Mini project: Data preprocessing

Aim: Preprocessing of data from data sheets in the form of pdf using various tools

Approach:

1. Read the data from the pdf
2. Differentiate between headers, subheaders, paragraphs and tables
3. Convert data to JSON format

Firstly I used pypdf library to read the data from the pdf but later on switched to pdfplumber library as it seemed to be a better option.
Pdfplumber had an inbuilt method which would return tables in the form of tuples.

The next step was to write some code that would help to identify headers, subheaders and paragraphs. This would have to be done considering the font sizes, number of words in each line and the font weights. This wasnt quite successful as I couldnt define these properties clearly. If an entity was detected the necessary function would be called to generate some sort of key value pairs.

Further to obtain key value pairs from paragraphs the idea was to use regular expressions or use spaCy. Like probably by using conditions and checking if a particular key was found then starting a search for the necessary values using regex.

For tables either the rows or the columns would have to be the keys and the other would have to be considered the values.

Progress:
I have an overall idea as to how the project has to be executed and the necessary libraries. The further steps would be to properly identify the required entities in the pdf file and find a way to arrange them in the required format.