

**Mini Project Report
On
CREATION OF SPELLCHECKER FOR IMAGE
PROCESSING ON TESSERACT-OCR AND
GOOGLE-DOCS**

**VIth SEMESTER
INFORMATION TECHNOLOGY**

Submitted by

**Parineeta Warade
Shruti Bangre
Rahul Raipurkar**

**Under the guidance of
Prof. Vikas Bhowate**

Academic Year 2019-2020

Department of Information Technology



**ST. VINCENT PALLOTTI COLLEGE OF ENGINEERING AND
TECHNOLOGY**

Wardha Road, Gavsi Manapur, Nagpur

ST. VINCENT PALLOTTI COLLEGE OF ENGINEERING AND TECHNOLOGY, NAGPUR

DEPARTMENT OF INFORMATION TECHNOLOGY

CERTIFICATE

Certified that this project report “**CREATION OF SPELL CHECKER USING OCR**” is the bonafide work of “**Parineeta Warade, Shruti Bangre, Rahul Raipurkar**” who carried out the mini project work under my supervision in partial fulfillment of VI Semester, Bachelor of Engineering in **INFORMATION TECHNOLOGY** of RASHTRASANT TUKADOJI MAHARAJ NAGPUR UNIVERSITY, NAGPUR.

Prof. M. V. Bramhe
Associate Professor

HEAD OF THE DEPARTMENT

Prof. Vikas Bhowate
Assistant Professor

GUIDE

PRINCIPAL
ST. VINCENT PALLOTTI COLLEGE OF ENGINEERING AND TECHNOLOGY

ACKNOWLEDGEMENT

Our major project seminar is titled, “**CREATION OF SPELLCHECKER FOR IMAGE PROCEESING ON TESSERACT-OCR AND GOOGLE-DOCS**”. Any project seminar requires a lot of hard work, sincerity and systematic work methodologies. We express our deepest gratitude to our Project Guide, **Prof. Vikas Bhowate**, for giving us an opportunity to be a part of this project seminar and for the guidance, provided throughout the span of first phase of our project.

We would also like to thank **Prof. M. V. Bramhe, Head of the Department of Information Technology** and all our faculty members who regularly evaluated our first phase of project and pointed out the shortcomings in the project. They also gave us important feedback for the further improvement of our project. We are highly indebted to them.

We are also grateful to the industry mentor, **Dr. Soma Paul, Research Assistant Professor, IIIT Hyderabad, Hyderabad** and alumni mentor, **Dr. Vineet Chaitanya at Language Translation Research Center(LTRC), IIIT Hyderabad**, for their immense support throughout the project completion process. We are highly obliged to them.

We are also grateful to the **Management of the College, Dr. Surendra Gole, Principal and Prof. R. B. Gowardhan, Vice-Principal** for the overwhelming support in providing us the facilities of computer lab and other required infrastructure. We would like to thank our Library Department for providing us useful books related to our project.

**Parineeta Warade
Shruti Bangre
Rahul Raipurkar**

ABSTRACT

Optical character recognition or optical character reader (OCR) is the electronic or mechanical conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, or from subtitle text superimposed on an image (for example from a television broadcast).

Widely used as a form of data entry from printed paper data records – whether passport documents, invoices, bank statements, computerized receipts, business cards, mail, printouts of static-data, or any suitable documentation – it is a common method of digitizing printed texts so that they can be electronically edited, searched, stored more compactly, displayed on-line, and used in machine processes such as cognitive computing, machine translation, (extracted) text-to-speech, key data and text mining. OCR is a field of research in pattern recognition, artificial intelligence and computer vision.

Early versions needed to be trained with images of each character, and worked on one font at a time. Advanced systems capable of producing a high degree of recognition accuracy for most fonts are now common, and with support for a variety of digital image file format inputs.^[2] Some systems are capable of reproducing formatted output that closely approximates the original page including images, columns, and other non-textual components.`

Image processing is an absolutely crucial area of work for those groups and industries that are working in areas such as medical diagnostics, astronomy, geophysics, environmental science, data analysis in laboratories, industrial inspection, etc. Although not originally developed for users who are visually impaired, Optical Character Recognition (OCR) technology has become an aid for inputting documents quickly by and for users with vision impairments. A complete OCR system consists of a scanner, the recognition component, and OCR software that interacts with the other components to store the computerized document in the computer. The process of inputting the material into the computer begins with the scanner taking a picture of the printed material. Then, during the recognition process, the picture is analyzed for layout, fonts, text and graphics. Finally, the picture of the document is converted into an electronic format that can be edited with an application software.

List of Figures

Chapter No	Title	Page No
3.2.1	Gantt Chart	8
3.4.1.	Flowchart of the OCR system	11
3.4.2.	Data Flow Diagram	12
3.4.3.	Class Diagram of OCR System	13
3.4.4	Use Case Diagram of system	14
3.4.5.	Component Diagram of system	15
3.4.6.	Deployment Diagram of system	16
3.5.1.	System Architecture	18
4.1.1	Implementation and Coding	20

LIST OF CONTENTS

Chapter No.	Title	Page no.
	ACKNOWLEDGEMENT	
	ABSTRACT	
	LIST OF FIGURES	
1.	INTRODUCTION 1.1 OVERVIEW 1.2 PROBLEM STATEMENT 1.3 OBJECTIVES 1.4 ORGANIZATION OF REPORT	1
2.	REVIEW OF LITERATURE 2.1 INTRODUCTION 2.2 LITERATURE SURVEY 2.3 FEASIBILITY STUDY	3

3.	<p>PROPOSED SYSTEM</p> <p>3.1 DRAWBACK OF CURRENT SYSTEM AND NEED OF PROPOSED SYSTEM</p> <p>3.2 PROJECT PLANNING AND SCHEDULING</p> <p>3.3 SYSTEM DESCRIPTION AND SRS</p> <p>3.4 SYSTEM ANALYSIS</p> <p>3.4.1. FLOWCHART OF SYSTEM</p> <p>3.4.2. DATA FLOW DIAGRAM</p> <p>3.4.3. CLASS DIGRAM</p> <p>3.4.4. USE CASE DIAGRAM</p> <p>3.4.5. COMPONENET DIAGRAM</p> <p>3.4.6. DEPLOYMENT DIAGRAM</p> <p>3.4.7. HARDWARE REQUIREMENTS</p> <p>3.4.8. SOFTWARE REQUIREMENTS</p> <p>3.5 SYSTEM DESIGN</p> <p>3.5.1. SYSTEM ARCHITECTURE</p> <p>3.5.2. FRONT END DESIGN</p> <p>3.5.3. BACK END DESIGN IF ANY</p>	7
4.	IMPLEMENTATION AND CODING	20

5.	TESTING 5.1 UNIT TESTING 5.2 VALIDATION 5.3 INVALIDATION	26
6.	CONCLUSION & FUTURE SCOPE	27
7.	References	28

1.INTRODUCTION

1.1 OVERVIEW

Optical character recognition or optical character reader (OCR) is the electronic or mechanical conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo (for example the text on signs and billboards in a landscape photo) or from subtitle text superimposed on an image (for example from a television broadcast).

Widely used as a form of data entry from printed paper data records – whether passport documents, invoices, bank statements, computerized receipts, business cards, mail, printouts of static-data, or any suitable documentation – it is a common method of digitizing printed texts so that they can be electronically edited, searched, stored more compactly, displayed on-line, and used in machine processes such as cognitive computing, machine translation, (extracted) text-to-speech, key data and text mining. OCR is a field of research in pattern recognition, artificial intelligence and computer vision.

Early versions needed to be trained with images of each character, and worked on one font at a time. Advanced systems capable of producing a high degree of recognition accuracy for most fonts are now common, and with support for a variety of digital image file format inputs. Some systems are capable of reproducing formatted output that closely approximates the original page including images, columns, and other non-textual components of the system called OCR.

1.2 PROBLEM STATEMENT

To render the text of high resolution images (such as in .tiff format) using OCR capable software namely Tesseract software & Google Docs. This includes the work on Dictionary and rendering text of Bhagvatgeeta. This task includes a thorough knowledge of the software and programming language such as python. To convert an any image given by user into editable text using Tesseract-OCR and Google Docs.

