Report

PDF anonymization and clinical analysis

1. Motivation

I had previously worked web scrapping using AutoScraper library in python. The AutoScraper scraps the required details from the specified website mentioned. As I worked on this project therefore, I chose this task.

2. Introduction about the task

As per the task there are 50 patient PDFs from that details I need to extract the data and check whether their any abnormal keywords in the PDF. If there are any abnormal keywords in the PDF scanned then the patient is abnormal.

As of my knowledge the above 50 patient medical details are related to the fetus.

3. Data extraction, preprocesses, and analysis

Here I used fitz from PyMuPDF library in python to scan and retrieve the data from the PDFs. And from the scanned data I stored only the required details in a list. And from the list I compared it with the abnormal keywords to check whether the patient is abnormal or not.

4. Results

By following the above procedure converted CSV file into JSON file based on the requirements.

5. Key findings

As I scanned the PDF we got all the details except the hospital logo. And as per the requirements we need to find the number of normal and abnormal patients in the given 50 patients. As I observed if any patient is abnormal then it will be mentioned in the PDF. So I used that word to check whether the word

is exist in the scanned PDF or not. If it is present then patient is abnormal otherwise normal.

.

6. Future work

In future I will try to store the scanned data in database in an organised manner and use it future purpose like preparing model based on the data and use it for predictions and so many.