ORIGIN HEALTH: RC

ALL TASKS: 1, 2, & 3

PRAJJWAL: DATA ENGIEER AND OPS

09.10.2023

REPORT

TASK 3

MOTIVATION

I like working with PDFs and photographs because they need a lot of visually appealing skills. Whether it's through manipulation or redaction, with a few lines of code, you can create and build lots of things on top of it and that's why I took this last optional task as a fun project.

INTRODUCTION ABOUT THE TASK

1. This task has two parts: the first one is PDF anonymisation and the second one is Clinical analysis.
2. PDF anonymisation: Because patient data is sensitive, this duty requires us to conceal patient personal information, whether through anonymisation or

blurring. I took a strategy that will make the required portion of the report difficult to read. After converting to an image, I discovered that every report has the same format, thus if I can determine the coordinates where to make the first image black, I can do it for the entire document.

3. Clinical analysis: Here with the basis of certain keywords associated with the normal reports, the tasks require us to make a conclusion. For this i used one simple method to extract all the text portions from the PDF and then compare them with our keywords, in this way we can find out if this report contains those keywords and categorise our report as abnormal and normal based on that.

DATA EXTRACTION, PREPROCESSES, AND ANALYSIS

PDF Anonymization : I first converted the PDF doc into jpg just to work with CV2 library . In that case, one easily finds out the coordinates and apply gaussian blur or fill the dark image. My initial approach was not this simple as i was trying to figure out a way where by just entering the key like patient value one can blur the image next to it. I used some of the regex techniques to located the text but it was taking a lot of my time hence shifted my focus towards completion of this basic technique.

Locating coordinates was an task through paint but i did try a technique which is shared in the task3b.ipynb where you can click teh image and get coordinates (x,y) . Furter tries to add two extra coordinated like width and height of rectangle but again time limit.

Clinical analysis : The keywords **were not** enough to classify reports, As i scanned through the documents there were keywords there but it has prefix like NO and sufic like found which can make our code obsolete. Hence I added this feature into my code to take care of that. Other stuff were basic pattern matching to the text which made easier by the library PyPDF2 pdfreader.

