

Welcome!

Smart India hackathon 2020

Organization: Government of Andhra Pradesh

Team Name :- Editha
Code :- KB145

Team Members:

Abhishek Kumar (Team Leader)
Ritik Gupta
Harshita Madhok
Abhishek Khatri
Vaibhav Kumar
Ashish

Mentors

Mohd Ilyas
Shobhit Bhatnagar



01

Brief Description

Problem Statement

AI & OCR: To search Telugu,Urdu & English words in pdf present in Unicode & images format.

Our Idea

We will make a online web portal where user can upload the pdf and image from which he wish to Extract and search the data. The portal will provide an interactive user interface and it will display the desired data from pdf/images with the help of OCR and Machine Learning. The platform will provide support for English, Telugu, Urdu languages.

02

FEATURES

- This portal will provide an interactive platform for any person to Extract and search data in pdf and images.
- Extract Telugu, Urdu, English, text in pdf, or images.
- User friendly(user can search for data in a specific language or in a combination of all 3 languages.
- This portal will allow users to search data in Unicode as well in image format.
- Voice Searching (User can search for data using voice).
- Fast text extraction and voice searching.
- It's available for windows as well as Linux. Extendable (we can add more languages also).
- It is very security (Data upload on a portal is an encrypted format)

03

Advantages

- Extract Telugu, Urdu, English, text in pdf, or images in single time.
- Voice Searching.
- User friendly
- This portal will allow users to search data in Unicode as well in image format.
- Fast text extraction and voice searching.
- It's available for windows as well as Linux.
- Extendable.
- It is very security.



04

MODULES



Front end:- All UI/UX part.







Backend:-

- Telugu OCR with image or pdf
- English OCR with image or pdf
- Urdu OCR with image or pdf
- Extract Telugu, Urdu, English text from image or pdf
- Text Searching (simple or voice)



