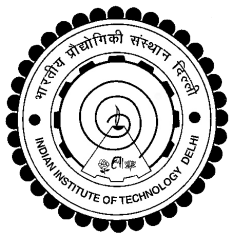
MTP2

K Laxman 2018CS50408

COD892 Project Proposal

Advisors: Prof.M.Balakrishnan, Prof.Volker Sorge, Abhishek Baghel
Supervisor: Prof. Vireshwar Kumar

January 04, 2023



Extracting table from pdf using layout parser and GCV(Google Clound vision)

Automated tool for converting documents containing the pdf files to extract text out of the image per page

1. **Problem** : extract structured data from a PDF file containing tables and text .
2. **Algorithm** :
3. Import required files and Load PDF file using pdf2image.
4. For each page in the PDF file:
5. a. Perform layout analysis using Detectron2 LayoutModel from layout parser.
6. b. Draw boxes around detected elements using lp drawbox function.
7. c. Extract tables from the layout analysis using lp.Layout() function.
8. d. Perform text recognition on the page image using the GCV API through the layoutparser library.

Extracting table from pdf using layout parser and GCV(Google Clound vision)

Automated tool for converting documents containing the pdf files to extract text out of the image per page

1. e. Draw text boxes around detected text using the lp.drawtext() function.
2. f. Extract text content and location coordinates using OCR agent from layoutparser.
3. Group text content into columns based on the x-coordinates
4. Convert the extracted table data into a pandas DataFrame and save as a CSV file.
5. end of loop and end of program

Extracting table from pdf using layout parser and GCV(Google Clound vision)

Group text content into columns based on the x-coordinates .How does it work ?

1. Initialize an empty dictionary to hold the columns
2. Loop through each text block:.
3. a. Determine the x-coordinate of the text block by finding the average of the minimum and maximum x-values of the block's bounding box
4. b. Check if there is already a column for that x-coordinate:
5. i. If there is, append the text content to the existing column
6. ii. If there is not, create a new column and add the text content to it
7. Determine the maximum length of any column
8. Remove any columns with fewer elements than the maximum length minus 3
9. Fill any missing values in each column with "None"
10. Create a pandas DataFrame with the resulting column-wise text content and save as CSV