1.3 OBJECTIVE

- The aim of the project is to extract all the text from the image in the work of English-Hindi Dictionary and Bhagvatgeeta using OCR software.
- Basic objective is to give user perfect output of the scanned image.
- To convert any scanned documents that is image in any language that is Hindi, Marathi, English etc into editable text, where user can read it correctly.

1.4 ORGANIZATION OF REPORT

In Chapter 1, we have the introduction that describes the overview of the Tesseract Tool, which includes the information about the methods that are to be developed for the rendering the text.

In Chapter 2, we have the review literature that discusses about what has been derived from Research Papers, online forums, and what previous work is done and need of the proposed system.

In Chapter 3, we have proposed system that describes the desired features and functions of the system. System analysis and system design has been explained in detail.

In Chapter 4, System Implementation and coding. It describes all the methods to run the system and extract the output and all the list of modules.

In Chapter 5, different types of testing performed are discussed including Validation testing, Invalidation Testing, etc.

In Chapter 6, Conclusion and Scope of the Project are explained. Procedure and steps to perform the execution are explained step by step. It also consist of the list of references for the project.

2. REVIEW OF LITERATURE

A Literature review is an evaluation report of information found in literature related to your selected area of study. The review should describe, summarize, evaluate and clarify this literature. It should give a theoretical base for the research and help you to determine the nature of your research.

When conducting research, a literature review is an essential part of the project because it covers all the previous research done on the topic and sets the platform on which the current research is based. No new research can be taken seriously without first reviewing the previous research done on the topic.

In our survey we found that the best way to debug the errors in the existing tool is to study the working of the TESSERACT software thoroughly and then developing a mechanism that could be implemented to all the images in one go to avoid more machine storage consumption.

2.1 INTRODUCTION

There is huge demand in “storing the information available on paper documents in to a computer storage disk and later use them.” Simple way to store the data is Scanning any paper document and store them as image, but its very difficult to reuse the information and difficult to read the individual content line by line. The reason behind this the font of characters on paper document which computer can not recognize while reading them.

Therefore, this concept of reading paper document called as document processing. But sometimes paper document maybe in some other language than English. For accessing other languages character recognition is used.

Our project is based on the Optical Character Recognition which is used to read the text basically from a high quality image. It includes various steps for performing this operation such as installing and running the software through linux terminal by commands.

2.2 LITERATURE SURVEY

Paper [1], titled “Anusaaraka : Machine Translation In Stages”, by Akshar Bharati, Vineet Chaitanya, Amba Kulkarni, explains the solutions the could be implemented. Different researchers have tried to give different answers to the problem. The most common approach has been to delimit the subject domain so that machine works in a narrow subject area, such as, weather reports aircraft maintenance manuals, computer manuals, etc. It has been hoped that by delimiting MT in a narrow area, one stands a better chance of using context, domain knowledge, etc. The system would perform badly when give a text outside the domain but that is a limitation one would have to live with. The real difficulty is in identifying a domain that is narrow enough that the system works well, and wide enough that enough real texts qualify to be in it, so that it is practically useful. The first module (called core anusaaraka) does language-based analysis: It takes all the information in the source text and presents it in its output, in an intermediate language which is quite close to the target language. The second module may do domain specific knowledge based processing, statistical processing, etc. in

which it may utilize world knowledge, frequency information, concordances, etc. to produce output in the target language.

Book[1], titled as “Natural Language Annotations for Machine Learning”, by Amber Stubbs and James Pustejovsky, Chapter 4, Building Your Model and Specification, guides on developing the modules smartly. Basically, the model is the practical representation of your goal: a description of your task that defines the classifications and terms that are relevant to project.

Book[2], titled as, “Modern Language Models and Computation Theory with Application, by Alexander Meduna and Ondrej Soukup, in Chapter 2, Modern Grammars, presents an overview of major modern types of grammars together with the corresponding computational modes formalized by them. It covers the most important grammars for regulated computations.

It discusses about the grammatical models for computation in parallel. Accordingly, these grammars generate their languages in parallel and thereby, accelerate the generation process enormously just like computation in parallel is much more faster than that made in an ordinary sequential way.

2.3 FEASIBILITY STUDY

A feasibility study is high level capsule version of the entire system analysis and design process. The study begins by classifying the problem, definition. Feasibility is to determine if its worth doing. Once an acceptance problem definition has been generated, the analyst develops a logical model of the system.

A feasibility study aims to objectively and rationally uncover the strengths and weaknesses of an existing business or proposed venture, opportunities and threats present in the environment, the resources required to carry through and ultimately the prospects for success.

A well designed feasibility study should provide a historical background of the business or the project, a description of the product or service, accounting statements, details of the operation and management, marketing research policies, financial data, legal requirements and tax obligations. It focuses on these major questions:

1. What are the user’s demonstrable needs and how does a candidate system meet them?
2. What resources are available for the given candidate system?
3. What are the likely impacts of the candidate system on the organization?
4. Whether it is worth to solve the problem?

During feasibility analysis for this project, the following primary areas of interest are to be considered. Investigation and generating ideas about a new system does this. □ Technical Feasibility □ Economical Feasibility □ Operational Feasibility □ Schedule Feasibility

➤ TECHNICAL FEASIBILITY :

A number of issues have to be considered while doing a technical analysis. Understanding different technology involved in the proposed system before connecting the project we have to be very clear about what are the technologies that are to be required for the development of the new system. Find out whether the organization currently passes the required technologies.

A technical feasibility study assesses the details of how you intend to deliver a product or service to customers. Think materials, labor, transportation, where your business will be located,

and the technology that will be necessary to bring all this together. The Anusaaraka language modules will be delivered as independent set of heuristic rules for separate data dictionaries.

➤ **ECONOMIC FEASIBILITY :**

Economically feasibility attempts to weigh the costs of developing and implementing a new system against the benefits that would accrue from having the new system in place. This study gives the top management the economics justification for the new system. A simple economics analysis which gives the actual comparison of costs and benefits are much more meaningful in the case. This proves to be useful point of reference to compare actual costs the project progress. It will include customer satisfaction, improvement in product quality better decision making timeliness of information better documentation and faster retrieve of information as well as record keeping.

Economic feasibility is a kind of cost-benefit analysis of the examined project, which assesses whether it is possible to implement it. ... It consists of market analysis, economic analysis, technical and strategic analysis. The tool is economical when considered through financial perspective. It requires as minimum as possible investment throughout the development cycle.

➤ **OPERATIONAL FEASIBILITY**

Project should be perform its operation in proper way therefore there are questions which will help test the operational feasibility of project:

1. Is there sufficient support for the project from management from users?
2. Are the current business method acceptable to the user?
3. Have the user been involved in the planning and deployment of the project?

Operational feasibility is the measure of how well a proposed system solves the problems, and takes advantage of the opportunities identified during scope definition and how it satisfies the requirements identified in the requirements analysis phase of system development.

➤ **SCHEDULE FEASIBILITY**

Schedule feasibility: The process of assessing the degree to which the potential time frame and completion dates for all major activities within a project meet organizational deadlines and constraints for affecting change. And then helps the schedule feasibility study.

3. PROPOSED SYSTEM

3.1 DRAWBACKS OF CURRENT SYSTEM AND NEED OF PROPOSED SYSTEM

Drawbacks:

Earlier OCR systems is that they only have the capability to convert document in English language specific. Current system has recovers the above drawback but required high quality resolution images to convert it into specific language. This system does not recognize the basic problems of language that is guessing the word and does not provide missing information of word to complete the sentence.

OCR (optical character recognition) is the use of technology to distinguish printed or handwritten text characters inside digital images of physical documents, such as a scanned paper document. The basic process of OCR involves examining the text of a document and translating the characters into code that can be used for data processing. OCR is sometimes also referred to as text recognition.

The process of OCR is most commonly used to turn hard copy legal or historic documents into PDFs. Once placed in this soft copy, users can edit, format and search the document as if it was created with a word processor.

Benefits of optical character recognition :

The main advantages of OCR technology are saved time, decreased errors and minimized effort. It also enables actions that are not capable with physical copies such as compressing into ZIP files, highlighting keywords, incorporating into a website and attaching to an email. While taking images of documents enables them to be digitally archived, OCR provides the added functionality of being able to edit and search those documents.

