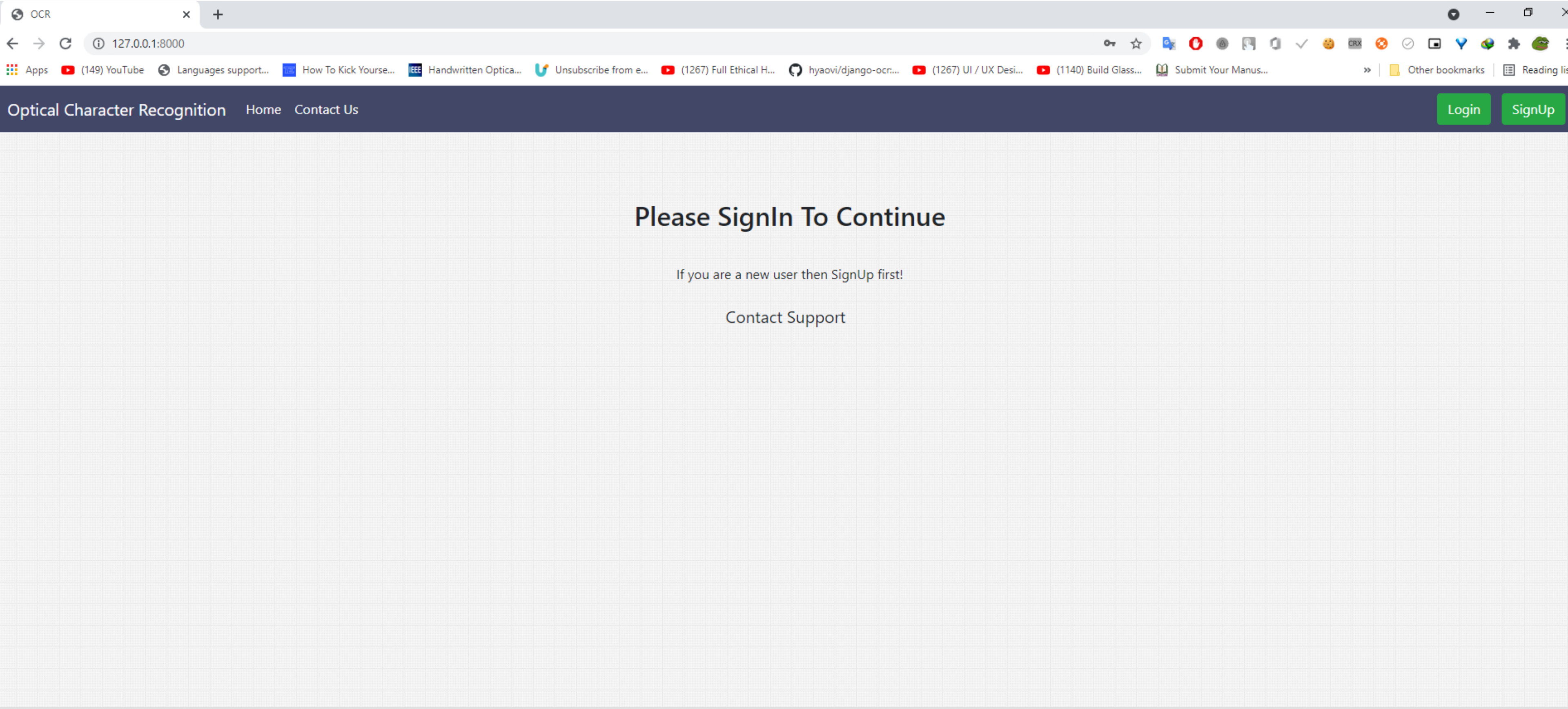
Step 1: In order to use the app user first need to register and sign in.

Step 2:Then the user have to upload the image of which they wanna extract the text.

Step 3: Then the user have to select the language of the text on the image

Step 4: As soon as the user clicks extract button , the extracted text will be displayed on the app and the user will also will able to copy the text using copy button and download the extracted text in text and pdf file using respective buttons

Landing Page

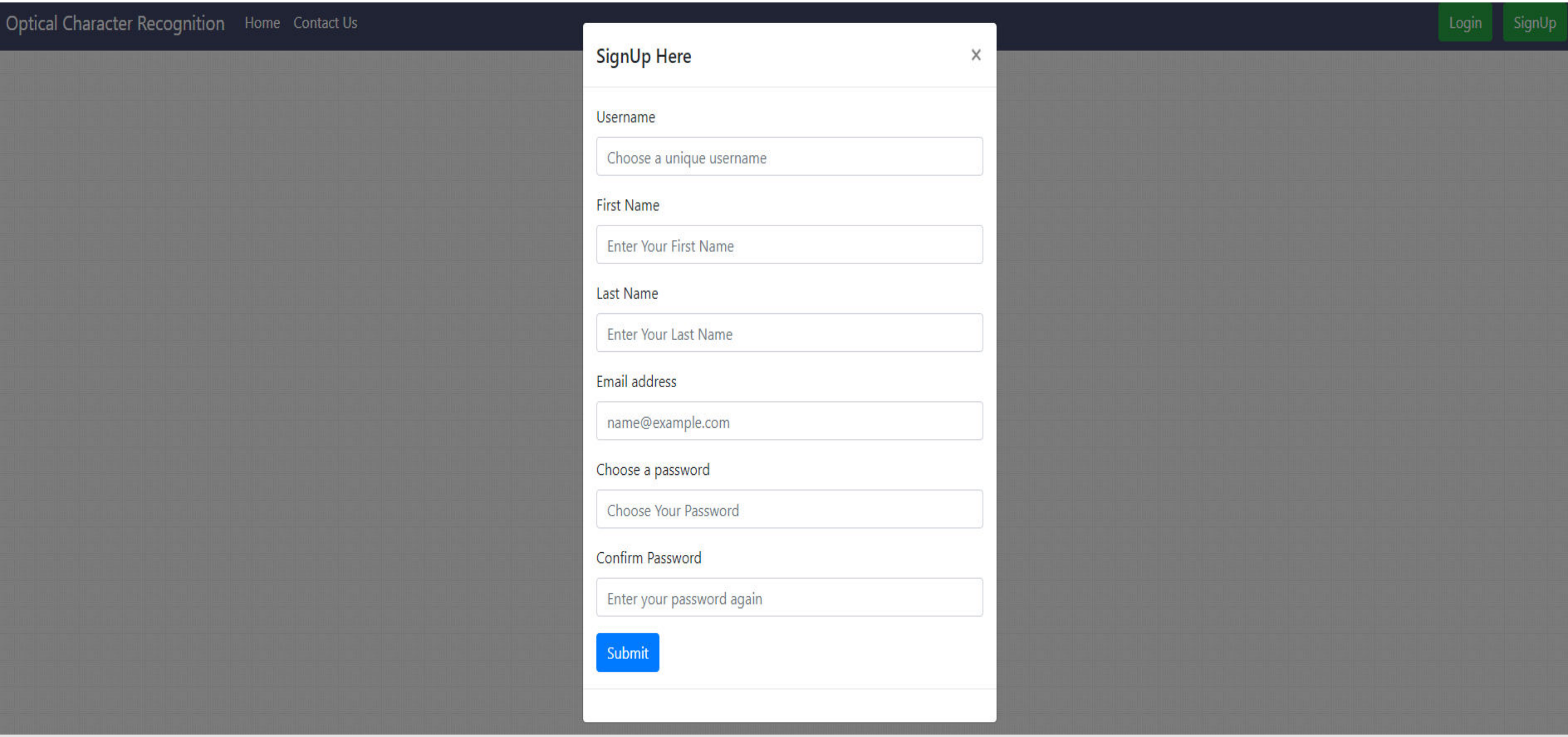


Please SignIn To Continue

If you are a new user then SignUp first!

