

# README for OCR and Document Search Web Application


This project is a web application that allows users to upload an image containing text (both English and Hindi), extract the text from the image using a vision-language model (Qwen2-VL-2B), and perform keyword search within the extracted text. The application uses Gradio for the interface and can be run locally. This README will guide you through setting up the environment, running the web application, and deploying it.

## Prerequisites

1. **Python 3.8 or above:** Ensure you have Python installed. You can check the version using:

```
python --version
```

bash

 Copy code

```
python --version
```

## Setting up the Environment


1. **Download the repository from** <https://github.com/MehwishSameer/Assignment>
2. **Create a virtual environment (optional but recommended):**

```
python -m venv venv
```

```
source venv/bin/activate # For Linux/Mac
```

```
venv\Scripts\activate # For Windows
```

bash

 Copy code

```
python -m venv venv
source venv/bin/activate # For Linux/Mac
venv\Scripts\activate # For Windows
```

### 3. Install dependencies:

Inside the project directory, run:

```
pip install -r requirements.txt
```

```
bash Copy code

pip install -r requirements.txt
```

## Running the Web Application Locally

### 1. Run the Gradio Application:

Once the environment is set up, run the following command:

```
python app.py
```

```
bash Copy code

python app.py
```

This will start the web server and the Gradio interface should automatically open in your web browser. If it doesn't, the terminal will display a local URL (e.g., `http://127.0.0.1:7860/``). Open this URL manually in your browser.

### 2. Using the Application:

- Upload an image with Hindi and/or English text.
- The text will be extracted and displayed in the "Extracted Text" box.
- Enter a keyword in the search box to highlight it in the extracted text. The search results and the time taken to perform the search will be displayed.

## Deploying the Application on Hugging Face

**1. Create a Hugging Face Account:** Sign up for an account at [Hugging Face](#). If you already have an account, log in.

### 2. Create a New Space:

- Navigate to the Spaces section of the Hugging Face website.
- Click on the "Create a Space" button.

- Fill in the necessary details, such as the Space name and the license. Choose "Gradio" as the SDK.

### 3. Upload Your Code:

- After creating your Space, you will be directed to the Space's repository.
- Upload Python script ( `app.py` ) and other required files ( `requirements.txt` , images folder) using the interface.

### 6. Deploy Your Space:

- After uploading all files, Hugging Face will automatically build and deploy your application.
- You can monitor the build process and see logs in real-time.

### 7. Access Your Application:

Once the deployment is complete, your Space will be live at a URL like <https://huggingface.co/spaces/Mehwish12/OCR>

## Input, Extracted Text and Search Output:

**Input:** The user uploads an image, and the extracted text is displayed in the "Extracted Text" box.

**Sample Image:**

विकिपीडिया

**मुम्बई**

महाराष्ट्र की राजधानी और जिला

भारत के पश्चिमी तट पर स्थित **मुम्बई** (पूर्व नाम बम्बई), भारतीय राज्य महाराष्ट्र की राजधानी है। इसकी अनुमानित जनसंख्या ३ करोड़ २९ लाख है जो देश की पहली सर्वाधिक आबादी वाली नगरी है।<sup>[2]</sup> इसका गठन लावा निर्मित सात छोटे-छोटे द्वीपों द्वारा हुआ है एवं यह पुल द्वारा प्रमुख भू-खंड के साथ जुड़ा हुआ है। मुम्बई बन्दरगाह भारतवर्ष का सर्वश्रेष्ठ सामुद्रिक बन्दरगाह है। मुम्बई का तट कटा-फटा है जिसके कारण इसका पोताश्रय प्राकृतिक एवं सुरक्षित है। यूरोप, अमेरिका, अफ्रीका आदि पश्चिमी देशों से

**Extracted Text:** मुम्बई महाराष्ट्र की राजधानी और जिला भारत के पश्चिमी तट पर स्थित मुम्बई (पूर्व नाम बम्बई), भारतीय राज्य महाराष्ट्र की राजधानी है। इसकी अनुमानित जनसंख्या 3 करोड़ 29 लाख है जो देश की पहली सर्वाधिक आबादी वाली नगरी है।[2] इसका गठन लावा निर्मित सात छोटे-छोटे द्वीपों द्वारा हुआ है एवं यह पुल द्वारा प्रमुख भू-खंड के साथ जुड़ा हुआ है। मुम्बई बन्दरगाह भारतवर्ष का सर्वश्रेष्ठ सामुद्रिक बन्दरगाह है। मुम्बई का तट कटा-फटा है जिसके कारण इसका पोताश्रय प्रकृतिक एवं सुरक्षित है। यूरोप, अमेरिका, अफ्रीका आदि पश्चिमी देशों से

**Keyword Search:** A keyword entered by the user is searched in the extracted text, and the result is displayed with highlighted keywords.

### Example Extraction and Keyword Searches:

OCR and Document Search Web Application Prototype

Upload Image

Enter a keyword to search  
संदिग्ध होना

हिंदी: संदिग्ध होना और अंगरेज़ी: Silence is the best answer to anger

OCR Processing Time  
OCR processing time: 135.48 seconds

OCR and Document Search Web Application Prototype

Upload Image

## 1000 English to Hindi Words

Words	Meanings	Roman
Chief	दार सर	Daar sar
Colony	कालोनी	Kaalonee
Clock	घड़ी	Ghadee
Mine	मेरी	Meree
Tie	गुलोबन्द	Guloband
Enter	दर्ज	Darj
Maior	प्रमुख	Pramukh

Enter a keyword to search  
प्रमुख Pramukh

Chief दार सर Daar sar Colony कालोनी Kaalonee Clock घड़ी Ghadee Mine मेरी Meree Tie गुलोबन्द Guloband Enter दर्ज Darj Major प्रमुख Pramukh Fresh ताज़ा Taaza Search खोज Khoj

OCR Processing Time  
OCR processing time: 613.44 seconds

OCR and Document Search Web Application Prototype

Upload Image

विकिपीडिया

### मुम्बई

महाराष्ट्र की राजधानी और जिला

भारत के पश्चिमी तट पर स्थित **मुम्बई** (पूर्व नाम बम्बई), भारतीय राज्य महाराष्ट्र की राजधानी है। इसकी अनुमानित जनसंख्या ३ करोड़ २९ लाख है जो देश की पहली सर्वाधिक आबादी वाली नगरी है।<sup>[2]</sup> इसका गठन लावा निर्मित सात छोटे-छोटे द्वीपों द्वारा हुआ है एवं यह पुल द्वारा प्रमुख भू-खंड के साथ जुड़ा हुआ है। मुम्बई बन्दरगाह भारतवर्ष का सर्वश्रेष्ठ सामुद्रिक बन्दरगाह है। मुम्बई का तट कटा-फटा है जिसके कारण इसका पोताश्रय प्राकृतिक एवं सुरक्षित है। यूरोप, अमेरिका, अफ्रीका आदि पश्चिमी देशों से

Enter a keyword to search

है एवं यह पुल द्वारा प्रमुख भू-खंड के साथ

मुम्बई महाराष्ट्र की राजधानी और जिला भारत के पश्चिमी तट पर स्थित मुम्बई (पूर्व नाम बम्बई), भारतीय राज्य महाराष्ट्र की राजधानी है। इसकी अनुमानित जनसंख्या 3 करोड़ 29 लाख है जो देश की पहली सर्वाधिक आबादी वाली नगरी है। [2] इसका गठन लावा निर्मित सात छोटे-छोटे द्वीपों द्वारा हुआ है एवं यह पुल द्वारा प्रमुख भू-खंड के साथ जुड़ा हुआ है। मुम्बई बन्दरगाह भारतवर्ष का सर्वश्रेष्ठ सामुद्रिक बन्दरगाह है। मुम्बई का तट कटा-फटा है जिसके कारण इसका पोताश्रय प्राकृतिक एवं सुरक्षित है। यूरोप, अमेरिका, अफ्रीका आदि पश्चिमी देशों से

OCR Processing Time

OCR processing time: 1356.93 seconds

OCR and Document Search Web Application Prototype

Upload Image

**Translation with Extra Knowledge**

एक इंसान हमेशा ही शांति और खुशी की तलाश में रहता है। वह अक्सर यह महसूस करता है कि उन्हें खुश रखने के लिए धन और दौलत जरूरी है। लेकिन यह पूरी तरह से सच नहीं है। हम लोग इस दुनिया में बहुत सारे अमीर लोगों को भी उदास पाते हैं। इंसान आमतौर पर उस चीज से संतुष्ट नहीं होता है जो उसके पास रहता है। वह अधिक और अधिक

Enter a keyword to search

लिए धन और दौलत

Translation with Extra Knowledge एक इंसान हमेशा ही शांति और खुशी की तलाश में रहता है। वह अक्सर यह महसूस करता है कि उन्हें खुश रखने के लिए धन और दौलत जरूरी है। लेकिन यह पूरी तरह से सच नहीं है। हम लोग इस दुनिया में बहुत सारे अमीर लोगों को भी उदास पाते हैं। इंसान आमतौर पर उस चीज से संतुष्ट नहीं होता जो उसके पास रहता है। वह अधिक और अधिक

OCR Processing Time

OCR processing time: 1431.82 seconds

Additional Notes

- **CUDA (GPU):** If CUDA is available, the application will run on the GPU for faster processing. Otherwise, it will run on the CPU.
- **Keyword Search:** The application supports case-insensitive keyword search and highlights all instances of the keyword.