3.2 PROJECT PLANNING AND SCHEDULING

3.2.1 PROJECT PLANNING

Project Planning is a part of project management, which relates to the use of schedules such as Gantt chart to plan and subsequently report progress within the project environment. Initially, the project scope is defined and the appropriate methods for completing the project are determined. Following this step, the durations for various tasks necessary to complete the work are listed and grouped into a work breakdown structure. Project planning is often used to organize different areas of a project, including project plans, workload and the management of teams and individuals. The logical dependency between tasks are defined using an activity network diagram that enables identification of the critical path. Project planning is inherently uncertain as it must be done before the project is actually started.

At this stage, the project schedule may be optimized to achieve the appropriate balance between resource usage and project duration to comply with the project objectives. Once established and agreed, the project schedule becomes what is known as the baseline schedule. Progress will be measured against the baseline schedule throughout the life of the project.

Sr.No.				Dec	Dec	Dec	Dec	Jan	Jan	Jan	Jan	Feb	Feb	Feb	Feb	Mar
		start	end	W1	W2	W3	W4	W1	W2	W3	W4	W1	W2	W3	W4	W1
1.	Domain and topic discussion															
2.	Project Finalized															
3.	Literature survey															
4.	Designing and architecture															
5.	Obtaining required software															
6.	Poat process data analyzed															
7.	Producing accurate output															
8.	Rechecking accuracy															
9.	Testing															
10.	Documentation															

Fig 3.2.1 Gantt Chart

3.3 SYSTEM DESCRIPTION AND SOFTWARE REQUIREMENT SPECIFICATION

A Software requirements specifications (SRS) is a description of a software system to be developed. It lays out functional and non-functional requirements, and may include a set of use cases that describes user interaction that the software must provide. The SRS document is prepared as per the SRS IEEE (Institute of Electrical and Electronics Engineers) template. This document is followed by three sections viz. Project Purpose, Scope of Project, Project Planning.

3.3.1 PURPOSE

The main purpose of OCR system based on grid infrastructure is to perform Document Image Analysis, Document Processing of electronic formats converted from paper formats more effectively. It increases accuracy of recognizing characters. The primary objective is speed up the

process of character recognition in document processing. As a result system can process huge number of documents within less time and hence it saves time.

It aims to recognize multiple heterogeneous characters that belongs to different universal language with different font properties and alignments. The purpose is to convert the images of different types (Printed, Handwritten etc.) into machine encoded text of dictionaries by Neural Networks and to debug the issues that occur during execution. The OCR technology is used at an industrial level. For example, banks can employ this technology to scan the bank statements, the travel industry can use this technology to scan passport and invoices, etc.

3.3.2 SCOPE OF PROJECT

Scope of our project which is bases on OCR is to provide efficient and enhanced software tools for user to perform Document image Analysis, document processing by reading and recognizing the characters in research, academic, governmental organizations, etc. that are having large documented scanned images. This can solve various documentation related problem. Irrespective size of documents and the type of characters in documents, the product is recognizing them, searching them and processing them faster according to the need of the environment.

3.4 SYSTEM ANALYSIS

The system analysis requirements gathering process is intensified and focused specifically on software. To understand the nature of the program(s) to be built, the software engineer ("analyst") must understand the information domain for the software, as well as required function, behavior, performance, and interface. Requirements for both the system and the software are documented and reviewed with the customer. The major objectives of system analysis are to find answers for each business process; what is being done, how it is being done, who is doing it, when is he doing it, why is it being done and how can it be improved? It is more of thinking process and involves the creative skills of the system analyst. It attempts to give birth to a new efficient system that satisfies the current needs of the user and the scope for future growth within the organizational constraints. The result of this process is the logical system design. System analysis is the iterative process that continues until a preferred and acceptable solution emerges.

Requirement analysis also provide software designer with a representation of information, function, and behavior that can be translated to data, architectural, interface, and component-level designs. Finally, the requirements specification provides the developer and the customer with the means to assess quality once software is built.

Software requirements analysis may be divided into five areas of effort: (1) Problem recognition, (2) evaluation and synthesis, (3) modeling, (4) specification, and (5) review. Initially, the analyst studies the system specification (if one exists) and the software project plan.

3.4.1 FLOWCHART OF THE SYSTEM

A flowchart is a type of diagram that represents an algorithm, workflow or process, showing the steps as boxes of various kinds, and their order by connecting them with arrows.

Step1: It will take input as an image of document.

Step 2: It will process the input that is scanned image.

- Step 3: Main function of OCR that is conversion will takes place using Leptonica And Trained Data Set.
- Step 4: It will perform operations after input that is post processing.
- Step 5: It will test the post data and give correct output.

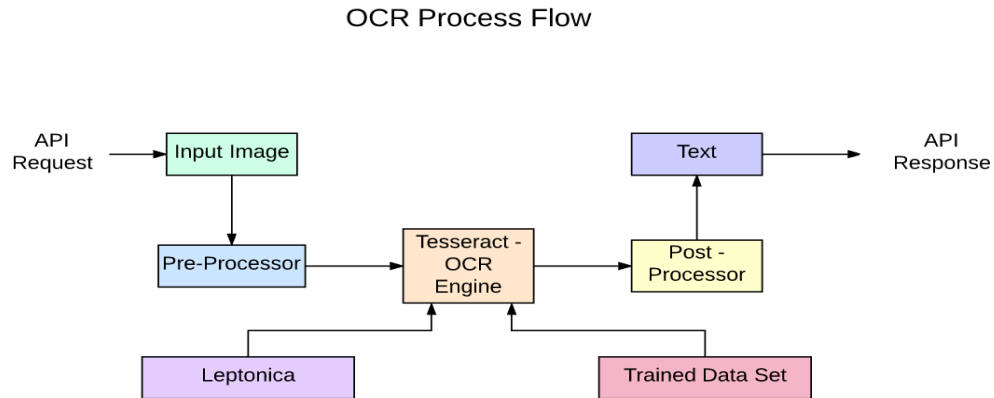


Fig 3.4.1. Flowchart of the OCR System

3.4.2 DATA FLOW DIAGRAM

A data flow diagram is a graphical representation of the ‘flow’ of data through an information system, modeling its process aspects. A DFD shows what kinds of information will be input to and output from the system, where the data will come from and go to, and where the data will be stored.

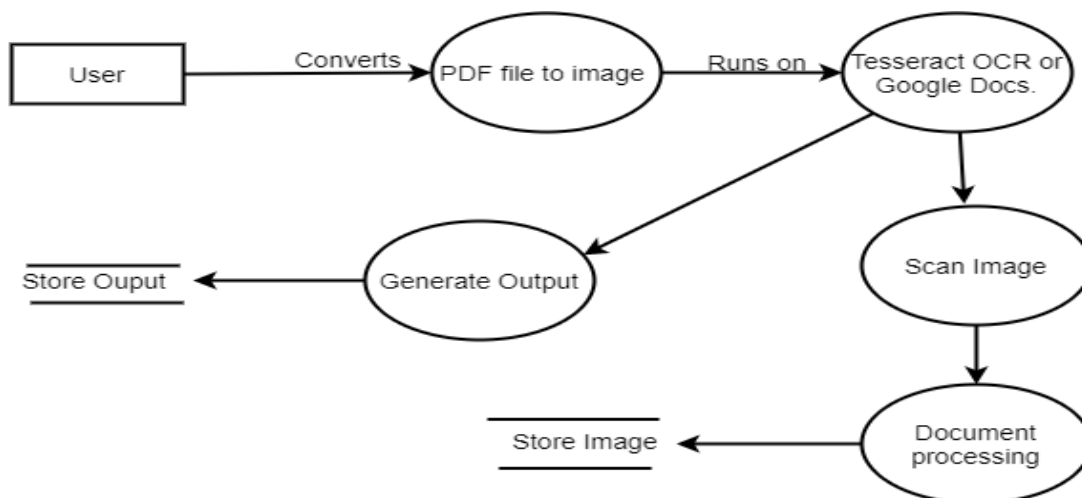


Fig 3.4.2. Data flow Diagram

The above DFD shows flow of the data that is the processes involve in the system. It is also known as context diagram. It's designed to be an abstraction view, showing the system as a single process with its relationship to external entities. It represent the entire system as single bubble with input and output data indicated by incoming/outgoing arrows.

3.4.3 CLASS DIAGRAM

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among objects.

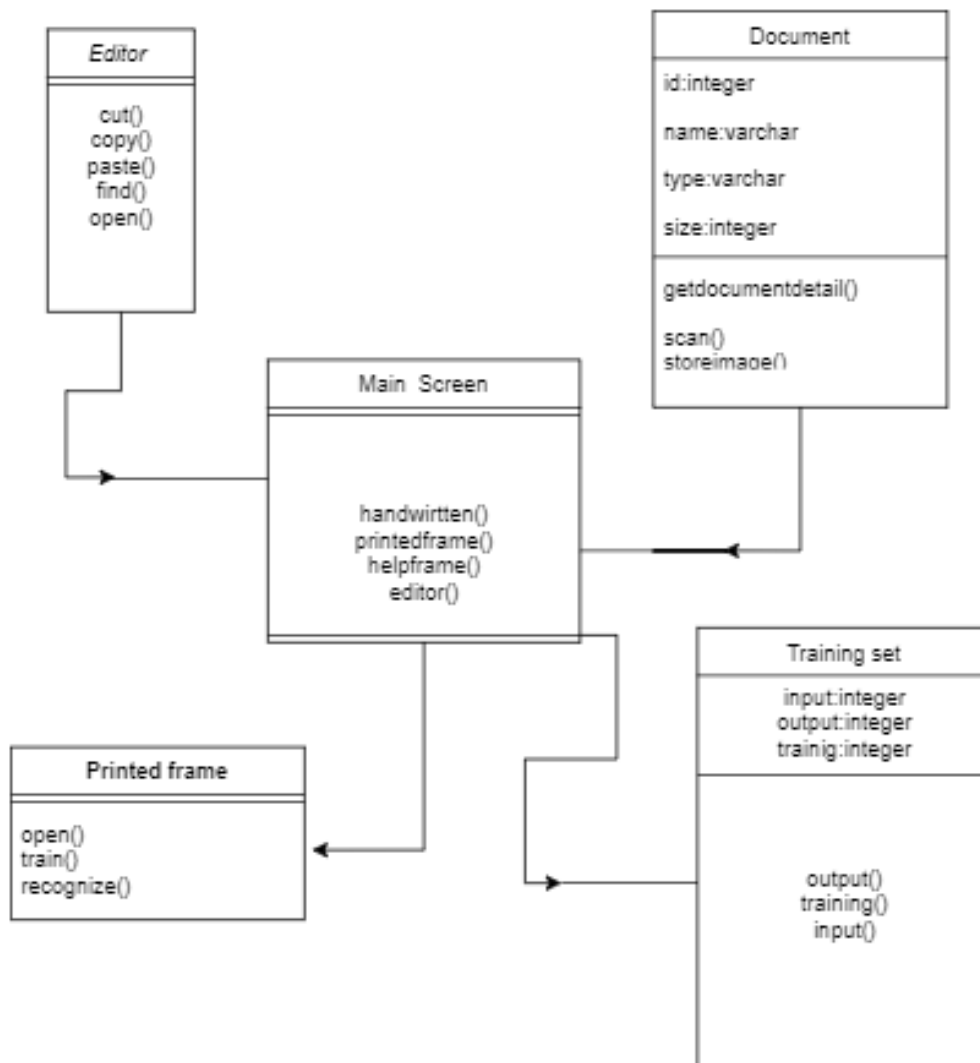


Fig 3.4.3. Class Diagram of OCR process

3.4.4 USE CASE DIAGRAM

A use case diagram at its simplest is a representation of a user's interaction with the system that shows the relationship between the user and the different use cases in which the user is involved.

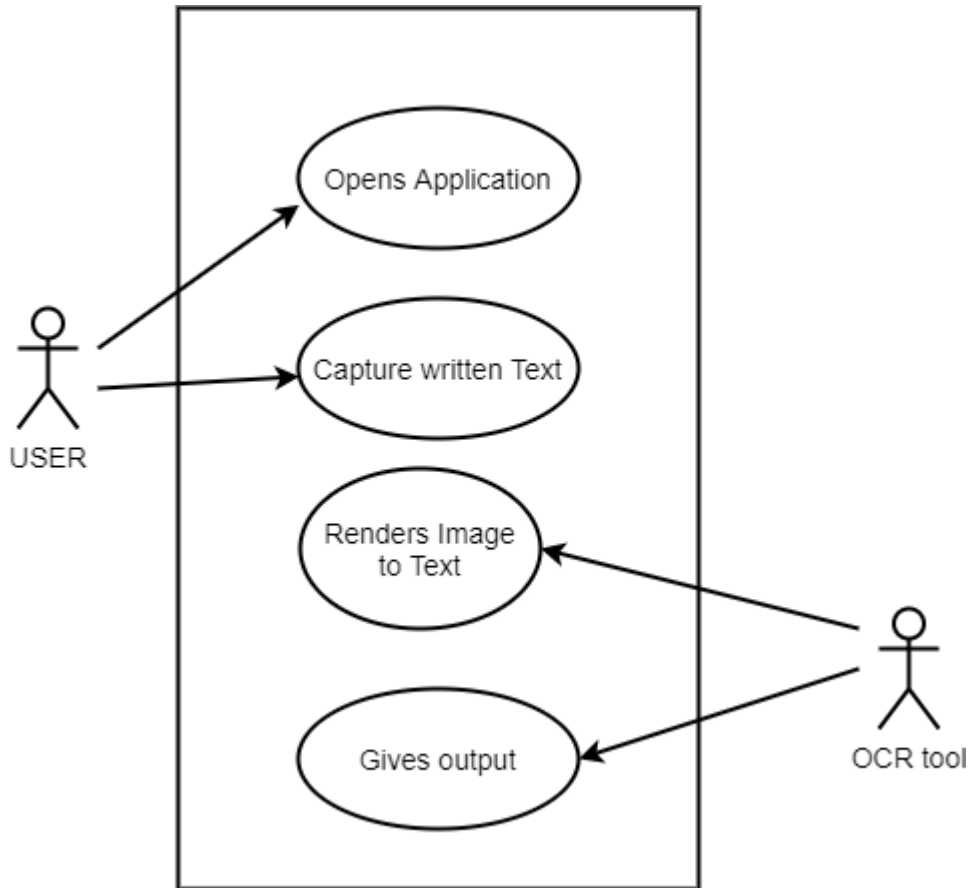


Fig 3.4.4. Use case Diagram

The use case diagram clearly explains the relationship between the user and OCR tool (In this case we are using Google Docs and Tesseract OCR Engine which is also developed by Google).

3.4.5 COMPONENT DIAGRAM

A component diagram, also known as a UML component diagram, describes the organization and wiring of the physical components in a system. Component diagrams are often drawn to help model implementation details and double-check that every aspect of the system's required functions is covered by planned development.

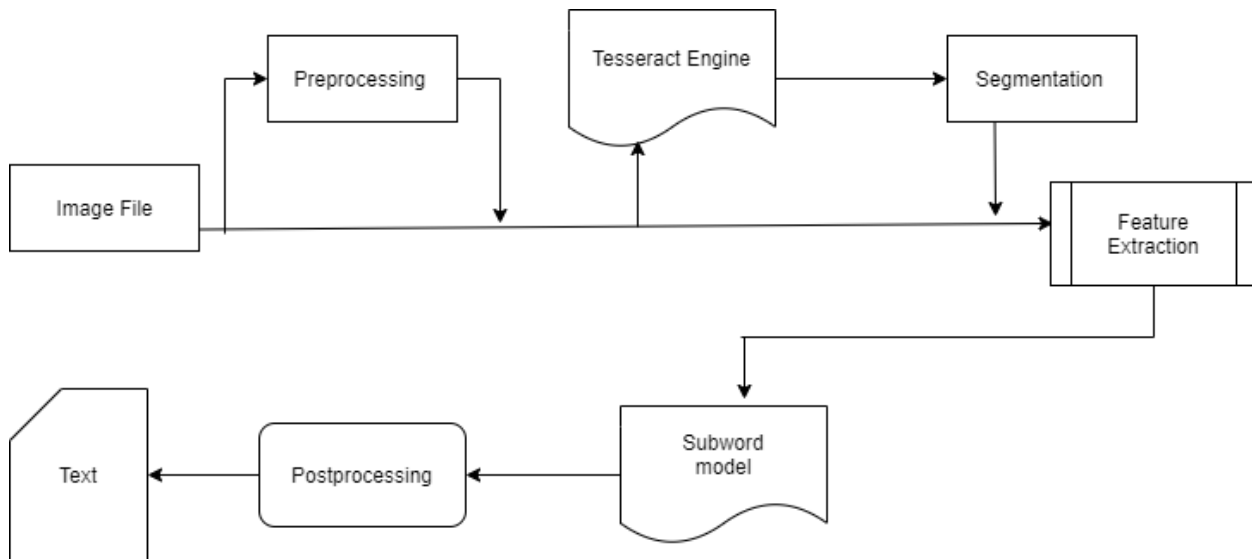


Fig.3.4.5 Component Diagram

3.4.6 DEPLOYMENT DIAGRAM:

A deployment diagram serves to model the physical deployment of artifacts on deployment targets. Deployment diagrams show “The allocation of Artifacts to nodes according to the deployment defined between them”

In the diagram of OCR system, the sever role is played by Admin called Librarian. There can be N number of clients who can access the digital library data content at a time. The clients here ,ay either the students or faculty.