Contact Support

SignUp Form



SignUp Here

Username

Choose a unique username

First Name

Enter Your First Name

Last Name

Enter Your Last Name

Email address

name@example.com

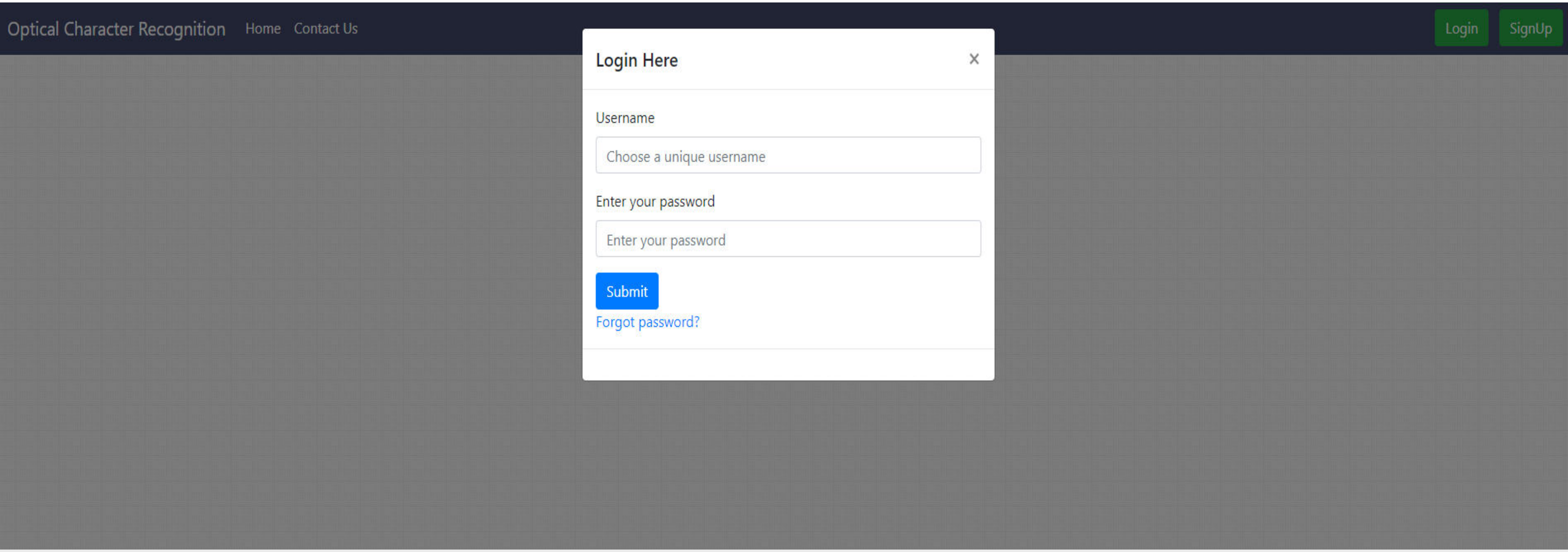
Choose a password

Choose Your Password

Confirm Password

