* Some documents are scanned and uploaded
* Reading text from pdf
* Splitting the text into segments.