**Smart Ocr Application For**

**Document Digitization**

**Team members*:***

**1.P.Praneeth kumar-19R11A04J0**

**2.Reddy Rajula Udaykiran-19R11A04J3**

**3.Shaik Mohammad Siddique-19R111A04J4**

**4.Somireddy Bhavani Prasad-19R11A04J6**

**5.Tanniru.Sravani-19R11A04J7**

**INTRODUCTION**

**Overview:**

**OCR, or Optical Character Recognition, is a process of recognizing text ould be of handwritten text, prinside images and converting it into an electronic form. These images cinted text like documents, receipts, name cards, etc., or even a natural scene photograph.**

* **There are 2 different techniques in optical character recognition: pattern recognition and feature extraction. Pattern recognition refers to recognize characters in their entirety. Feature extraction means detecting the individual lines and strokes characters are made from and identifying them that way**
* **OCR accuracy can be measured by the portion of characters in a text that the OCR tool can extract extract without mistakes. For example, 99% accuracy means that 990 out of 1000 characters are correctly recognized.**
* **At the start of the digital revolution, when most of the printed information was being uploaded on the web, manual data entry of such humongous printed data (like those of newspaper collections) became a task that required time and patience. This data entry task was also prone to human errors. The outcome of this problem was in the form of the birth of OCR. It was invented in the early eighties. OCR is now mature enough to grab characters and words from images to extract meaningful information. This technology has now attained a near-perfect text detection accuracy.**

**Purpose:**

* **Increased Security: Physical documents cannot be tracked but scanned documents can be tracked. The access to the digitized documents can also be restricted.**
* **Saves Space: Digital storage spaces are way cheaper to take on rent than physical spaces. Eliminating the need for paper storage can help save more space, thereby reducing the requirement of physical space.**
* **Risk Mitigation: In case of a natural or manmade disaster, the risk of losing physical documents is mitigated by digitizing them.**
* **Ease of access: Digitized documents can be easily accessed as compared to physical ones. The search for a physical document in storage takes way more time than the search for a digitized document on a digital storage device or network.**
* **Cost Efficiency: The cost of producing digital documents is fractional compared to the humongous cost of paper, printer, ink, storage all put together.**
* **Ease of data search: Ease of data search inside a bunch of documents is possible in case of digitized documents but one has to physically examine each physical document to search particular data. Digitizing documents can save tons of time.**
* **Environment Friendly: Digitizing documents saves paper, ink(chemicals), etc., saving the a lot of energy resources..**
* **Data Sharing: Easier and quicker data sharing is possible in the case of digitized data but not in case of physical data.**

**LITERATURE SURVEY**

**Problem statement:**

* **The problem here for the software system to recognize characters in computer system.**
* **when information is scanned through the paper documents as we know that we have number of news papers and books which are in printed format related to different subjects. Whenever we scan the documents through the scanners, the document are stored as images such as jpeg,gif etc.., in the computer systems .these images cannot or edited by the user.**
* **But to resue these information it is very difficult to read the individuals contents and searching the contents from these documents line-by-line and word by word .these days there is is a huge demand in "storing the information available in this paper documents in to a computer storage disk and then later editing or reusing this information by searching process".**

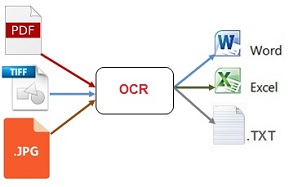
**Proposed Solution:**

**Our proposed System is OCR .Which is a character recognition system that supports recognition of characters of multiple languages. OCR software processes a digital image by locating and recognizing**[**characters**](https://techterms.com/definition/character)**, such as letters, numbers, and symbols. Some OCR software will simply export the text, while other programs can convert the characters to editable text directly in the image. Advanced OCR software can export the size and formatting of the text as well as the layout of the text found on a page. It supports multiple functionalities to be performed on the document .**