Enter your password again

SignIn Form



Optical Character Recognition Home Contact Us Login SignUp

Login Here ×

Username

Choose a unique username

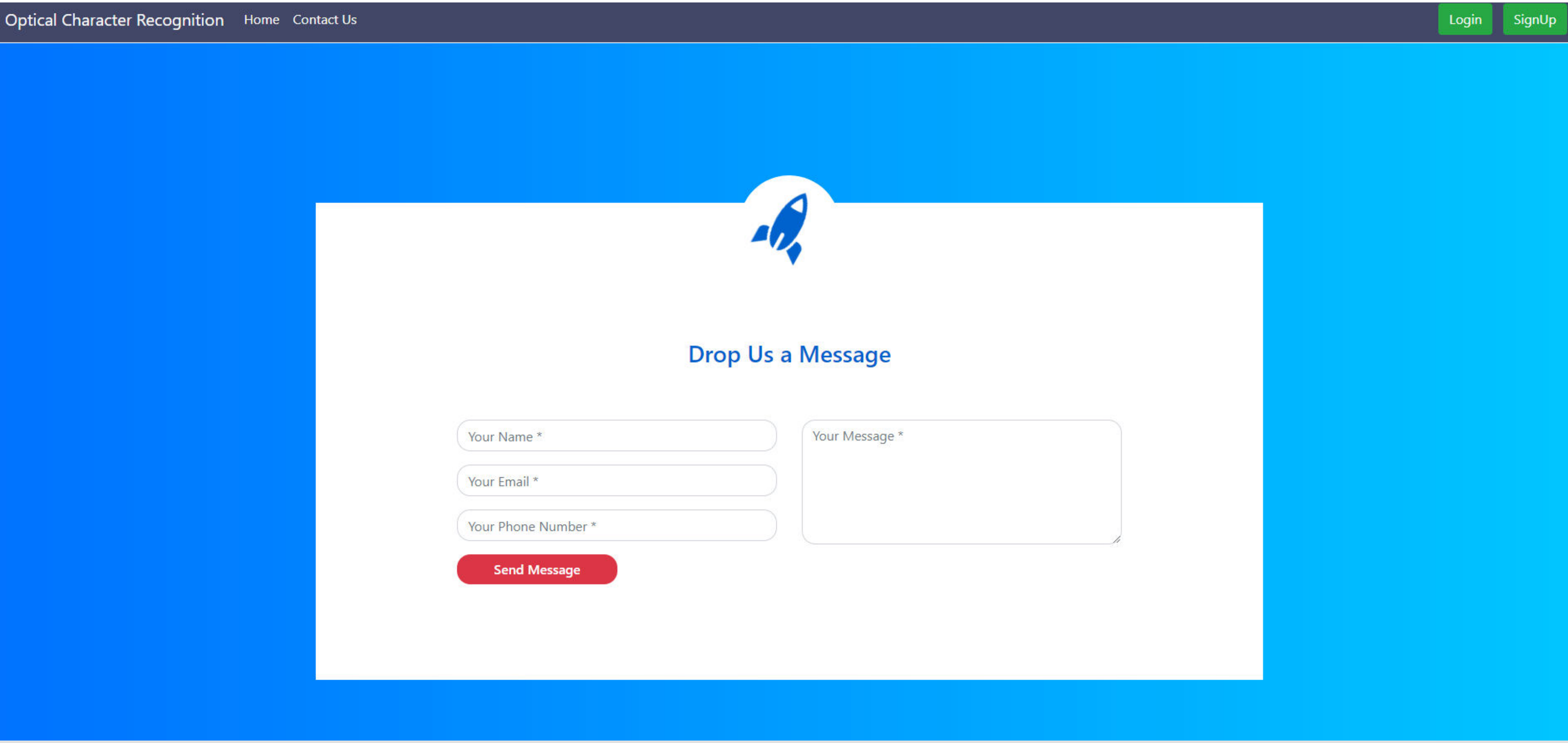
Enter your password

Enter your password

Submit

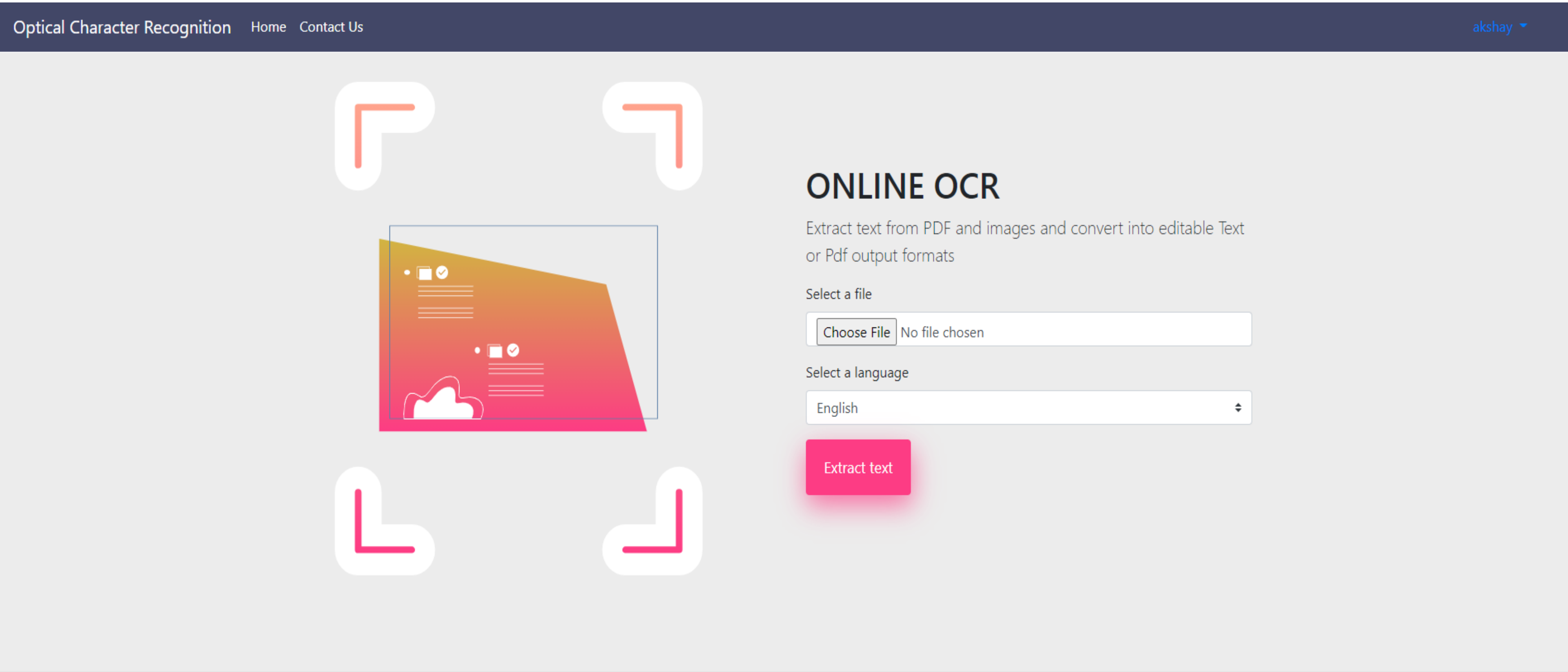
Forgot password?

