* Some documents are scanned and uploaded
* Reading text from pdf
* Splitting the text into segments. (still needs to refine)

Software needed:

* Tesseract-ocr
* Poppler
* Pdf2image (pip)
* Pillow (pip)
* pyTesseract (pip)
* numpy(pip)