**READ\_ME ICR\_Preperation\_Implementation (Databricks)**

**Overview:**

The Python script processes text files within a specified folder, aiming to extract sections on project preparation and implementation. It employs regular expressions to remove table of contents, find specific patterns, and clean the extracted text. Each file's project ID is determined and associated with the relevant text. The script iterates over all .txt files in the folder, performing these operations and aggregating the results in a list. Finally, this data is saved into a pickle file at a designated path, creating a structured dataset from the unstructured text files.

**Running the Code**:

There are two cells that can be run. Run the first one if the files are year 2018 and onwards. Run the second part if the files are from before the year 2018. Please follow the steps below.   
  
1) Set Up a Databricks Workspace:

If you don't already have a Databricks account, create one.

Log in to your Databricks workspace.

2) Import Data:

Since the script processes files from a specific folder, you need to upload your .txt files to Databricks.

You can upload data to DBFS (Databricks File System) or mount a storage (like AWS S3, Azure Blob Storage) to DBFS.

3) Create a Notebook:

In your workspace, create a new notebook.

Choose Python as the language for the notebook.

4) Adapt the Script:

Copy the Python script into cells in the notebook.

Modify the folder\_path and output\_pickle\_path in the script to point to the correct paths in DBFS or your mounted storage.

Ensure that all dependencies (like the re and pickle modules) are available in your Databricks environment. These are standard Python libraries and should be available by default.

5) Run the Notebook:

Attach the notebook to the cluster you created.

Run the cells in the notebook sequentially.

Monitor the output for any errors or issues.

6) Accessing Output:

After the script runs successfully, the output pickle file will be saved in the path specified by output\_pickle\_path.

You can access this file directly in Databricks or download it to your local machine.