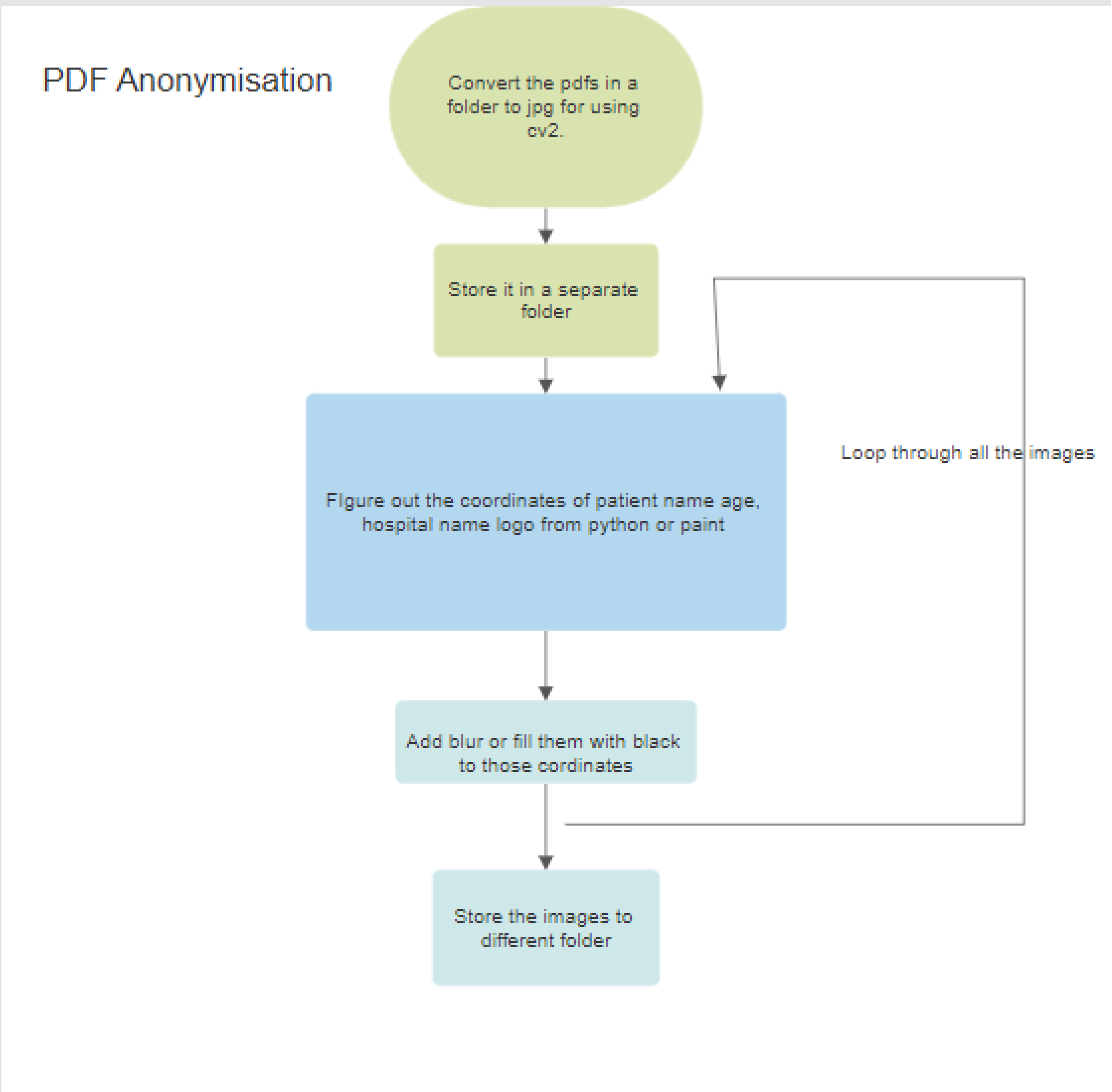
PDF Anonymisation
Convert the pdfs in a folder to jpg for using cv2.
Store it in a separate folder
FIgure out the coordinates of patient name age, hospital name logo from python or paint
Loop through all the images
Add blur or fill them with black to those cordinates
Store the images to different folder

