1. Import Required Libraries

Code:

import nltk

import pandas as pd

from nltk.corpus import stopwords

* **nltk:** Natural Language Toolkit, a library used for text processing in Python.
* **pandas:** A powerful data manipulation library used for handling structured data (like CSV or Excel files).
* **stopwords:** A list of common words (e.g., "the", "is", "and") that are often removed in text preprocessing as they don’t carry meaningful information.

1. Define File Paths

Code:

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

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

* **tokenized\_file\_path:** The file path where the tokenized hotel reviews are stored in an Excel file.
* **cleaned\_file\_path:** The file path where the cleaned reviews (after stopword removal) will be saved.

1. Load the Tokenized Reviews Data

Code:

tokenized\_reviews = pd.read\_excel(tokenized\_file\_path)

* **pd.read\_excel():** Loads the Excel file containing tokenized reviews into a Pandas DataFrame.
* **tokenized\_reviews:** This variable holds the loaded DataFrame that contains the tokenized hotel reviews.

1. Define Stopwords

Code:

stop\_words = set(stopwords.words('english'))

* **stopwords.words('english'**): Retrieves a list of common English stopwords (like "the", "is", etc.) from the nltk.corpus.stopwords module.
* **set():** Converts the list of stopwords into a set for faster lookups during stopword removal.

1. Define a Function to Remove Stopwords

Code:

def remove\_stopwords(tokens):

return [word for word in tokens if word not in stop\_words]

* **remove\_stopwords(tokens):** This is a custom function that takes a list of tokens (words) as input.
* **word for word in tokens if word not in stop\_words:** This is a list comprehension that iterates through each word in the tokens list and only keeps words that are not in the stopwords set.
* **Returns:** A list of words after removing stopwords.

1. Convert String Representation of List to Actual List and Remove Stopwords

Code:

tokenized\_reviews['Cleaned Reviews'] = tokenized\_reviews['Tokenized Reviews'].apply(lambda x: eval(x))

* **tokenized\_reviews['Tokenized Reviews']:** Refers to the column in the DataFrame that contains tokenized reviews, likely stored as strings representing lists of words.
* **eval(x):** Converts the string representation of a list (e.g., "[word1, word2]") into an actual Python list so that further processing can be done.
* .**apply(lambda x: eval(x)):** Applies the eval function to every row in the Tokenized Reviews column, converting each string into a list.

Code:

tokenized\_reviews['Dhaka Regency'] = tokenized\_reviews['Cleaned Reviews'].apply(remove\_stopwords)

* **tokenized\_reviews['Cleaned Reviews'].apply(remove\_stopwords):** Applies the remove\_stopwords function to each list in the Cleaned Reviews column, removing stopwords from each tokenized review.
* **tokenized\_reviews['Dhaka Regency']:** A new column created in the DataFrame to store the cleaned reviews (after stopword removal) for "Dhaka Regency".

1. Save the Cleaned Data to a New Excel File

Code:

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

* **[['Cleaned Reviews']]:** Selects only the Cleaned Reviews column (which contains the tokenized reviews after stopword removal).
* **.to\_excel():** Saves the DataFrame with cleaned reviews to the new Excel file specified by cleaned\_file\_path.
* **index=False:** Ensures that row indices are not included in the saved file.
* **sheet\_name='Dhaka Regency':** Specifies the name of the sheet in the Excel file as "Dhaka Regency".

1. Print Confirmation

Code:

print(f"Stop Words Removed saved to {cleaned\_file\_path}")

* **print(f"..."):** Prints a confirmation message indicating that the stopword-removed reviews have been saved to the specified file path.

Summary of Workflow

1. Load tokenized reviews from an Excel file.
2. Define a function to remove stopwords.
3. Apply the function to the tokenized reviews to remove stopwords.
4. Save the cleaned reviews (after stopword removal) to a new Excel file.