ReadMe: CLRR Data Extraction (Databricks)

Overview

The code extracts specific information from PDF files in a given folder, processes the data, and saves it into an Excel file. The code leverages a custom class PDFDataExtractor to extract information such as country names, dates, fiscal years, and other key project-related information. After extraction, the data is combined into a DataFrame and then exported to an Excel file in the specified output folder.

Getting Started

To run this script in a Databricks environment, follow the steps below:

1. Set Up Databricks Workspace

* If you do not have a Databricks account, create one.
* Log in to your Databricks workspace.

2. Install Required Libraries

The script uses Pandas and OpenPyxl for data manipulation and Excel file handling. Ensure these libraries are installed in your Databricks cluster by running the following command in your notebook:

‘‘%pip install pandas openpyxl’

3. Add Custom Python Module

* Ensure that PDFDataExtractor.py is in the same directory as your Databricks notebook. This custom module contains the class used for extracting data from PDFs.

4. Set File Paths

* PDF Folder Path: Update the folder\_path variable with the path to the folder containing your PDF files.
* Output Folder: Set the output\_folder variable to the desired output directory.
* Output File: Provide a name for the resulting Excel file in the output\_file variable.

5. Execute the Notebook

* Once you have configured the paths, execute the cells in the notebook to extract and save the data.
* Monitor the output for any errors or messages. Databricks provides detailed logs that can help with troubleshooting.

6. Access the Output

* After the script completes successfully, the Excel file will be saved in the specified output folder.
* You can download the Excel file or work with the data within Databricks.

7. Additional Comments

* If you receive errors indicating missing modules, ensure that PDFDataExtractor.py is in the same directory as your notebook.
* The script assumes PDF files have a consistent format for extraction. If the structure of the PDF files varies, additional modifications may be needed.

Troubleshooting

* If the script fails to extract data, ensure the specified folder path is correct and contains PDF files.
* If the script does not generate the expected output, check for error messages in the output logs to identify the source of the problem.
* Make sure to check the output folder to confirm the Excel file was created successfully.