Fig1: TASK 3 process flowchart

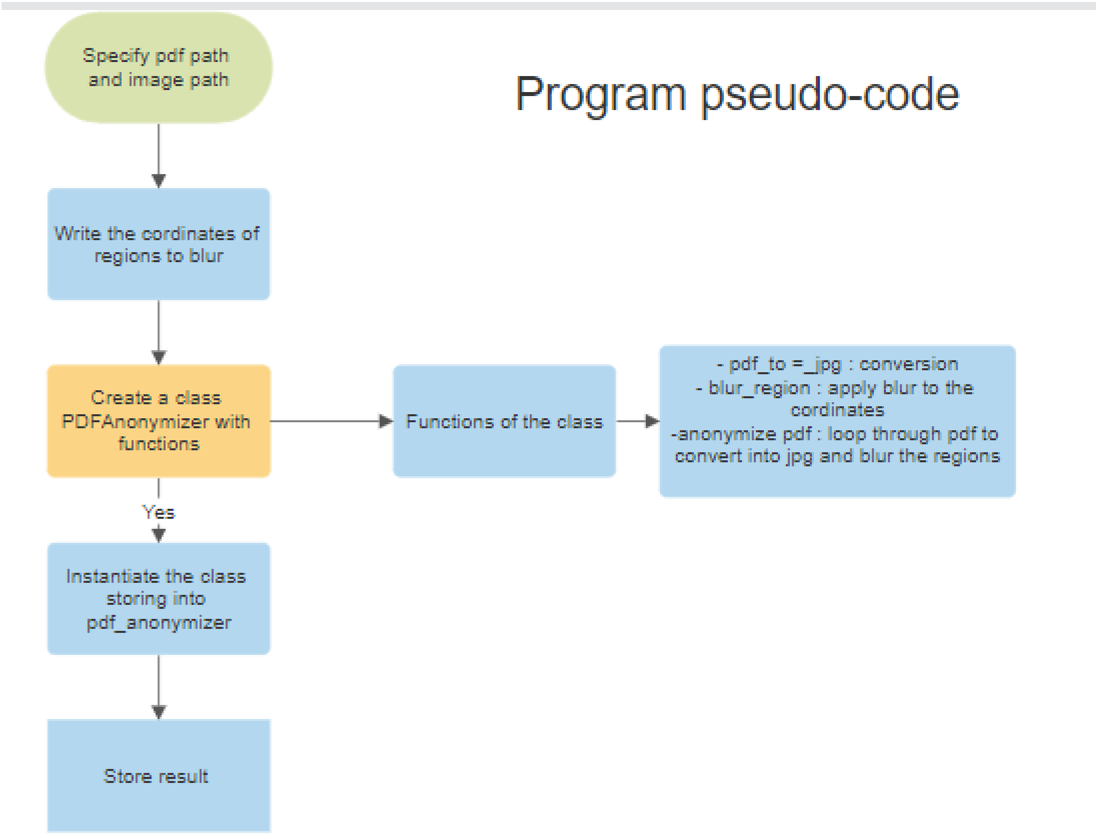
Program pseudo-code
Specify pdf path and image path
Write the cordinates of regions to blur
Create a class PDFAnonymizer with functions
Functions of the class
- pdf_to =_jpg : conversion
- blur_region : apply blur to the cordinates
-anonymize pdf : loop through pdf to convert into jpg and blur the regions
Yes
Instantiate the class storing into pdf_anonymizer
Store result