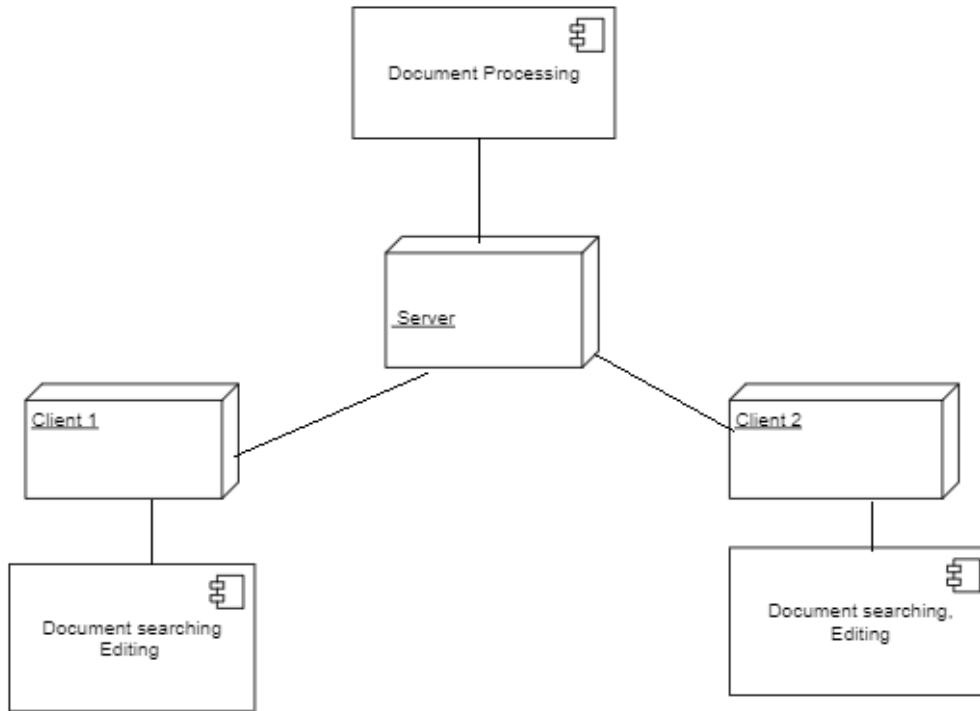


Fig.3.4.6 Deployment diagram of OCR

Actions performed by the administrator are document processing, searching and editing whereas the actions performed by the end user are only documents searching and editing.

3.4.7 HARDWARE REQUIREMENT

- Laptop – at least 4GB RAM,
- Quadcore Processor,
- CPU core i3 and above
- Storage 256GB SSD and Hard Disk – 20GB for alignment composition
- Processor P4 and above

3.4.8 SOFTWARE REQUIREMENT

- Linux 16.01 or higher versions.
- Tesseract OCR Engine should be installed in Operating Systems.
- Python (2.7 or above) Interpreter to be installed.
- Test Data tool to be installed from Github
- Online PDF to Image Convertor
- GITHUB

3.5 SYSTEM DESIGN

3.5.1 SYSTEM ARCHITECTURE:

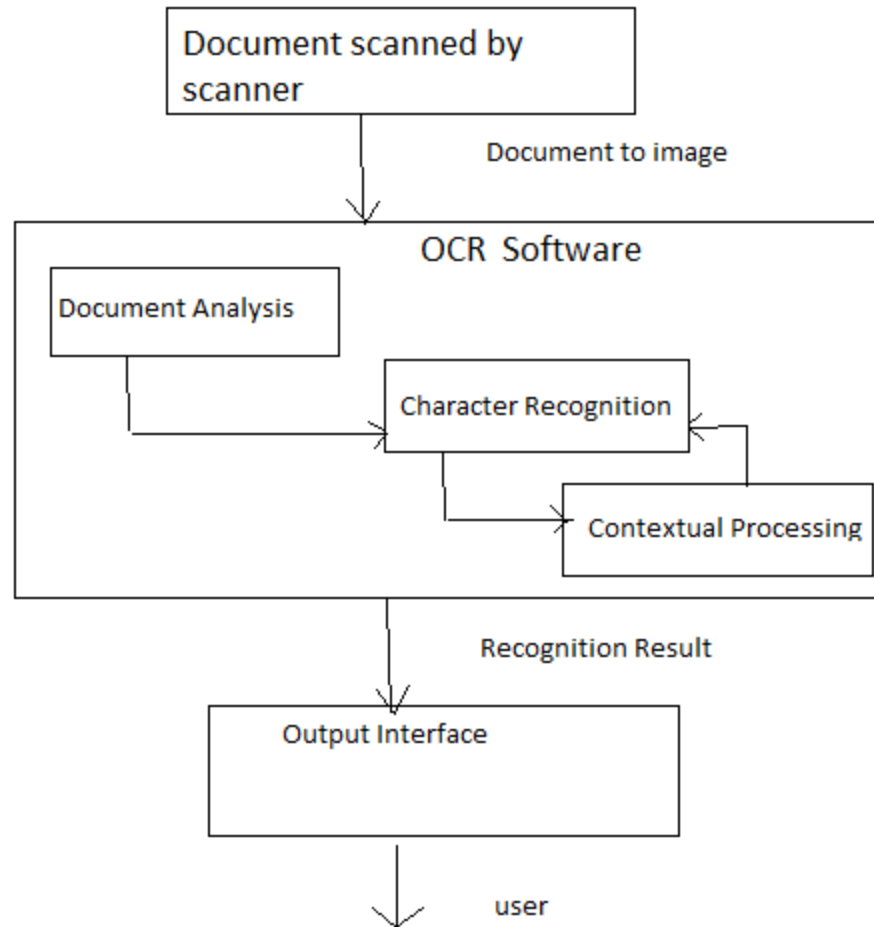


Fig.3.5.1. System Architecture

1. Accepting input form the user, which converts documents into images by scanner or any scanning machine available. The next step is storing the image of the document in .jpg,.tiff format only as shown in figure.
2. In the next stage where OCR software tool is used for conversion of document into recognition of text. Next step will first make analysis of document . Character Recognition will recognize the character and pass to contextual processing.
3. Third and last step is after creating a recognizable code it will give an output to the user without fail.

3.5.2 FRONT END DESIGN

The design of the application is highly important in order to fulfill the requirements and functionalities of the project. The design of the tesseract is quiet simple.

It is based on command line interface which is used for processing the data.

The Tesseract OCR engine can be access through the terminal in Linux System (Ubuntu) which is same as that of command prompt in windows operating system.

Files can be modified in the Tesseract Engine by using various terminal commands.

The output is shown in the terminal itself .

Google Docs. , which is also used in recognizing the characters is based on graphical user interface. The work can be done in online mode.

The tesseract OCR consist of various packages which is to be link with libtesseract library. Basically this also includes the python 2.7 or above versions to implement basic commands or installing the packages in the system. The package links to the libtesseract C++ library and works out of the box on Windows and Mac without installing any third party software.