**The multiple functionalities include editing and searching.**

**THEORITICAL ANALYSIS**

**Block Diadram:**

****

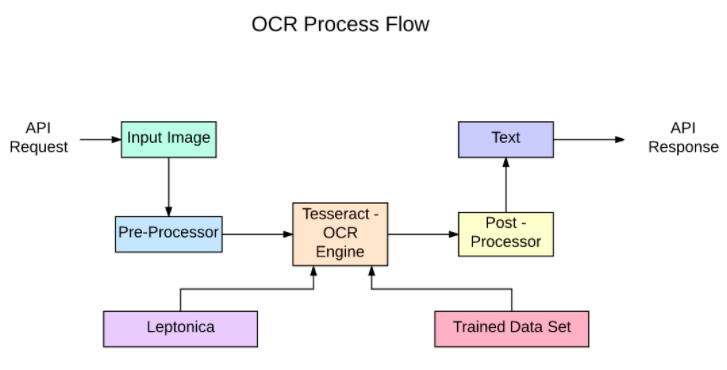
**Hardware/Software Designing:**

**Anaconda/ PyCharm IDE is Ideal to complete this project.**

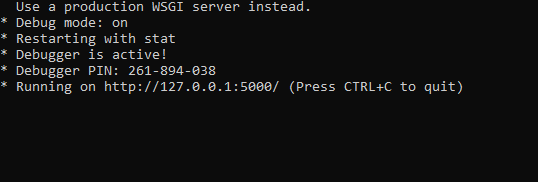
Make sure you have installed all the following necessary libraries:

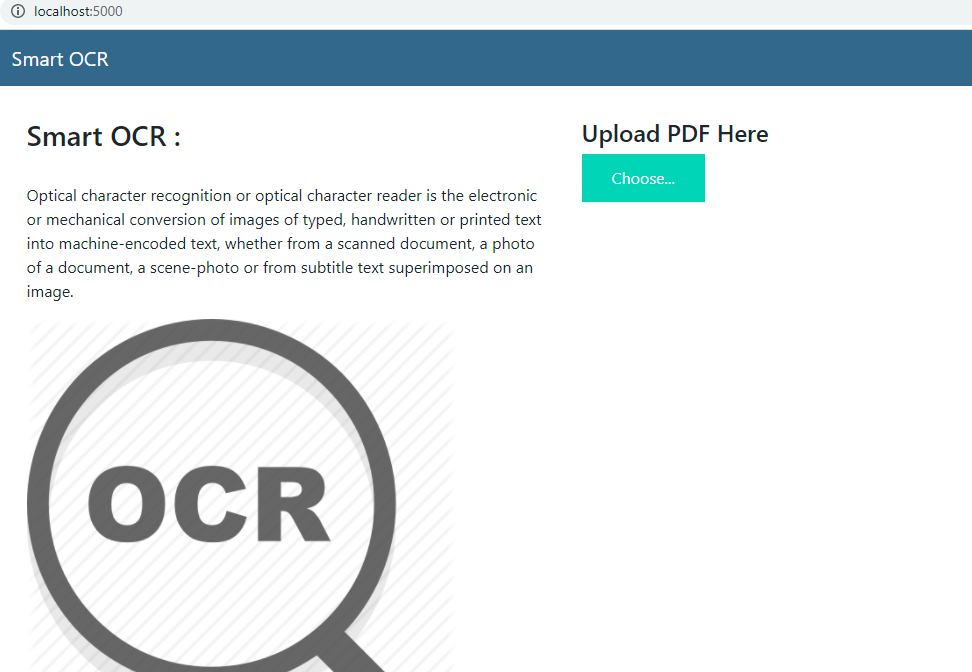
* **pytesseract -**OCR package in python
* **pdf2image**- Converting PDF to Image
* **tesseract-ocr execution file** -Backend used for pytesseract
* **poppler**-Supporting file for pdf2image package
* **Open**-**cv-**image processing

