1. Library Imports

Code:

library(openxlsx)

library(readxl)

library(stringr)

library(tm)

* **openxlsx:** Used for reading from and writing to Excel files.
* **readxl:** Another library for reading Excel files.
* **stringr:** Used for string manipulations, like trimming spaces.
* **tm:** Text mining package used to preprocess the text corpus (e.g., removing punctuation, stopwords, stemming, etc.).

1. Reading Data

Code:

directory <- "D:/AIUB/Thesis/GIT/Dataset"

file\_name <- "Hotel Review Data Table.xlsx"

sheet\_index <- 1

file\_path <- file.path(directory, file\_name)

read\_data <- read\_excel(file\_path, sheet = sheet\_index)

dataset <- read\_data

View(dataset)

* Define directory where the dataset is located.
* Define name of the Excel file that contains the dataset.
* Refers to the sheet number in the Excel file that will be read.
* Combines the directory and file name into a full file path.
* Reads the specified Excel file and sheet into R.
* Assigns the data read to a variable called 'dataset'.
* View the dataset.

1. Missing Value Handling

Code:

any(is.na(dataset))

clean\_data <- na.omit(dataset)

* Checks if there are any missing values (NA) in the dataset. Returns TRUE if there are missing values.
* Removes rows with any missing values in the dataset.

1. Special Character Removal function

Code:

removeSpecialChars <- content\_transformer(function(x) gsub("[^[:alnum:][:space:]]", "", x))

* **content\_transformer:** This is a function provided by the tm (text mining) package in R. It is used to apply custom transformations to text data. It allows us to create a custom transformation function to be applied to each text element in a corpus.
* **function(x):** This defines an anonymous function in R. The variable x will represent the text input (such as a string or document) that is passed to this function.
* **gsub:** This is a built-in R function used to replace (substitute) patterns in text. In this case, it's being used to remove certain characters from the text. The syntax of gsub is: gsub(pattern, replacement, text)
* **pattern:** A regular expression that defines what you're looking for in the text.
* replacement: The text that will replace the matched pattern. Here, it's an empty string "", meaning the matching pattern will be removed.
* **text:** The actual text (x) that you are performing the substitution on.
* **[^[:alnum:][:space:]]:** This is the regular expression pattern used to identify characters to remove:
* **[:alnum:]:** This is a POSIX character class, which matches any alphanumeric characters (letters and digits). So [:alnum:] matches letters (A-Z, a-z) and numbers (0-9).
* **[:space:]:** This matches any whitespace characters, including spaces, tabs, and line breaks.
* **[^...]:** The ^ inside square brackets negates the pattern, meaning it will match anything except what is specified inside the brackets.
* [^[:alnum:][:space:]] means "anything that is not an alphanumeric character or space."
* "", x. The second argument to gsub is "", which means that any special characters found by the pattern will be replaced with an empty string, i.e., they will be removed. The third argument x is the input text that the function is working on.

1. Text Processing

Code:

corpus <- Corpus(VectorSource(dataset$Review))

corpus <- tm\_map(corpus, content\_transformer(tolower))

corpus <- tm\_map(corpus, removePunctuation)

corpus <- tm\_map(corpus, removeNumbers)

corpus <- tm\_map(corpus, removeSpecialChars)

corpus <- tm\_map(corpus, removeWords, stopwords("english"))

corpus <- tm\_map(corpus, stemDocument)

corpus <- tm\_map(corpus, stripWhitespace)

* Creates a text corpus from the 'Review' column of the dataset.
* Converts all text to lowercase to ensure uniformity.
* Removes punctuation marks from the text.
* Removes numbers from the text.
* Removes any special characters.
* Removes common stopwords (e.g., "the", "is", "and") from the text.
* Reduces words to their root form (e.g., "running" becomes "run").
* Removes any extra spaces that may have been introduced.

1. Trimming

Code:

corpus$Review <- str\_trim(corpus$Review)

* Removes any leading or trailing white spaces from the 'Review' column of the corpus.

1. Corpus to DataFrame

Code:

corpus\_df <- data.frame(Text = sapply(corpus, as.character), stringsAsFactors = FALSE)

* **data.frame():** This function creates a data frame in R, which is a table-like structure where columns can contain different types of data (numeric, character, etc.). In this case, you're creating a data frame from the corpus.
* Inside the data.frame() function, you provide a column called Text that will hold the text data from the corpus.
* **sapply():** This function applies a specified function to each element of a list or vector. It simplifies the output into a vector or matrix if possible. In this case, you're applying a function to each element of the corpus (which is a collection of text documents).
* **corpus:** This is the corpus object, which contains the preprocessed text documents (after converting to lowercase, removing punctuation, stop words, etc.). Each element in the corpus represents a single document or review.
* **as.character:** This is the function that is applied to each element of the corpus using sapply(). It converts each document in the corpus into a character string (text). By default, each document in the corpus might be represented as a more complex structure, so as.character ensures that each document is treated as plain text.
* **stringsAsFactors = FALSE:** In R, by default, character strings can be converted to factors (a data type used for categorical data). However, in this case, you don't want the text to be treated as categorical data (factors), so you set stringsAsFactors = FALSE to keep the text as plain character strings. This ensures that the Text column remains as text and doesn't get converted into factors.

1. Exporting the Processed Data

Code:

write.xlsx(corpus\_df, file = "D:/AIUB/Thesis/GIT/Dataset/Processed Data.xlsx", sheetName = "Dhaka Regency", rowNames = FALSE)

* Saves the cleaned text data into a new Excel file at the specified location with the sheet name "Dhaka Regency". No row names are included.