4. IMPLEMENTATION AND CODING

IMPLEMENTATION AND CODING:

The package links to the libtesseract C++ library and works out of the box on Windows and Mac without installing any third party software.

```
install.packages("tesseract")
```

On Linux you first need to install libtesseract (readme) which ships with every popular distribution (Debian, Ubuntu, Fedora, CentOS, etc). The package itself is very simple. The ocr function takes a URL or path or raw vector with image data. On most platforms the image should either be in png or jpeg or tiff format.

```
library(tesseract)
text <- ocr("http://jeroenooms.github.io/images/testocr.png")
cat(text)
```

The OCR method used by tesseract uses language specific training data to optimize character recognition. The default language is English, training data for other languages are provided via the official tessdata repository directory. On Linux these can be installed directly with the yum or apt package manager.

```
# Low quality:
text1 <- ocr("http://jeroenooms.github.io/files/dog_lq.png")
cat(text1)
```

```
# High quality:
text2 <- ocr("http://jeroenooms.github.io/files/dog_hq.png")
cat(text2)
```

- **Steps for Implementation:**

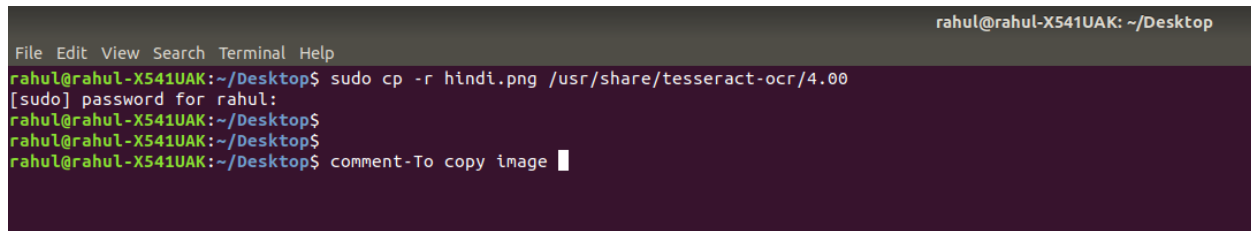
Steps1: First of all the dictionary (i.e. English to Hindi Dictionary) is to be converted from PDF to Image format using an online convertor tool.

Step2: Open the Tesseract OCR engine. It will take input as an image of document.

Step 3: It will process the input that is scanned image.

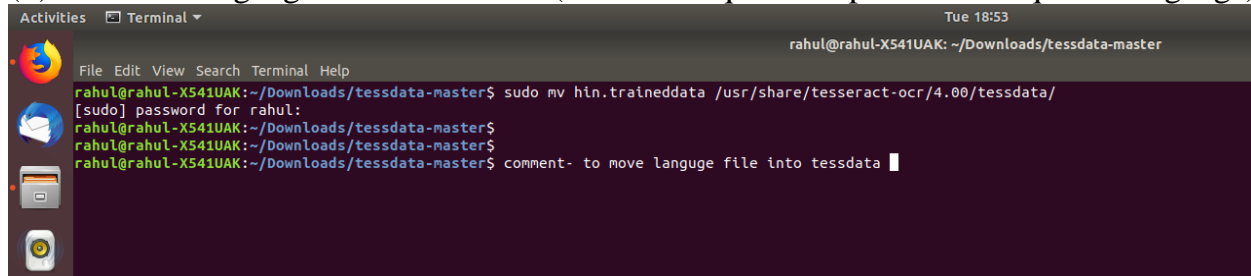
Step 4: Following commands are used for processing the the image and generating the output:

(i)First copy image file to the root folder.



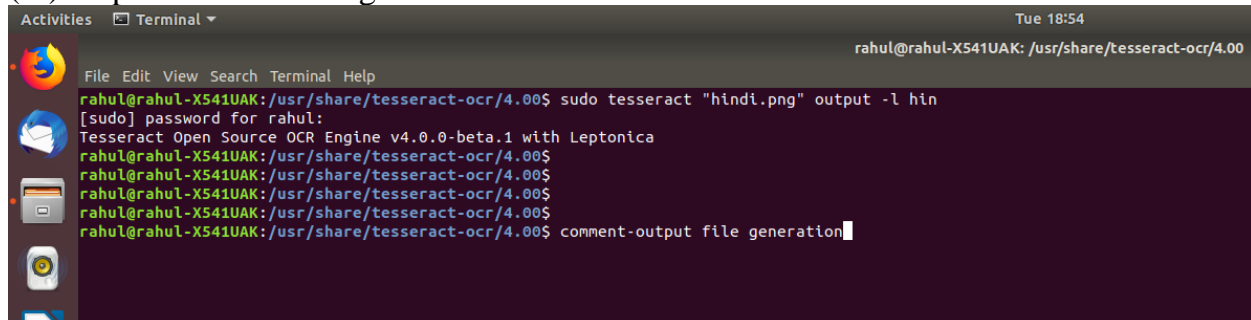
```
rahul@rahul-X541UAK: ~/Desktop
File Edit View Search Terminal Help
rahul@rahul-X541UAK:~/Desktop$ sudo cp -r hindi.png /usr/share/tesseract-ocr/4.00
[sudo] password for rahul:
rahul@rahul-X541UAK:~/Desktop$
rahul@rahul-X541UAK:~/Desktop$
rahul@rahul-X541UAK:~/Desktop$ comment-To copy image
```

(ii)Move the language file into tessdata(which is required to process in required language)



```
Activities Terminal
Tue 18:53
rahul@rahul-X541UAK: ~/Downloads/tessdata-master
File Edit View Search Terminal Help
rahul@rahul-X541UAK:~/Downloads/tessdata-master$ sudo mv hin.traineddata /usr/share/tesseract-ocr/4.00/tessdata/
[sudo] password for rahul:
rahul@rahul-X541UAK:~/Downloads/tessdata-master$
rahul@rahul-X541UAK:~/Downloads/tessdata-master$
rahul@rahul-X541UAK:~/Downloads/tessdata-master$ comment- to move language file into tessdata
```

(iii)Output Code for File generation.



```
Activities Terminal
Tue 18:54
rahul@rahul-X541UAK: /usr/share/tesseract-ocr/4.00
File Edit View Search Terminal Help
rahul@rahul-X541UAK:/usr/share/tesseract-ocr/4.00$ sudo tesseract "hindi.png" output -l hin
[sudo] password for rahul:
Tesseract Open Source OCR Engine v4.0.0-beta.1 with Leptonica
rahul@rahul-X541UAK:/usr/share/tesseract-ocr/4.00$
rahul@rahul-X541UAK:/usr/share/tesseract-ocr/4.00$
rahul@rahul-X541UAK:/usr/share/tesseract-ocr/4.00$
rahul@rahul-X541UAK:/usr/share/tesseract-ocr/4.00$
rahul@rahul-X541UAK:/usr/share/tesseract-ocr/4.00$ comment-output file generation
```

- Step 3: Main function of OCR that is conversion will takes place using Leptonica and Trained Data Set.
- Step 4: It will perform operations after input that is post processing.
- Step 5: It will test the post data and give correct output.

INPUT IMAGE (In TESSERACT):

