# **ReadMe: URL-Based Bulk Document Download**

## **Overview**

This script facilitates the bulk download of documents from specified URLs. It is designed to download files in text or binary format (PDF), rename them based on extracted project IDs and GUIDs, and create an Excel summary detailing the downloaded files' project IDs and file types. Additionally, the script handles a variety of World Bank document types, provides URL mapping, and extracts relevant information from downloaded files.

## **Required Libraries**

Ensure the following Python libraries are installed:

* requests: For HTTP requests.
* re: For regular expressions.
* os: For interacting with the operating system.
* time: For time-related operations.
* json: For handling JSON data.
* shutil: For file operations.
* pdfplumber: For extracting content from PDF files.
* pandas: For data manipulation.
* openpyxl: For creating and manipulating Excel workbooks.
* urllib.parse: For parsing URLs.

Install any missing libraries with:

‘pip install requests pdfplumber pandas openpyxl’

## **Running the Script**

### **1. Databricks Workspace Setup**

* Create or log in to a Databricks account.
* Set up the necessary notebook environment.

### **2. Define Input Parameters**

* Create an Excel file with a column named "URL" containing the URLs of the documents to be downloaded.
* Download Folder Path: Path to save the downloaded files.
* Excel Path: Path to the Excel file containing the URLs for downloading (under URL column name).

### **3. Execute the Script**

Set the following variables and execute the script to:

* Download Files: Downloads files from the specified URLs.
* Rename Files: Renames files based on project ID and GUID.
* Create Excel Summary: Creates an Excel summary detailing the downloaded files' project IDs and file types.

### **4. Access the Output**

After running the script, the outputs include:

* Downloaded Files: Files saved in the specified folder.
* Renamed Files: Files renamed based on extracted information.
* Excel Summary: A summary of downloaded files with their project IDs and file types.

## **Troubleshooting**

* If the script fails to download files, ensure the URLs are correct and accessible.
* If file renaming does not work as expected, verify the project IDs and GUIDs in the Excel or JSON files.
* If there are errors during execution, check the error messages for hints on what went wrong.