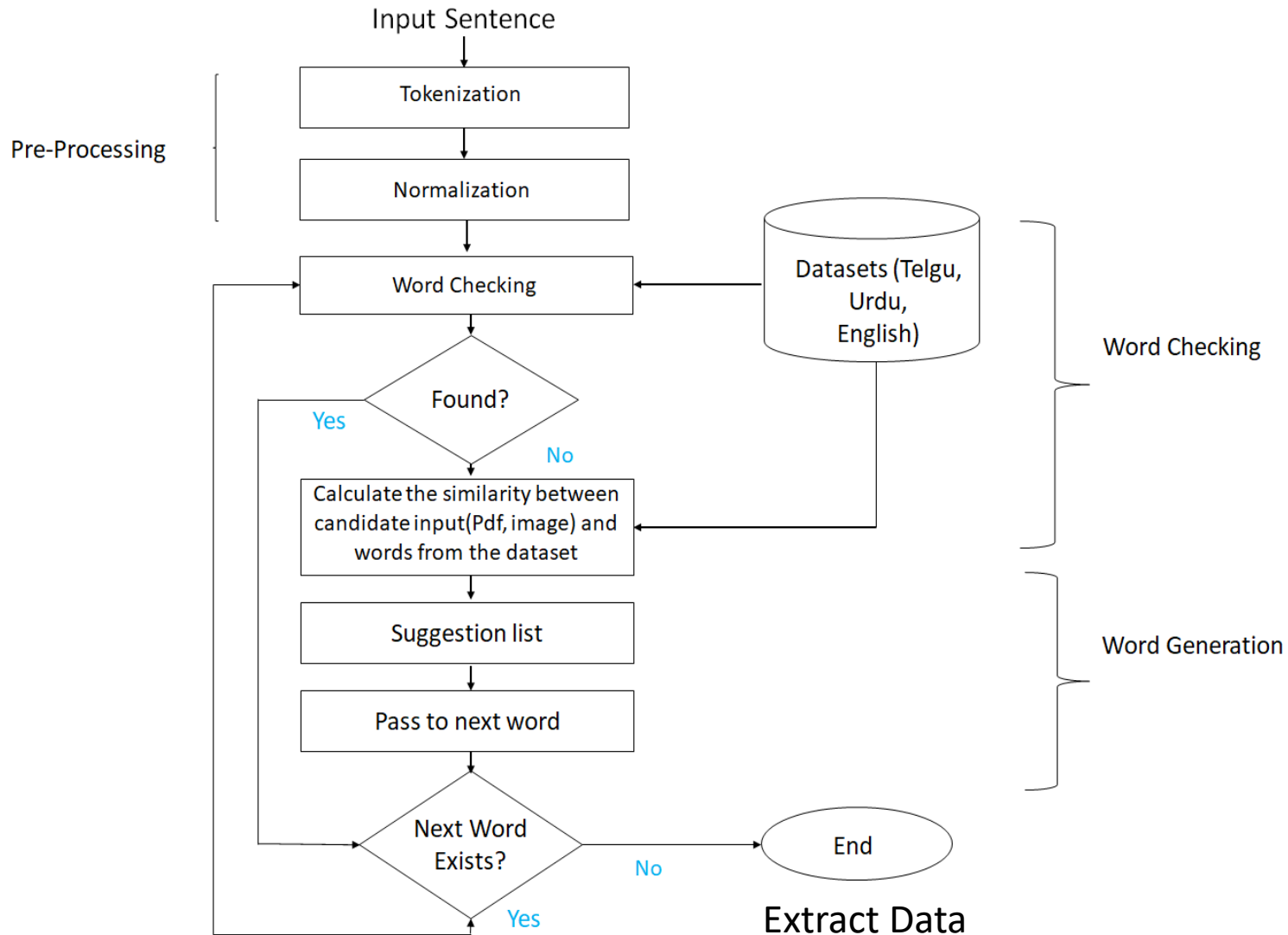
05

LIBRARIES AND FRAMEWORK USED

-  **Tesseract** is an optical character recognition engine for various operating systems.
-  For pdf rendering we use poppler for window.
-  For Front end :- HTML, CSS, JS, AJAX
-  For backend :- Python, Flask, ML

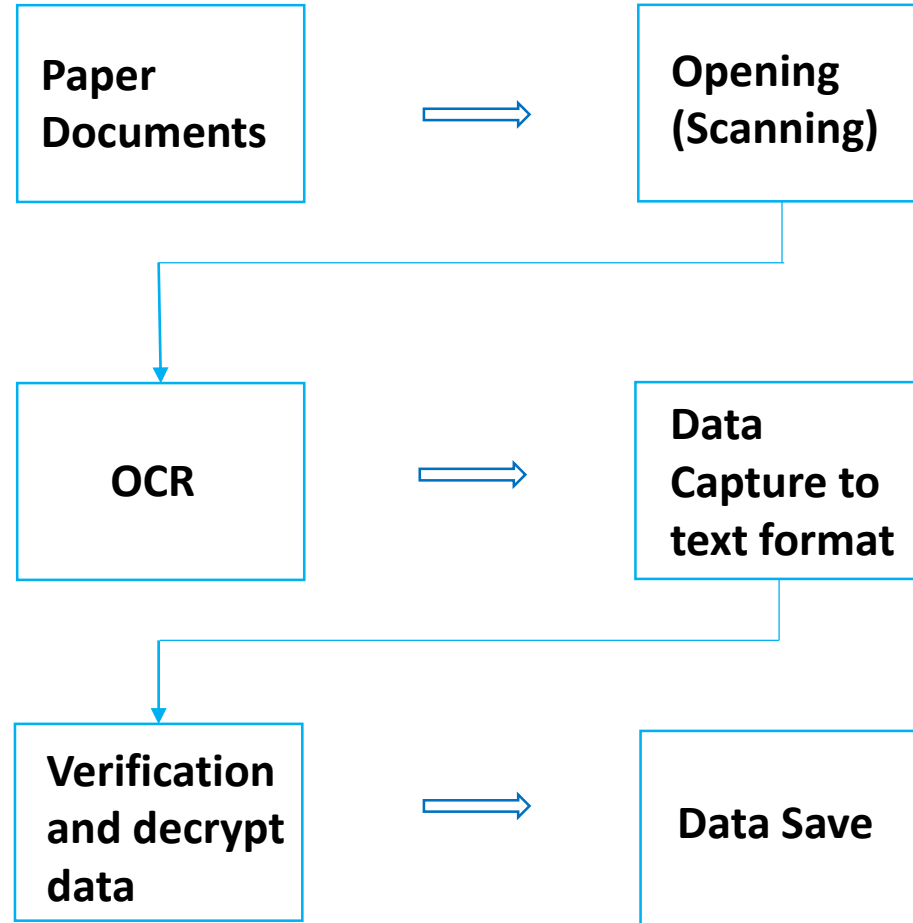
06

BASIC FLOW OF DATA EXTRACTION



07



OCR PROCESS





08

FUTURE SCOPE

-  We can add all other language also in future.
-  We convert one to another language.

09

THANK **Y**OU!