**WORK REPORT**

**(Task: OCR Text Highlight)**

* **Normal Pdf & OCR content Extraction :**

1. Used PyPDF2 library to extract text content from normal PDF files.
2. And Integrated Tesseract OCR functionality to process scanned PDF files, converting image-based text into machine-readable text.
3. Applied pre-processing techniques such as noise reduction and binarization to improve OCR accuracy for scanned PDFs.

* **Text Highlighting in PDFs :**

1. **Similarity Search for Meaning-Based Highlighting:**

* Implemented Similarity Search technique to highlight text based on its meaning.
* This approach enhances text highlighting by considering semantic similarity, improving comprehension and usability.
* But issue at some level it highlight the text of normal pdf but facing issue for the response that is in summarise manner.

1. **Manhattan Distance for Refining Highlighting:**

* Utilized Manhattan distance to identify and highlight the most similar sentences or text segments.
* This method aids in reducing unnecessary highlighting, enhancing the precision of highlighted content.

1. **Jaccard Distance :**

* Used Jaccard distance measure to refine highlighting is a strategic choice. It calculates the dissimilarity between sets by comparing their intersection and union, making it suitable for measuring text similarity.
* But for this method I not get an expected output.

1. **Cosine Similarity using SciPy:**

* In this Cosine Similarity it compute similarity between text segment.
* For Cosine similarity I achieve some text highlighting on threshold 0.75 and 0.81, in this range some text is highlighting but some extra unnecessary text also highlight.

1. **OCR pdf highlight :**

* For OCR pdf highlight I tried using pytesseract and OpenCV(cv2) python libraries
* Issue: it only highlight single word in OCR file, not more than that.
* **Conclusion:**
* Normal and OCR pdf text extraction done successfully.
* For Normal pdf text highlight I tried some similarity search technique also some distances, but not get an expected output.
* Then I tried on OCR file for text highlight but same issue occur, means in my side only some character is highlight in OCR file.