**DOCUMENT ANALYZER**



The Project submitted to

Sant Gadgebaba Amravati University, Amravati

towards partial fulfillment of the Degree of

Bachelor of Engineering

in

Information Technology

**Guided by Submitted by**

**Prof. Sumit Muddalkar Kiran Kolte**

**Ayush Batra**

**Ayush Naphade**

**Neha Lokhande**

DEPARTMENT OF INFORMATION TECHNOLOGY

SHRI SANT GAJANAN MAHARAJ COLLEGE OF ENGINEERING,

SHEGAON (M.S.)

**2021- 2022**

SHRI SANT GAJANAN MAHARAJ COLLEGE OF ENGINEERING

# **SHEGAON**



**2021-2022**

**CERTIFICATE**

This is to certify that **Mr. Kiran Kolte, Mr. Ayush Batra, Mr. Ayush Naphade, Ms. Neha Lokhande**---- students of final year B.E. (Information Technology) in the year 2021-2022 of Information Technology Department of this institute has completed the project work entitled “**Document Analyzer**”based on syllabus and has submitted a satisfactory account of his/her work in this report which is recommended for the partial fulfillment of degree of Bachelor of Engineering in Information Technology.

|  |  |
| --- | --- |
| **Name of Guide** (Prof. Sumit Muddalkar) | |
| **Prof. A S Manekar**  Head, Dept. of IT  S.S.G.M. College of Engineering,  Shegaon | **Dr. S. B. Somani**  Principal  S.S.G.M. College of Engineering,  Shegaon |

**ACKNOWLEDGEMENT**

**ABSTRACT**

To create an application which can extract data from any documents and store them. Amazon Textract uses Optical Character Recognition (OCR) technology to automatically detect printed text, handwriting, and numbers in a scan or rendering of a document, such as a legal document or a scan of a book. Form Extraction, Amazon Textract enables you to detect key-value pairs in document images automatically so that you can retain the inherent context of the document without any manual intervention. Table Extraction, Amazon Textract preserves the composition of data stored in tables during extraction. This is helpful for documents that are largely composed of structured data, such as financial reports or medical records that have column names in the top row of the table followed by rows of individual entries. Handwriting Recognition, many documents such as medical intake forms or employment applications contain both handwritten and printed text. Amazon Textract can extract printed text and handwriting from documents written in English with high confidence scores, whether it is free-form text or text embedded in tables and forms. Documents can also contain a mix of typed text or handwritten text. Document Text Detection, it can help to detect all the words and hence sentences from the document. This can be used to further process or get the information from the form.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

1. **INTRODUCTION**
   1. **Preface**

In today’s world everything is getting digitalized