# ReadMe: World Bank Document Bulk Download (Date)

## **Overview**

This script facilitates bulk downloading of World Bank documents, categorization, renaming, and summarization of downloaded files. It uses the World Bank API to fetch documents based on a variety of criteria, including document type, date range, or project ID. The script also includes functions for renaming downloaded files based on specific metadata and generating an Excel summary of downloaded files.

## **Required Libraries**

The following Python libraries are required for this script:

* requests: For making HTTP requests.
* os: For interacting with the operating system.
* re: For regular expressions.
* time: For time-related operations.
* json: For handling JSON data.
* shutil: For file operations.
* openpyxl: For working with Excel files.

Ensure these libraries are installed in your environment before running the script.

## **Running the Script**

### **1. Setting Up the Environment**

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

### **2. Install Required Libraries**

To ensure all dependencies are installed, run the following command in your Databricks notebook:

‘pip install requests openpyxl’

### **3. Prepare the Data**

* Determine the root path where the downloaded files will be stored.
* Identify the document types you wish to download. These can include various World Bank documents like ICRR, ICR, PPAR, PAD, and others.
* Define the start and end dates for document downloads..
* Choose whether to download text (txt) or PDF (pdf) versions of the documents.

### **4. Execute the Script**

* Set the necessary input parameters in the input\_dict, including the root path, document type(s), start and end dates, and the desired download version.
* Run the script. This will:
  + Download documents based on the specified criteria.
  + Rename the downloaded files based on metadata from the World Bank API.
  + Generate an Excel summary of the downloaded files.

### **5. Access the Output**

After the script completes, you will have the following outputs:

* Downloaded Documents: The downloaded files, saved in the specified folder.
* Renamed Files: Files renamed based on metadata, ensuring a consistent naming convention.
* Excel Summary: An Excel file summarizing the downloaded files, including project ID and file type.

## **Script Functionality**

The script contains the following main components:

* Document Download: Uses the World Bank API to download documents based on criteria such as date range, document type, and project ID.
* File Management: Renames downloaded files based on specific metadata, ensuring consistency and clarity.
* Excel Summary: Generates an Excel summary of the downloaded files, providing an overview of the project IDs and file types.

### **Output Report**

The Excel summary contains the following columns:

* Proj.Id: The project ID associated with the downloaded document.
* File Type: The type of document, such as ICRR, PAD, or others.

## **Troubleshooting**

* If the script fails to find the specified folder, ensure the root path is correct and accessible.
* If the downloads do not complete successfully, check the start and end dates, document types, and ensure the World Bank API is accessible.
* If errors occur during execution, review the error messages to identify the cause and adjust the script or input parameters accordingly.