ReadMe: Keyword-Based Project Text Extraction (Databricks)

**Overview:**

The code extracts context sentences containing specified keywords from text and PDF files. It can handle two scenarios: one where there's only a list of primary keywords, and another with both primary and secondary keywords. The script reads the text/PDF files, identifies relevant sentences with the specified keywords, collects surrounding context, and saves the extracted information in Excel and pickle formats.

**Required Libraries:**

This script requires the following Python libraries:

* pandas: For data manipulation and handling Excel and CSV files.
* nltk: For natural language processing, particularly tokenization and part-of-speech tagging.
* tika: For parsing PDF files.
* re: For regular expression operations.
* pickle: For serializing Python objects.

Ensure these libraries are installed in your Databricks 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 met, run the following command in your Databricks notebook:

Python ‘%pip install pandas nltk tika’

3. Prepare the Data

* Place the text and PDF files in a specified folder for processing.
* Create an Excel file containing the keywords. This file should have at least one column named "Primary Keyword" (for primary keywords) and optionally a "Secondary Keyword" column (for secondary keywords).
* Set the correct paths for the folder containing data files, the Excel file with keywords, and the desired output files (Excel and pickle).

4. Execute the Script

* Ensure the required libraries are installed and the data is prepared.
* Set the paths for the input data folder, the Excel file with keywords, and the desired output files.
* Run the script cells in your notebook.
* Monitor the output for any error messages. Databricks provides detailed logs to aid in troubleshooting.

5. Access the Output

* After the script completes successfully, you will have two output files: an Excel file containing the extracted data and a pickle file for further processing.
* You can download the Excel file for further analysis or work with the data within Databricks.
* You can also load the pickle file to quickly retrieve the extracted data.

**Script Functionality:**

The script processes text and PDF files to extract sentences containing specified keywords.

It handles two cases: primary keywords only and primary with secondary keywords.

For each file, it identifies the project ID (if available), removes unwanted sections (like table of contents), and cleans text from noise and unnecessary elements.

It tokenizes the text into sentences and collects surrounding context for each sentence with a specified keyword.

It ensures that the extracted context has a minimum number of sentences with more than 4 words.

**Output Report:**

The output report contains the following information:

* Project ID: The unique identifier for each project.
* Primary Keyword: The primary keyword that triggered the context extraction.
* Secondary Keyword (optional): The secondary keyword, if applicable.
* Distance: The distance between the primary and secondary keywords in the context.
* Main Sentence: The sentence that contains the primary keyword.
* Text: The extracted context, ensuring at least 5 sentences with more than 4 words.