**Practical 7: Measuring Similarity Among Documents and Detecting Passages Which Have Been Reused**

Codes:

# Install necessary packages

install.packages("tm")

require("tm")

install.packages("ggplot2")

install.packages("textreuse")

install.packages("devtools")

# Load in corpus and preprocess text

my.corpus <- Corpus(DirSource("C:/Users/asif0/Documents/New folder"))  # Load in corpus from directory

my.corpus <- tm\_map(my.corpus, removeWords, stopwords("english"))  # Remove stop words from corpus

# Create term-document matrix

my.tdm <- TermDocumentMatrix(my.corpus)  # Create term-document matrix from corpus

#inspect(my.tdm)  # Inspect term-document matrix (optional)

# Create document-term matrix

my.dtm <- DocumentTermMatrix(my.corpus, control = list(weighting = weightTfIdf, stopwords = TRUE))  # Create document-term matrix from corpus, using TF-IDF weighting and removing stop words

#inspect(my.dtm)  # Inspect document-term matrix (optional)

# Convert document-term matrix to data frame and scale data

my.df <- as.data.frame(inspect(my.tdm))  # Convert document-term matrix to data frame

my.df.scale <- scale(my.df)  # Scale data using z-score normalization

# Perform hierarchical clustering and plot dendrogram

d <- dist(my.df.scale, method = "euclidean")  # Calculate distance matrix using Euclidean distance

fit <- hclust(d, method = "ward")  # Perform hierarchical clustering using Ward's method

plot(fit)  # Plot dendrogram

OUTPUT:

