PDF Table Extraction and Export to Excel

Overview

This script extracts tabular data from PDF files and exports it as Excel files. It processes individual PDFs or batches of files within a directory.

Dependencies

- `pdfplumber` for extracting text from PDF files
- `pandas` for data manipulation and exporting to Excel
- `os` for handling file operations

Functions

1. `segment_text_by_lines(word_data, line_threshold=3)`

Organizes extracted words into lines based on their vertical positioning.

Parameters:

- `word_data` (list): List of word dictionaries extracted from a PDF page.
- `line_threshold` (int): Maximum vertical difference for grouping words into the same line.

Returns:

- `structured_lines` (list): List of grouped words forming lines.

2. `arrange_text_into_columns(text_segments, column_spacing=5)`

Clusters words into columns based on horizontal spacing.

Parameters:

- `text_segments` (list): List of words from a structured line.
- `column_spacing` (int): Minimum horizontal gap to separate columns.

Returns:

- `column_structure` (list): List of text columns.

3. `extract_table_from_pdf(file_path)`

Extracts tabular text content from a given PDF file.

Parameters:

- `file_path` (str): Path to the PDF file.

Returns:

- `extracted_data` (list): List of extracted rows, where each row is a list of column values.

4. `export_to_excel(data_matrix, save_path)`

Exports extracted tabular data into an Excel spreadsheet.

Parameters:

- `data_matrix` (list): 2D list of extracted tabular data.
- `save_path` (str): File path to save the Excel file.

5. `batch_process_pdfs(source_directory, destination_directory)`

Processes multiple PDF files from a directory and saves extracted tables as Excel files.

Parameters:

- `source_directory` (str): Directory containing input PDF files.
- `destination_directory` (str): Directory where Excel files will be saved.

Execution

The script can be executed by setting the input and output directories and running:

```
if __name__ == "__main__":
    input_directory = "Input_pdfs"  # Directory holding PDF files
    output_directory = "extracted_tables"  # Directory for output Excel files
    batch_process_pdfs(input_directory, output_directory)
```

Notes

- The script extracts words and organizes them into structured lines and columns before exporting them as tables.
- Ensure the input directory exists and contains valid PDF files.
- The script automatically creates the output directory if it does not exist.