Please use csv files to understand the dataset structure.

You can use this colab file to download the dataset: [data\_ingestion\_code\_gdrive](https://colab.research.google.com/drive/1xKc5pfvxeeg4N7r5AzIDy-m3rVDqNqhB)

Total Stats for 'markdown\_content' column in DataFrame:

Total Words: 173706

Total Characters: 1074790

### **Dataset Column Descriptions**

1. **id***Description*: A unique identifier assigned to each document, starting from 0 and incrementing by 1 for each subsequent document.
2. **class***Description*: Defines the type of the document. Currently, there are four classes: ACT, Circulers, Regulation, and Guideline. There may be future changes to consolidate these into two classes: **Regulatory** and **Guidance**.  
   *Note*: Misspellings from the original dataset (e.g., "Circulers") have been retained for error detection purposes.
3. **filename***Description*: The name of the PDF file, including the .pdf extension. If document has only text, the text filename is added. This column is used with conjunction with **PDF\_or\_text** to select paths.
4. **path***Description*: The location of the PDF on the drive, either as an absolute path or relative to the root folder, depending on where the CSV file is uploaded.
5. **url***Description*: The URL from which the PDF document was scraped or collected.
6. **language\_label***Description*: The language of the document as detected by the **LangDetect** library. Currently, only English ('en') is supported. Other languages or dual-language documents are excluded from the dataset.
7. **required\_ocr***Description*: Indicates whether Optical Character Recognition (OCR) was necessary to extract text from the PDF.
   * **'no'**: Text was readable without OCR.
   * **'yes'**: OCR was required (via **pytesseract**).
   * **'unextractable'**: The document could not be processed by OCR or text extraction methods.  
     *Note*: Documents marked as 'unextractable' have been removed from the dataset.
8. **has\_tables***Description*: Indicates whether the document contains tables, detected via the **gmft** library.
   * **'yes'**: Tables were detected.
   * **'no'**: No tables were detected.
9. **text\_path***Description*: The file path to the text extracted from the PDF using **py2PDF** or **pytesseract**. This column is also used for documents that were directly added as text (such as ACTs) instead of PDFs.
10. **quality***Description*: The quality of the extracted text.

* **'bad'**: If no text was extracted or if manual inspection revealed issues such as noise, skewed pages, or multiple pages merged into one.
* **'good'**: If text extraction was successful and the document quality is acceptable.

1. **data\_origin***Description*: Indicates the source of the document.

* **'original\_dataset'**: The document is part of the dataset provided by the Omdena team.
* **'scraped'**: The document was collected from the internet.

1. **retrieved\_date\_of\_issuance***Description*: The date the document was issued, extracted from HTML tags during scraping.
2. **issuing\_authority***Description*: The department or organization that issued the document. The document class (e.g., **Circulers**) is appended for easier identification in later stages of the pipeline (e.g., "TRI Circulers").
3. **retrieved\_topic***Description*: The topic or subject of the document, extracted either from HTML tags during scraping or inferred from the document name.
4. **text\_content***Description*: The actual text content extracted from the document. This is typically stored in the text\_path text files. Only for understanding that we need advanced parsing since this does not have layout info and correct table parsing, and is also unclean. This text is extracted on an as is basis.
5. **PDF\_or\_text***Description*: Specifies whether the document is a **PDF** or **direct text**. This column helps to determine the correct source for reading the document (from the path for PDFs or text\_path for text files).
6. **llama\_markdown\_content***Description*: The cleaned document content with layout information extracted using **llamaparse.**

Use <https://markdownlivepreview.com/> to view markdown content to see formatting

Use this for further RAG

1. **llama\_title***Description*: The main title of the document, extracted using **llamaParse**.
2. **llama\_issue\_date***Description*: The issue date of the document (not the publishing date), as extracted via **llamaParse**.
3. **llama\_reference\_number***Description*: The reference, serial, or circular number of the document, often containing dates or other details that may require further exploration.