**FLOW CHART**

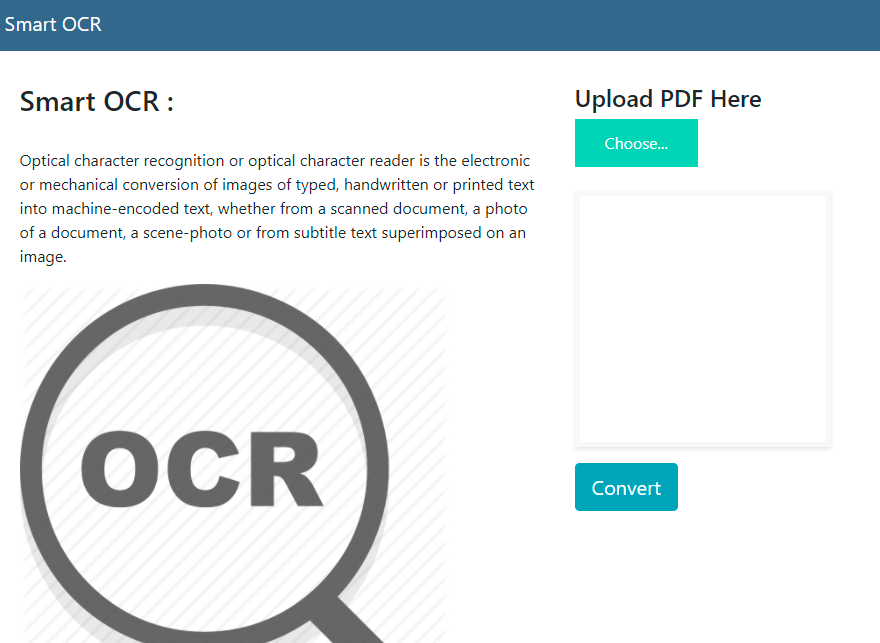
****

**OUTPUT:**

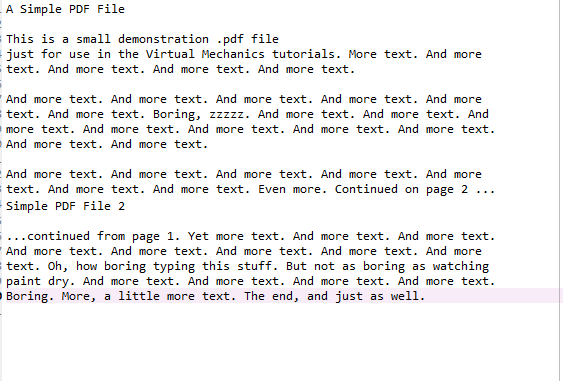
****



* **Choose the pdf from the folder called upload**

****

* **click on convert .**
* **When we click on convert.**
* **It will extracted in to the text format.**

****

**Advantages :**

**processing of OCR information is fast. Large quantities of text are often input quickly.**

* **A paper based form are often became an electronic form which is straightforward to store or send by mail.**
* **It is cheaper than paying someone amount to manually enter great deal of text data. Moreover it takes less time to convert within the electronic form.**
* **The latest software can re-create tables also as original layout.**
* **This process is much faster as compared to the manual typing the information into the system**
* **Advanced version can even Re create tables, columns and even produce sites.**

**Disadvantages :**

**OCR text works efficiently with the printed text only and not with handwritten text. Handwriting must be learnt by the pc.**

* **OCR systems are expensive.**
* **There is the need of lot of space required by the image produced.**
* **The quality of the image can be lose during this process.**
* **Quality of the ultimate image depends on quality of the first image.**
* **All the documents got to be checked over carefully then manually corrected.**
* **Not 100% accurate, there are likely to be some mistakes made during the method.**

**Applications:**

* [**Data entry**](https://en.wikipedia.org/wiki/Data_entry_clerk)**for business documents, e.g.**[**Cheque**](https://en.wikipedia.org/wiki/Check_clearing)**, passport, invoice, bank statement and receipt**
* [**Automatic number plate recognition**](https://en.wikipedia.org/wiki/Automatic_number_plate_recognition)**.**
* **In airports, for passport recognition and**[**information extraction**](https://en.wikipedia.org/wiki/Information_extraction)**.**
* **Automatic insurance documents key information extraction.**
* **Traffic sign recognition.**
* **Extracting business card information into a contact list.**
* **More quickly make textual versions of printed documents, e.g.**[**book scanning**](https://en.wikipedia.org/wiki/Book_scanning)**for**[**Project Gutenberg**](https://en.wikipedia.org/wiki/Project_Gutenberg)**.**
* **Make electronic images of printed documents searchable, e.g.**[**Google Books**](https://en.wikipedia.org/wiki/Google_Books)**.**
* **Converting handwriting in real-time to control a computer (**[**pen computing**](https://en.wikipedia.org/wiki/Pen_computing)**).**

**Conclusion:**

**OCR results depend on the input data quality. A clean segmentation of the text and no noise in the background gives better results. In the real world, this is not always possible, so we need to apply multiple pre-processing techniques for OCR to give better results.**

**FUTURE SCOPE:**

**Recently, a new generation of engineers is rebooting OCR in a way that would astonish Edmund Edward Fournier d’Albe. Built using artificial intelligence-based machine learning technologies, these new technologies aren’t limited by the rules-based character matching of existing OCR software. With machine learning, algorithms trained on a significant volume of data learn to think for themselves. Instead of being restricted to a fixed number of character sets, these new OCR programs will accumulate knowledge and learn to recognize any number of characters.**

**Amazon has also released a powerful OCR engine, Textract. Made available through Amazon Web Services in May of this year, the technology already has a reputation as being among the most accurate to date.**

**These readily-available technologies have certainly, vastly reduced the cost of building an OCR with enhanced quality. Still, they don’t necessarily solve the problems that most OCR users are looking to fix.**

**BIBILOGRAPHY**

**https://github.com/Guided-Projects/Smart-OCR-for-Document-Digitization**

**APPENDIX**