OCR RESULTS ON PDFS

Dataset Description:

The dataset is currently divided into two parts:

1. Normal PDFs: There are 9 PDF documents classified under this category. These PDFs predominantly consist of textual content without complex formatting such as tables or diagrams.
2. PDFs with Tables and Diagrams: This category comprises 24 PDF documents. These PDFs contain textual information along with tables and diagrams.

Evaluation Metrics:

1. Character Error Rate (CER): measures the rate of wrong characters compared to the ground truth
2. Word Error Rate (WER): measures the proportion of incorrectly recognized words relative to the ground text

Tesseract OCR Results:

Tesseract OCR was applied to both categories of PDFs (all 9 for normal pdfs and on 4 of them from the pdfs with Tables) with the following outcomes:

|  | CER | WER |
| --- | --- | --- |
| plain pdfs (9) | 2.0801 | 3.98 |
| Pdfs tables (4) | 0.3953 | 6.0438 |

PyMuPdf Results:

PyMuPdf was applied to both PDF groups. It is an open-source Python library used to read and modify PDFS.

|  | CER | WER |
| --- | --- | --- |
| plain pdfs (9) | 0.7256 | 2.2042 |
| Pdfs tables (5) | 0.1815 | 0.1787 |

Resources:  
1. TO RUN OCR ON PDFS (MODELS)  
<https://www.reddit.com/r/LocalLLaMA/comments/172k9q2/best_model_for_document_layout_analysis_and_ocr/>: 3 resources in here:

1. PyMuPDF Library with Fitz Module:
2. DONUT
3. Nougat

2. DOCUMENTATION BLOCK DIAGRAM HELP:

<https://www.researchgate.net/figure/A-block-diagram-of-a-typical-OCR-system-main-stages-The-diagram-shows-intermediate-steps_fig1_346164817>

3. <https://medium.com/@dr.booma19/extracting-text-from-pdf-files-using-ocr-a-step-by-step-guide-with-python-code-becf221529ef>

4. PYTHON LIBRARY FOR OCR: <https://pypi.org/project/PyPDF2/>

5. Ocrmypdf tool: uses Tesseract and also performs image processing. Extra step to convert img to pdf involved. [<https://github.com/ocrmypdf/OCRmyPDF?tab=readme-ov-file>]

6.<https://medium.com/@tam.tamanna18/deciphering-accuracy-evaluation-metrics-in-nlp-and-ocr-a-comparison-of-character-error-rate-cer-e97e809be0c8>