School Board Library Copilot - Document Classification Logic

Overview

This document outlines the classification logic used by the School Board Library Copilot to categorize documents based on their file type and content. The classification process involves a primary classification based on the file extension, followed by secondary and sometimes tertiary classification steps to analyze the content in more detail.

Classification Process

1. **Primary Classification:**
   - The primary classification is based on the file extension (e.g., .pdf, .docx, .jpg).
   - The extension is converted to uppercase and used as a preliminary document type (e.g., PDF, DOCX, JPEG).
2. **Secondary Classification:**
   - Secondary classification functions are called to analyze the content of the document in more detail.
   - These functions use appropriate libraries to extract text, images, tables, and other relevant information.
   - The specific functions called depend on the primary classification and are defined in the decision_tree.json file.
3. **Tertiary Classification (if applicable):**
   - In some cases, further classification steps might be performed based on the results of the secondary classification.
   - For example, image-only PDFs might be further classified as single-page or multi-page.

Decision Tree

The decision_tree.json file defines the mapping between file extensions, primary classifications, and secondary classification functions.

JSON

```
{
  "txt": {
    "primary": "TXT",
    "secondary": ["classify_text_document"]
  },
  "csv": {
```

```
    "primary": "CSV",
    "secondary": ["classify_csv_document"]
  },
  "docx": {
    "primary": "DOCX",
    "secondary": ["classify_docx_document"]
  },
  "xlsx": {
    "primary": "XLSX",
    "secondary": ["classify_excel_document"]
  },
  "pdf": {
    "primary": "PDF",
    "secondary": ["classify_pdf_for_images", "classify_pdf_for_text",
"classify_pdf_for_tables"]
  },
  "jpg": {
    "primary": "JPEG",
    "secondary": ["classify_image_document"]
  },
  "jpeg": {
    "primary": "JPEG",
    "secondary": ["classify_image_document"]
  },
  "png": {
    "primary": "PNG",
    "secondary": ["classify_image_document"]
  },
  "gif": {
    "primary": "GIF",
    "secondary": ["classify_image_document"]
  },
  "bmp": {
    "primary": "BMP",
    "secondary": ["classify_image_document"]
  },
  "tiff": {
    "primary": "TIFF",
    "secondary": ["classify_image_document",
"classify_tiff_as_multipage"]
  },
  "rtf": {
    "primary": "RTF",
    "secondary": ["classify_rtf_document"]
  },
  "html": {
    "primary": "HTML",
    "secondary": ["classify_html_document"]
```

```
  },
  "xml": {
    "primary": "XML",
    "secondary": ["classify_xml_document"]
  },
  "zip": {
    "primary": "ZIP",
    "secondary": ["classify_zip_contents"]
  },
  "pptx": {
    "primary": "PPTX",
    "secondary": ["classify_pptx_document"]
  },
  "odt": {
    "primary": "ODT",
    "secondary": ["classify_odt_document"]
  }
}
```

Document Types

The following table outlines the possible document types after classification:

File Extension	Primary Classification	Secondary Classification Functions	Possible Document Types
.pdf	PDF	classify_pdf_for_images, classify_pdf_for_text, classify_pdf_for_tables	Text-Only, Text-with-Images, Text-with-Tables, Image-Only
.docx	DOCX	classify_docx_document	Text-Only, Text-with-Images, Text-with-Tables

File Extension	Primary Classification	Secondary Classification Functions	Possible Document Types
.jpg, .jpeg	JPEG	classify_image_document	Image-Only
.png	PNG	classify_image_document	Image-Only
.gif	GIF	classify_image_document	Image-Only
.bmp	BMP	classify_image_document	Image-Only
.tiff	TIFF	classify_image_document, classify_tiff_as_multipage	Image-Only-Single-Page, Image-Only-Multi-Page
.pptx	PPTX	classify_pptx_document	Text-Only, Text-with-Images, Image-Only
.txt	TXT	classify_text_document	Text-Only
.csv	CSV	classify_csv_document	Text-Only
.xlsx	XLSX	classify_excel_document	Text-Only
.rtf	RTF	classify_rtf_document	Text-Only
.html	HTML	classify_html_document	Text-Only
.xml	XML	classify_xml_docu	Text-Only

File Extension	Primary Classification	Secondary Classification Functions	Possible Document Types
		ment	
.zip	ZIP	classify_zip_contents	Text-Only
.odt	ODT	classify_odt_document	Text-Only

Secondary Classification Functions (Illustrative)

Here are brief descriptions of some secondary classification functions. The actual implementation will depend on the specific libraries used.

- **classify_pdf_for_images(filepath):** Detects if a PDF contains images using PyMuPDF.
- **classify_pdf_for_text(filepath):** Attempts to extract text from a PDF using PyMuPDF or pdfminer.six.
- **classify_pdf_for_tables(filepath):** Detects tables in a PDF using tabula-py.
- **classify_docx_document(filepath):** Parses a DOCX file using python-docx and identifies paragraphs, headings, images, and tables.
- **classify_image_document(filepath):** Performs basic image analysis to potentially determine if an image is a photograph, diagram, or scanned document.
- **classify_tiff_as_multipage(filepath):** Determines if a TIFF file contains multiple pages.
- **classify_pptx_document(filepath):** Extracts text and images from slides in a PPTX file using python-pptx.
- **classify_text_document(filepath):** Analyzes plain text content to determine the structure of the document.
- **classify_csv_document(filepath):** Analyzes CSV structure.
- **classify_excel_document(filepath):** Analyzes .xlsx content.
- **classify_rtf_document(filepath):** Analyzes rtf content.
- **classify_html_document(filepath):** Parses HTML content.
- **classify_xml_document(filepath):** Parses XML content.
- **classify_zip_contents(filepath):** Extracts the contents of a ZIP archive.
- **classify_odt_document(filepath):** Analyzes ODT content.