CHRIST (DEEMED TO BE UNIVERSITY), NCR-201003
SCHOOL OF SCIENCES

Internship Activity Register for Students - Weekly Report

Name:	ANAV TRESA ROY	Date:	08-06-2024
Register Number:	22215016	Total Hours Spent:	30 hrs

Summary of activity done by the Student:

- Tried making changes to solve logical error in the program to make the files downloadable
- Made changes and now files are downloadable from the browser and pushed the code to git
- Tried to extract information from the pdf using PyPDF extract_text function, searched about other ways of data extraction using OCR techniques ,looked up on softwares such as 'tesseract' and its python integration "pytesseract" ,EasyOCR, Keras-OCR which allows processing of image/other document format and extract information into text format
- Installed easyocr,pdf2image libraries and also tried camelot library for table extraction from a pdf.
- Leant more about camelot ,resolved pypdf2 and camelot version conflict, looked into yolov2 model for object detection which can be modified to detect tables in pdf ,and camelot can be used to specify the area/region in a pdf ,which can be then passed on as a specific part from which data needs to be extracted. Looked into ghostscript which is a dependency of camelot.All these libraries can be used together to detect a table in a pdf more precisely.

Summary of Information/Discussion gathered from Internship Mentor: