1. Import Required Libraries

Code:

import nltk

import pandas as pd

from nltk.tokenize import word\_tokenize

* **nltk:** The Natural Language Toolkit, used for text processing tasks such as tokenization, stemming, lemmatization, etc.
* **pandas:** A powerful data manipulation library used for handling structured data like CSV, Excel, etc.
* **word\_tokenize:** A function from the nltk.tokenize module that splits a string into individual words (tokens).

1. Define File Paths

Code:

cleanedtext\_file\_path = 'D:/AIUB/Sentiment Analysis Research/Data Sets/Cleaned Text Hotel Reviews.xlsx'

tokenized\_file\_path = 'D:/AIUB/Sentiment Analysis Research/Data Sets/Tokenized Hotel Reviews.xlsx'

* **cleanedtext\_file\_path:** The file path of the Excel file that contains the cleaned hotel reviews.
* **tokenized\_file\_path:** The file path where the tokenized reviews will be saved as a new Excel file.

1. Load the Cleaned Reviews Data

Code:

cleaned\_reviews = pd.read\_excel(cleanedtext\_file\_path)

* **pd.read\_excel():** This function reads the Excel file into a Pandas DataFrame. Here, the cleaned hotel reviews are loaded from the specified cleanedtext\_file\_path.
* **cleaned\_reviews:** A variable that holds the DataFrame containing the cleaned hotel reviews.

1. Tokenize the Cleaned Reviews

Code:

cleaned\_reviews['Tokenized Reviews'] = cleaned\_reviews['Cleaned Reviews'].apply(word\_tokenize)

* **cleaned\_reviews['Tokenized Reviews']:** This creates a new column named Tokenized Reviews in the DataFrame.
* .**apply(word\_tokenize**): This applies the word\_tokenize function to every row in the Cleaned Reviews column. The function splits the text in each row into individual words (tokens).
* **word\_tokenize:** A function from nltk that breaks down a string into a list of words.
* **For example:**

Original text: "This is a hotel review."

Tokenized text: ["This", "is", "a", "hotel", "review", "."]

1. Save the Tokenized Reviews to a New Excel File

Code:

cleaned\_reviews[['Tokenized Reviews']].to\_excel(tokenized\_file\_path, index=False, sheet\_name='Dhaka Regency')

* **cleaned\_reviews[['Tokenized Reviews']]:** This selects only the Tokenized Reviews column from the DataFrame to be saved to the Excel file.
* **.to\_excel():** This function saves the DataFrame to the Excel file specified by tokenized\_file\_path.
* **index=False:** Ensures that row numbers (indices) are not written to the Excel file.
* **sheet\_name='Dhaka Regency':** Sets the name of the sheet in the Excel file to "Dhaka Regency".

1. Print Confirmation

Code:

print(f"Tokenized saved to {tokenized\_file\_path}")

* **print(f"..."):** This prints a confirmation message to the console, indicating that the tokenized reviews have been saved to the specified file path.

Summary of Workflow

1. Load the cleaned reviews from an Excel file.
2. Tokenize each review by splitting it into individual words using nltk.word\_tokenize.
3. Save the tokenized reviews to a new Excel file, with a specific sheet name.