ContactUs



Drop Us a Message

Main App



ONLINE OCR

Extract text from PDF and images and convert into editable Text or Pdf output formats

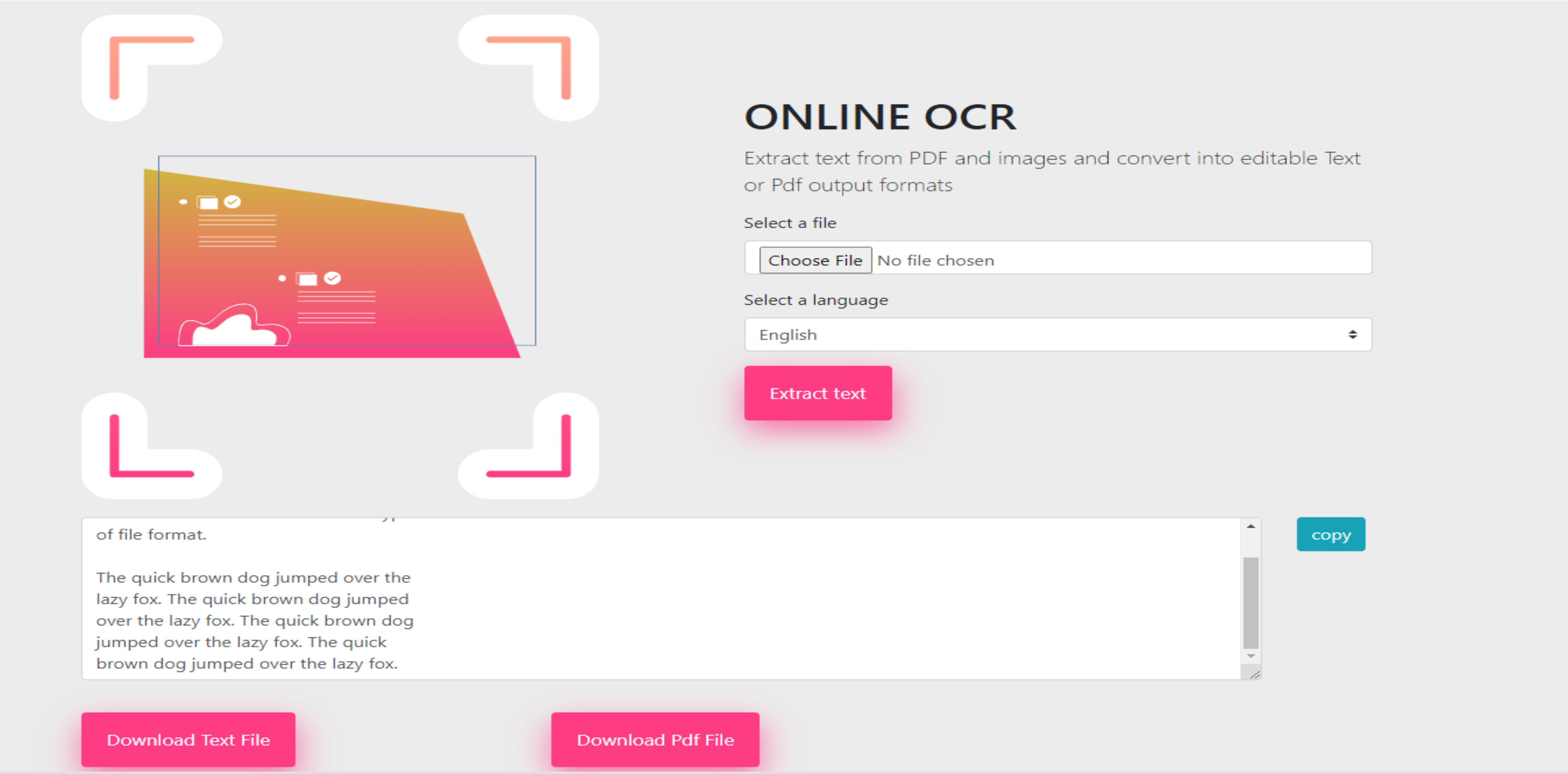
Select a file

No file chosen

Select a language

English

