**PROGRAMMERS DIARY**

**WEEK 1:**

(15-03-2022 to 22-03-2022)

Went through related scientific/technical papers and articles regarding different methods of redaction used for born-digital documents, and learnt about the use of regular expression for finding text of a certain pattern.

**WEEK 2:**

(23-03-2022 to 30-03-2022)

Tried implementation of josh pdf-redactor. Discussion about finding PIIs from different buckets. Tried finding regexes of common PIIs in plain text like PAN, Aadhar, IFSC Code, Bank account number, etc. Also implemented basic methods of conversion of pdf to plain text(Text Extraction).

**WEEK 3:**

(31-03-2022 to 07-04-2022)

Searched for different modules and methods available for modification of pdf, tried extraction of pdf files with python modules. Basic implementation of regex patterns that would be applicable.

**WEEK 4:**

(08-04-2022 to 15-04-2022)

Pattern identification of first 3 buckets of different types implemented. Identification of orphan text with salutations can be identified. PDF layout aware modification tried.

**WEEK 5:**

(16-04-2022 to 23-04-2022)

Explored about metadata of PDF files. Tried code related to annotations. Tested different pdf modification techniques based on annotations and implemented different regrex patterns on the sample files.

**WEEK 6:**

(24-04-2022 to 01-05-2022)

Tried code related to characters stream extraction and editing. Learnt different methods of modifying the characters stream.

**WEEK 7:**

(02-05-2022 to 09-05-2022)

Tried and tested code on some other sample bank statements, and made minor corrections in regular expressions. Tried to redact orphan data like account numbers and names.

**WEEK 8:**

(10-05-2022 to 17-05-2022)

Tried PyMuPDF codes and went through pdfrw to perform modifications in pdf files PyMupdf.

**WEEK 9:**

(18-05-2022 to 25-05-2022)

PyMuPDF utilities explored, also explored other functionalities of pdfrw for pdf modification. Layout analyzer utility of PyMuPDF implemented to extract the layout of pdf files. Other demos and examples explored.

**WEEK 10:**

(26-05-2022 to 02-06-2022)

Tried test cases with pdfrw. Tested the code with some samples.Mid term presentation slides preparation.

**WEEK 11:**

(03-07-2022 to 10-07-2022)

Mid Term presentation. Tried pdfrw approach with the new sample files. Found a new method using pymupdf. Tried to redact using MuPDF. Tried the MuPDF approach with all test files. Measured and compared times for both the methods.

**WEEK 12:**

(11-07-2022 to 18-07-2022)

Optimized the PyMuPDF approach. Tried to reduce overlapping issues with MuPDF by reducing redundant regexes. Understood about flaskx, flaskrestplus and other related libraries.

**WEEK 13:**

(19-07-2022 to 26-07-2022)

Tried a few more test cases. Applied the project structure. Learned about swagger.ui and made API calls to run the redaction as a web api service. Prepared for final presentation.

**WEEK 14:**

(27-07-2022 to 30-07-2022)

Final presentation.