Fig2: TASK3 code flowchart

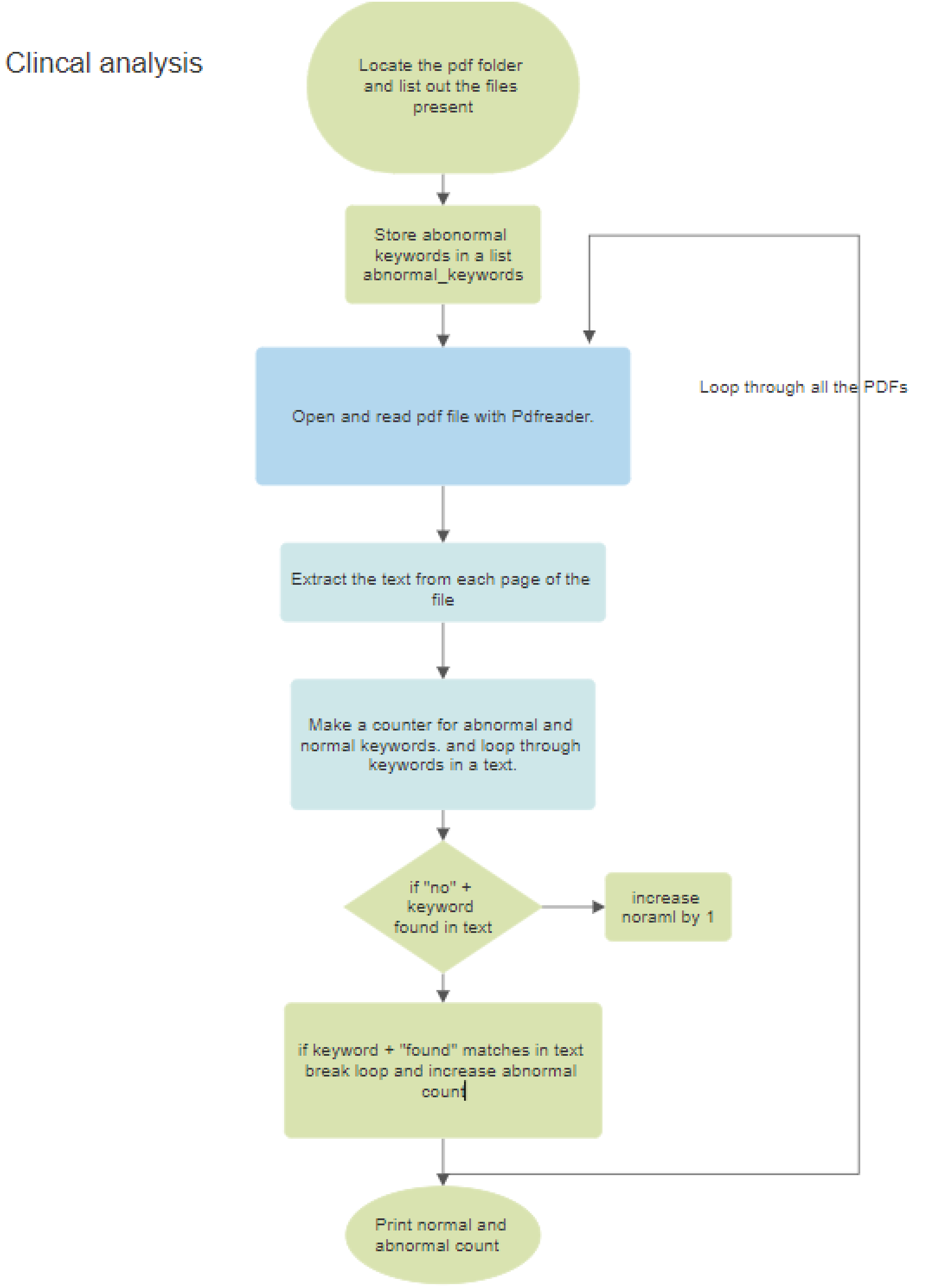
Clincal analysis
Locate the pdf folder and list out the files present
Store abonormal keywords in a list abnormal_keywords
Open and read pdf file with Pdfreader.
Loop through all the PDFs
Extract the text from each page of the file
Make a counter for abnormal and normal keywords. and loop through keywords in a text.
if "no" + keyword found in text
increase noraml by 1
if keyword + "found" matches in text break loop and increase abnormal count
Print normal and abnormal count

Fig 3 : TASK 3b - Clinical analysis procedure

RESULTS

Pdf were anonymized properly to hide sensitive information of patient and got the result into jpg files. If there is need of converting them into pdf again the this can be done easily too.
For the clinical analysis task, the number of normal cases was **35** and abnormal ones were **15**. Through the keywords provided, although the information in the pdf are limited and these keywords are enough, we can add extra keyword to it through domain knowledge after consulting with the analyst. I listed all the abnormal files and confirmed with the original for checking the validation of the code

KEY FINDINGS

PDF Anonymization: For one single hospital reports fromat will remain same and hence after figuring out the coordinate for one single file we can scale this code to anonymize all the reports present.

Clinical analysis: In this case also finding some keywords associated with disease we can rule out the possibility of abnormal cases. It doesn’t gives us that insights but for screening purposes of disease it can play a vital role. Through the power of reproducibility of code, we can see how we can infer the details of all the patients. Saving patients and doctors lots of time and money.

FUTURE WORK

As mentioned in the analysis part of the anonymisation task , how it will work for a particular hospital report only. Further case studies can be led to the development of singl;e model which can anonymise the section of the given information utilising the power of NLP or regex etc . I tried to search with regex with suitable arguments it can locate the given text but it distorts the format. With NLP advance techniques you can proceed for further analysis.

With the analyst and doctor consultation more robust approach can be built to detect disease also from the provided reports which can help in starting the diagnosis also.And wIth more data we can have statistical significant test for our results also. For example, applying t-test to null hypothesis to test our results are significant or not.