Output 1



ONLINE OCR

Extract text from PDF and images and convert into editable Text or Pdf output formats

Select a file

Choose File No file chosen

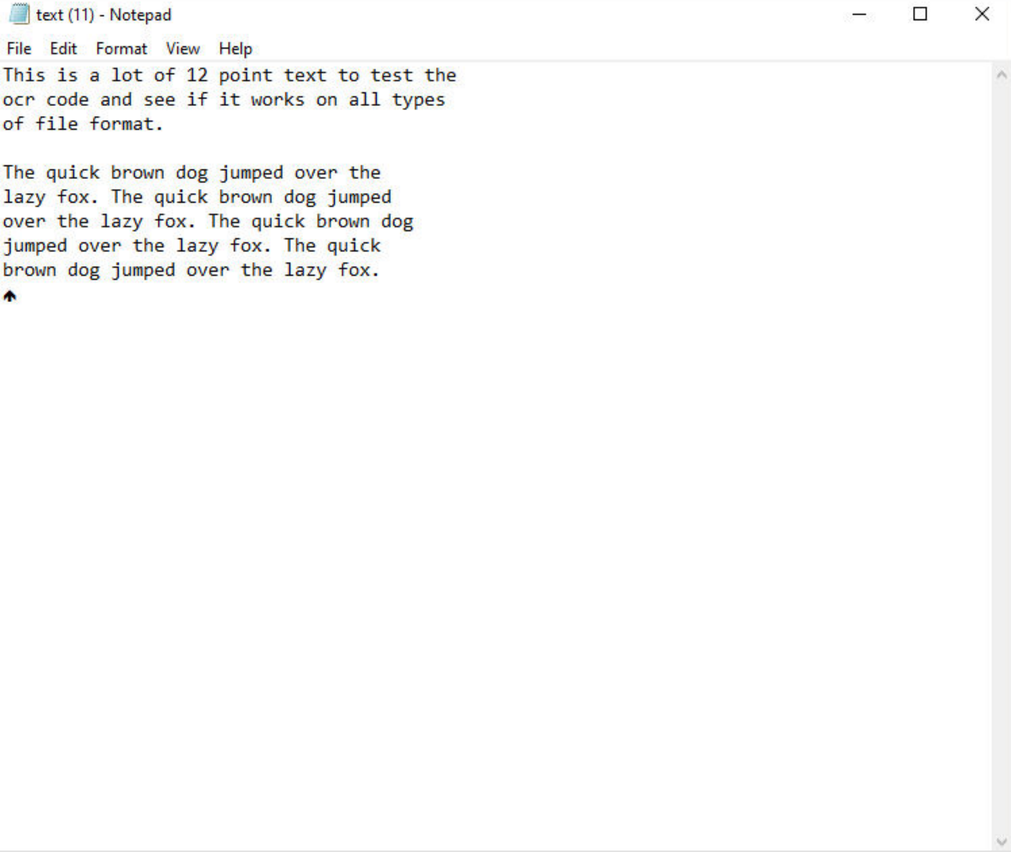
Select a language

English

of file format.

