# **ReadMe: Bulk Download Project ID**

## **Overview**

This script allows for bulk downloading of World Bank documents based on specific project IDs. It interacts with the World Bank API to retrieve documents of various types, such as Implementation Completion Reports, Project Appraisal Documents, Systematic Country Diagnostics, and more. The script includes functions for renaming downloaded files based on project ID and GUID, and generating an Excel summary of the downloaded files.

## **Required Libraries**

This script requires the following Python libraries:

* requests: For HTTP requests.
* re: For regular expressions.
* os: For interacting with the operating system.
* time: For time-related operations.
* json: For working with JSON data.
* shutil: For file operations.
* pandas: For data manipulation.
* openpyxl: For creating and manipulating Excel workbooks.

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

## **Running the Script**

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

* 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 pandas openpyxl’

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

* Place the required Excel file with project IDs (Proj.Id column) in a specified location.
* Set the download folder path, the desired document types, and the download format (either txt or pdf).
* Specify the Excel file path to read the list of project IDs.

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

* Set the necessary input parameters, including the root path, document type(s), download version, and project ID list.
* Run the script. This will:
  + Download documents based on the specified project IDs.
  + Rename files based on project ID and GUID.
  + 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 project ID and GUID.
* Excel Summary: An Excel file summarizing the downloaded files, including project ID and file type.

## **Script Functionality**

The script includes the following main components:

* Document Download: Uses the World Bank API to download documents based on project ID.
* File Management: Renames files based on project ID and GUID, ensuring a consistent naming convention.
* Excel Summary: Generates an Excel summary of downloaded files, providing an overview of the project IDs and file types.

### 

### **Key Functions**

* SingleTypeBulkDownload: A class that facilitates downloading documents based on project IDs.
* read\_catalog\_files: Reads catalog files and yields query file dictionaries for further processing.
* generate\_catalog\_for\_projects: Generates catalog files for project-based document retrieval.
* download\_documents\_for\_projects: Downloads documents based on project IDs.
* extract\_document\_id: Extracts the document ID from a filename.
* generate\_renamed\_filename: Generates a renamed filename based on project ID and GUID.
* rename\_actual\_file: Renames a file based on project ID and GUID.
* rename\_actual\_files: Renames all files in a folder based on information from JSON sources.
* create\_excel\_summary: Creates an Excel summary for downloaded files.

### **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 or Excel file, ensure the paths are correct and accessible. Moreover the Project IDs shall be in the column ‘Proj.Id’ of the excel sheet.
* If the downloads do not complete successfully, ensure the World Bank API is accessible and that the project IDs are correct.
* If errors occur during execution, review the error messages to identify the cause and adjust the script or input parameters accordingly.