ॐ श्रीपरमात्मने नमः

भूमिका

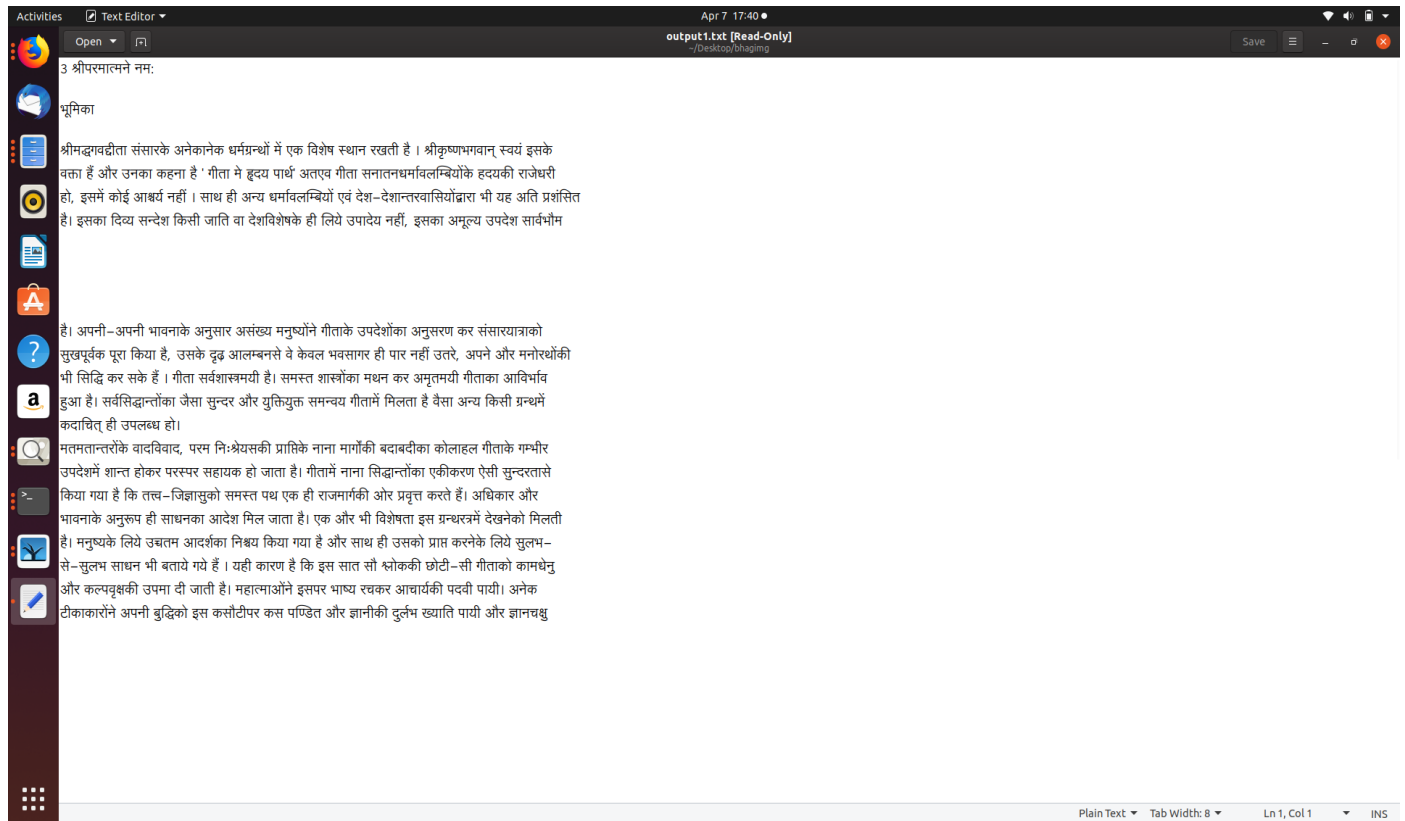
श्रीमद्भगवद्गीता संसारके अनेकानेक धर्मग्रन्थोंमें एक विशेष स्थान रखती है। श्रीकृष्णभगवान् स्वयं इसके वक्ता हैं और उनका कहना है 'गीता मे हृदयं पार्थ' अतएव गीता सनातनधर्मावलम्बियोंके हृदयकी राजेश्वरी हो, इसमें कोई आश्चर्य नहीं। साथ ही अन्य धर्मावलम्बियों एवं देश-देशान्तरवासियोंद्वारा भी यह अति प्रशंसित है। इसका दिव्य सन्देश किसी जाति वा देशविशेषके ही लिये उपादेय नहीं, इसका अमूल्य उपदेश सार्वभौम है। अपनी-अपनी भावनाके अनुसार असंख्य मनुष्योंने गीताके उपदेशोंका अनुसरण कर संसारयात्राको सुखपूर्वक पूरा किया है, उसके दृढ़ आलम्बनसे वे केवल भवसागर ही पार नहीं उतरे, अपने और मनोरथोंकी भी सिद्धि कर सके हैं। गीता सर्वशास्त्रमयी है। समस्त शास्त्रोंका मथन कर अमृतमयी गीताका आविर्भाव हुआ है। सर्वसिद्धान्तोंका जैसा सुन्दर और युक्तियुक्त समन्वय गीतामें मिलता है वैसा अन्य किसी ग्रन्थमें कदाचित् ही उपलब्ध हो।

मतमतान्तरोंके वादविवाद, परम निःश्रेयसकी प्राप्तिके नाना मार्गोंकी बदाबदीका कोलाहल गीताके गम्भीर उपदेशमें शान्त होकर परस्पर सहायक हो जाता है। गीतामें नाना सिद्धान्तोंका एकीकरण ऐसी सुन्दरतासे किया गया है कि तत्त्व-जिज्ञासुको समस्त पथ एक ही राजमार्गकी ओर प्रवृत्त करते हैं। अधिकार और भावनाके अनुरूप ही साधनका आदेश मिल जाता है। एक और भी विशेषता इस ग्रन्थरत्नमें देखनेको मिलती है। मनुष्यके लिये उच्चतम आदर्शका निश्चय किया गया है और साथ ही उसको प्राप्त करनेके लिये सुलभ-से-सुलभ साधन भी बताये गये हैं। यही कारण है कि इस सात सौ श्लोककी छोटी-सी गीताको कामधेनु और कल्पवृक्षकी उपमा दी जाती है। महात्माओंने इसपर भाष्य रचकर आचार्यकी पदवी पायी। अनेक टीकाकारोंने अपनी बुद्धिको इस कसौटीपर कस पण्डित और ज्ञानीकी दुर्लभ ख्याति पायी और ज्ञानचक्षु प्रदानकर इसके तत्त्वानुसन्धानमें साधारण गतिके लोगोंको इसका मर्म हृदयङ्गम करनेमें सहायता प्रदान की। विद्याका परमलाभ गीताके रहस्यको समझना ही माना गया है।

आचार्योंने अपने-अपने सिद्धान्तोंकी प्रामाणिकता स्थापन करनेमें गीताको एक मुख्य आधार माना है। गीतापर भाष्य रच अपने सिद्धान्तोंको गीता-सम्मत बताना ही उनका लक्ष्य रहा है। गीता-विरोधी किसी धर्म वा सम्प्रदायका प्रचार वे असम्भव समझते और जिस धर्म, आचार वा सिद्धान्तको ब्रह्मरूपा गीतासे सिद्ध कर दिया, वह अवश्य ही सर्वशास्त्र और वेद-सम्मत मान लिया जाता है।

सम्प्रदाय, जाति और देशकी भिन्नताका निराकरण करनेवाला गीता एक सार्वभौम सिद्धान्तप्रतिपादक ग्रन्थ-रत्न है। उसके उपदेश और निर्दिष्ट साधनोंने मानव-जातिके लिये एक महान् धर्मकी नींव डाली है, उसके प्रचारसे प्राणिमात्रका कल्याण सम्भव है। हृदय-दौर्बल्यपर विजयी होकर गीताके उपदेशसे मनुष्य कर्मरत हो सकता है। वह भक्तिरसामृतका आस्वादन करता हुआ ज्ञानी बन सकता है। ऐहिक और पारमार्थिक दोनों ही सुखोंकी प्राप्ति उसे अल्प प्रयाससे ही उपलब्ध होनेमें कोई सन्देह नहीं रहता। आधुनिक कालमें जो अनेकानेक जटिल प्रश्न नित्यप्रति समाज और व्यक्तिके समक्ष उपस्थित होते रहते हैं और बुद्धिको चकरा

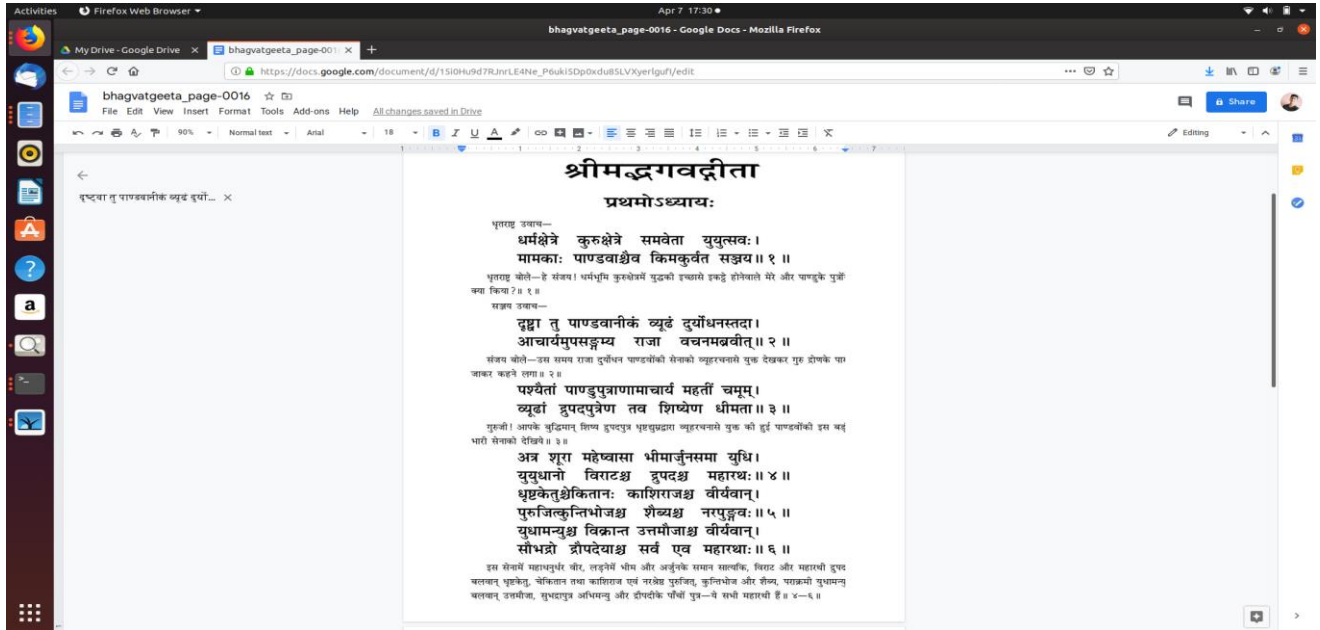
OUTPUT:



