Extract text from PDF File using Python

All of you must be familiar with what PDFs are. In fact, they are one of the most important and widely used digital media. PDF stands for Portable Document Format. It uses .pdf extension. It is used to present and exchange documents reliably, independent of software, hardware, or operating system.

Extracting Text from PDF File

Python package PyPDF can be used to achieve what we want (text extraction), although it can do more than what we need. This package can also be used to generate, decrypting and merging PDF files.

Id	Name	CGPA
1	Rasel	3.69
2	Jamil	3.80
3	Ruhul	3.78
4	Riaz	3.66

How to Extract Text from PDF

In this blog, we are going to examine the most popular libraries for processing PDFs with Python. A lot of information is shared in the form of PDF, and often we need to extract some details for further processing.

To assist it in my research in identifying the most popular python libraries, I looked across StackOverflow, Reddit and generally lots of google searches. I identified numerous packages, each with its own strengths and weakness. Specifically, users across the internet seem to be using: PyPDF2, Textract, tika, pdfPlumber, pdfMiner.