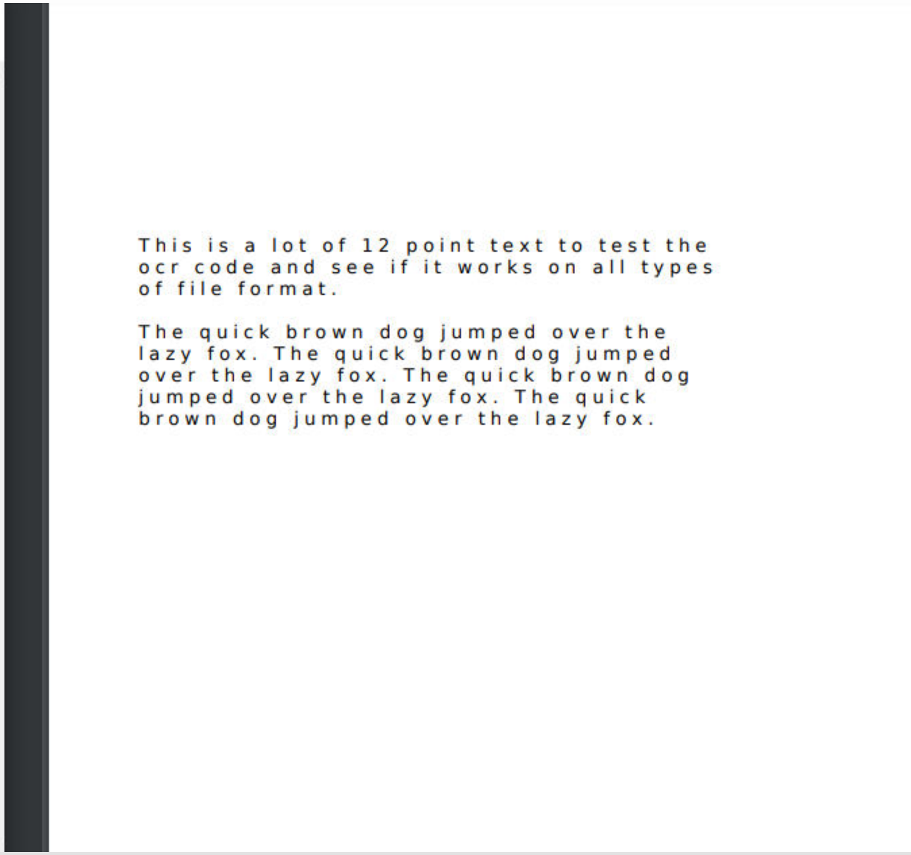
The quick brown dog jumped over the lazy fox. The quick brown dog jumped over the lazy fox. The quick brown dog jumped over the lazy fox. The quick brown dog jumped over the lazy fox.

Output 2



```
This is a lot of 12 point text to test the
ocr code and see if it works on all types
of file format.

The quick brown dog jumped over the
lazy fox. The quick brown dog jumped
over the lazy fox. The quick brown dog
jumped over the lazy fox. The quick
brown dog jumped over the lazy fox.
```


This is a lot of 12 point text to test the
ocr code and see if it works on all types
of file format.

The quick brown dog jumped over the
lazy fox. The quick brown dog jumped
over the lazy fox. The quick brown dog
jumped over the lazy fox. The quick
brown dog jumped over the lazy fox.

System Requirements

- Processor: 1 gigahertz (GHz) frequency or above
- RAM: A minimum of 500mb
- Active Internet Connection:
- OS: windows xp and above
- Browser: Preffered google chome

ADVANTAGES

- High Speed
- Good Accuracy
- Multilingual
- Local Language Support
- Output in txt and pdf format available

Disadvantages

- Less accuracy against local languages
- Less accuracy against special characters

Results

- In this project, we were successfully able to develop a robust and modular web application for image text extraction and multilingual translation using the Pytesseract based OCR Engine.

- The system was able to extract text from handwritten and printed documents with high accuracy which further strengthens the fact that OCR based applications can bring a lot of convenience to our daily activities and streamline a lot of workflows that can result in the efficient storage, retrieval, sharing and back up of the information.

- Although the results look promising, the OCR systems need to be made customizable so that they can be trained so as to efficiently recognize the characters from all sorts of handwritten data sources

REFERENCES

- [1] Smith, R. (2007, September). An overview of the Tesseract OCR engine. In Ninth international conference on document analysis and recognition (ICDAR 2007) (Vol. 2, pp. 629-633). IEEE.

- [2] Rekha, M. (2021). Educational Training For Processing Invoice Of Vendor Identification And Payments Using Python-Tesseract. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(11), 224-228.

- [3] https://www.reportlab.com/docs/reportlab-userguide.pdf.

- [4] https://pypi.org/project/pytesseract/

THANK YOU