1. Import Required Libraries

Code:

import json

import pandas as pd

from sklearn.feature\_extraction.text import TfidfVectorizer

* **json:** A library to handle JSON data for saving the TF-IDF output.
* **pandas:** Used to read and manipulate the Excel file.
* **TfidfVectorizer:** A class from sklearn.feature\_extraction.text that converts a collection of text documents to a matrix of TF-IDF features, used in feature extraction.

1. Load the Lemmatized Dataset

Code:

file\_path = 'D:/AIUB/Sentiment Analysis Research/Data Sets/Lemmatized Hotel Reviews.xlsx'

lemmatized\_reviews = pd.read\_excel(file\_path)

* **file\_path:** Specifies the path where the lemmatized data is stored.
* **pd.read\_excel(file\_path):** Loads the Excel file into a Pandas DataFrame. This DataFrame contains the lemmatized hotel reviews.

1. Initialize the TF-IDF Vectorizer

Code:

tfidf\_vectorizer = TfidfVectorizer()

* **TfidfVectorizer():** Creates an instance of the TfidfVectorizer class. This will be used to convert the text data into a numerical matrix where each value represents the importance of a word in a review using the TF-IDF method.

1. Apply TF-IDF Vectorization

Code:

tfidf\_matrix = tfidf\_vectorizer.fit\_transform(lemmatized\_reviews['Lemmatized Reviews'])

* **lemmatized\_reviews['Lemmatized Reviews']:** Refers to the column in the DataFrame that contains the lemmatized reviews (processed text).
* **fit\_transform():** Fits the TfidfVectorizer to the data and transforms the lemmatized text into a sparse matrix of TF-IDF features. Each row in the matrix represents a document (review), and each column represents a word with its respective TF-IDF value.

1. Convert the TF-IDF Matrix to an Array

Code:

tfidf\_array = tfidf\_matrix.toarray()

* **toarray():** Converts the sparse matrix from the TF-IDF vectorization into a dense array. Each row in the array represents a document, and each column represents the TF-IDF value of a specific word.

1. Get the Feature Names (Words)

Code:

feature\_names = tfidf\_vectorizer.get\_feature\_names\_out()

* **get\_feature\_names\_out():** Retrieves the list of feature names (words) corresponding to the columns of the TF-IDF matrix. These are the words present in the lemmatized reviews.

1. Convert the TF-IDF Matrix to JSON

Code:

tfidf\_data = []

for doc\_index in range(tfidf\_array.shape[0]):

doc\_tfidf = {feature\_names[word\_index]: tfidf\_array[doc\_index, word\_index]

for word\_index in range(len(feature\_names))}

tfidf\_data.append(doc\_tfidf)

* **tfidf\_data = []:** Initializes an empty list to store the TF-IDF data for each document.
* **for doc\_index in range(tfidf\_array.shape[0]):** Iterates over each document (row) in the TF-IDF array.
* **doc\_tfidf = {feature\_names[word\_index]:** tfidf\_array[doc\_index, word\_index]}: Creates a dictionary where each word (from feature\_names) is a key, and its corresponding TF-IDF score is the value. This dictionary represents the TF-IDF scores for a single document.
* **tfidf\_data.append(doc\_tfidf):** Appends the dictionary to the list tfidf\_data for each document.

1. Convert the Data to JSON Format

Code:

tfidf\_json = json.dumps(tfidf\_data, indent=4)

* **json.dumps(tfidf\_data, indent=4):** Converts the list of dictionaries (TF-IDF scores for all documents) into a JSON-formatted string. The indent=4 argument makes the output more readable by adding indentation.

1. Save the JSON Data to a File

Code:

json\_file\_path = 'D:/AIUB/Sentiment Analysis Research/JSON/TFIDF Hotel Reviews.json'

with open(json\_file\_path, 'w') as json\_file:

json\_file.write(tfidf\_json)

* **json\_file\_path:** Specifies the path where the JSON file will be saved.
* **with open(json\_file\_path, 'w') as json\_file:** Opens the specified file in write mode ('w'). If the file does not exist, it will be created.
* **json\_file.write(tfidf\_json):** Writes the JSON string (tfidf\_json) to the file.

1. Print Confirmation

Code:

print(f"TF-IDF data saved to {json\_file\_path}")

* **print():** Outputs a confirmation message indicating that the TF-IDF data has been saved successfully.

Summary of Workflow

1. Load the lemmatized hotel reviews from the Excel file.
2. Apply the TF-IDF vectorizer to convert the reviews into a matrix of word importance scores.
3. Convert the matrix to an array and map the words to their respective scores.
4. Transform the TF-IDF scores into a JSON format.
5. Save the JSON data to a file for further use.