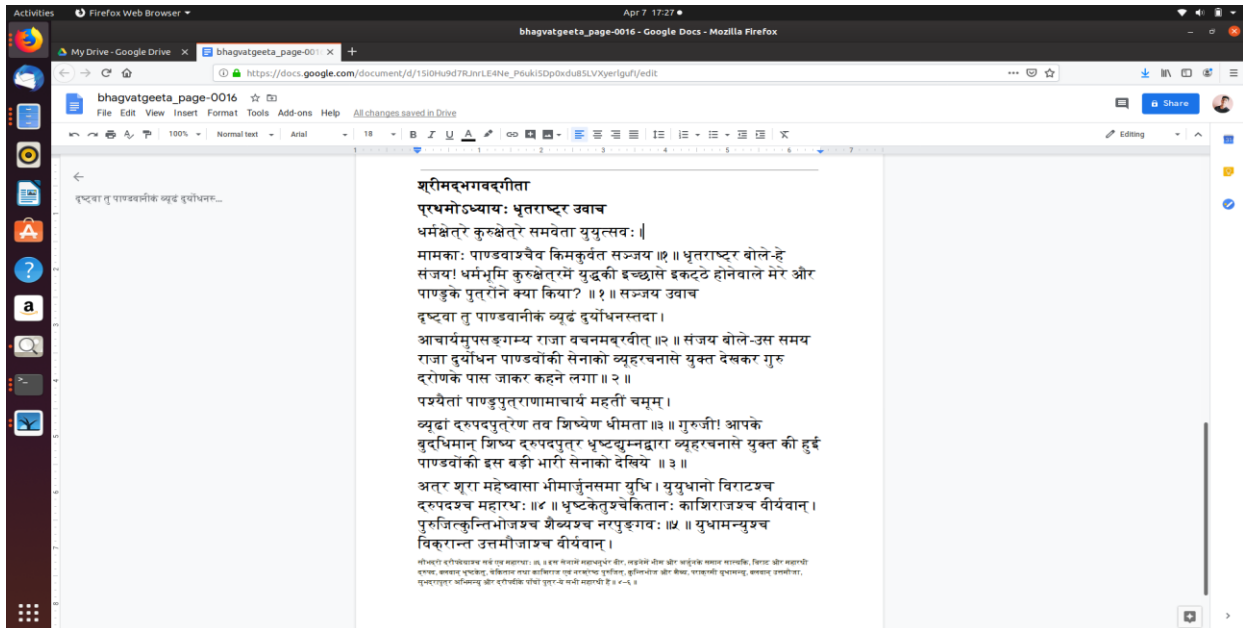
IN GOOGLE DOCS. , simply follow the below steps:

1. Convert the pdf to image file.
2. Open Google Drive , and upload the image on it.
3. And simply right click the image and select option “open with Google docs.”
- 4.The image will open along with its output.

Input Image:



Output Screen:



5. TESTING

5.1 UNIT TESTING

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and testing to be conducted as two distinct phases.

TEST OBJECTIVES:

- All field entries was work properly
- Pages must be activated from the identified link
- The entry screen, messages and responses must not be delayed.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

5.2 VALIDATION TESTING

Validation testing is the process of ensuring if the tested and developed software satisfies the client /user needs. The business requirement logic or scenarios have to be tested in detail. One such method that helps in detail evaluation of the functionalities is the Validation Process. When software is tested, the motive is to check the quality regarding the found defects and bugs. When defects and bugs are detected, developers fix them. After that, the software is checked again to make sure no bugs are left. In that way, the software product's quality scales up.

The aim of software testing is to measure the quality of software in terms of a number of defects found in it, the number of tests run and the system covered by the tests. When bugs or defects are found with the help of testing, the bugs are logged and the development team fixes them. Once the bugs are fixed, testing is carried out again to ensure that they are indeed fixed and no new defects have been introduced in the software. With the entire cycle, the quality of the software increases.

Test Result: Validation Testing is performed successfully in our project. No defects encountered.

5.3 INVALIDATON

As validation testing takes place invalidation plays an important role in it. If any user scanning any document and storing with .pdf file and try to convert with OCR software tool then there is no way to do this. This performs invalidation. It will not allowed user to take an any invalid input to convert into text. For this system invalidation performs role of checking input type while conversion of document into images.

Test Results: Invalidation testing is done successfully during the time period of our project.

No defects encountered

6. CONCLUSION AND FUTURE SCOPE

An Optical Character Recognition system could be developed by considering the multiple font style in use. Our approach is very much useful for the font independent case. Because, for font or character size, it finds the string and the strings are parsed to recognize the character. Once character is identified, the corresponding character could be ejected through an efficient editor. Efforts have been taken to develop a compatible editor for Tamil and English.

Except Bangla and Hindi, all other Indian languages require development of an OCR for printed characters, and for handwritten characters, OCR has to be developed for all languages (including Bangla and Hindi). Of course, OCR for printed characters are easy when compared to the handwritten cursive scripts. Even for the printed document recognition, an OCR should be able to perform the all features besides character recognition, such as spell check, sentence and grammar check, Also an editor with key board encoding and font encoding is required.

With this approach the printed and handwritten characters are recognizable easily with less effort and more accuracy. A module for Skew correction and line separation, word and character separation along with an editor with spell checker and grammar checker could be designed for 'developing a complete OCR. Further, with a little fine tuning on the modules, such as, skew correction and line separation, word and character separation, a complete OCR could be designed for handwritten scripts of any language for that matter. It is proposed to apply the approach to all manuscripts recognition of South-Indian languages. Since some of the characters in some of the languages have similar characters viz, Tamil and Malayalam have similar features among few characters, and Telugu and Kannada have similarity among most of the characters, our approach could be applied for these languages and could be extended to all other languages.

7. REFERENCES

[1] MACHINE TRANSLATION IN STAGES :

By Dr. Vineet Chaitanya, P.Kulkarni Rajeev Sangal Satyam, from School of Applied Information Systems , IITH.

[2] A view of Artificial Neural Network

By Manish Mishra ,at 2014 International Conference on advances in Engineering Technology Research (ICAETR - 2014) .

[3] OCR.Space: [The OCR Software Blog](#)

[4] ABBYY: [What is OCR and OCR Technology](#)

[5] <http://sampark.iiit.ac.in/anusaaraka/>

ANNEXURE

STAKE HOLDER'S DETAILS

Project Title: Creation of Spell Checker Using Optical Character Recognition(OCR)

Project Type: Industry Based

SR. NO	STAKEHOLDERS	NAME	COMPANY	DESIGNATION	EMAIL-ID
1.	Industry mentor	Dr. Soma Paul	IIIT-H	PROFESSOR	soma@iiit.ac.in
2.	Project Guide	Prof. Vikas Bhowate	SVPCET	ASST. PROFESSOR	vbhowate@stvincent.edu.in

PROJECT MEMBER DETAILS

Sr. No.	NAME	ROLL NO.	MOBILE NO.	EMAIL ID
1.	Parineeta Warade	20	9011365129	parineetawarade@gmail.com
2.	Shruti Bangre	29	8605925710	shrutibangre22@gmail.com
3.	Rahul Raipurkar	53	9075250275	rahulraipurkar735@gmail.com