

Python is widely used for analyzing the data but the data need not be in the required format always.

In such cases, we convert that format (like PDF or JPG etc.) to the text format, in order to analyze the data in better way. Python offers many libraries to do this task.

There are several ways of doing this, including using libraries like [PyPDF2 in Python](#). The major disadvantage of using these libraries is the encoding scheme.

PDF documents can come in a variety of encodings including UTF-8, ASCII, Unicode, etc. So, converting the PDF to text might result in the loss of data due to the encoding scheme.