Introduction to Document Parsing

Document parsing is the process of extracting meaningful text and data from documents such as PDF and DOCX files. It is widely used in applications like resume screening, data extraction, and digital archiving.

PDF files are commonly used because they preserve formatting across devices. However, extracting text from PDF files is not always easy due to different layouts, fonts, and embedded images.

To solve this problem, several PDF parsing libraries are available in Python. Each library uses a different technique to read and extract text from documents.

In this project, PDF parsing libraries such as PDFPlumber and PyMuPDF are evaluated. The accuracy of the extracted text is measured using the Word Error Rate (WER) metric.

Word Error Rate calculates the difference between the extracted text and the correct reference text. A lower WER value indicates better accuracy.

The selected parser is then used to build a document conversion module. This module extracts text from PDF and DOCX files and converts it into text (.txt) or markdown (.md) formats.

This approach helps automate document processing tasks and improves efficiency in handling large numbers of files.