**Text Mining Report: Preprocessing and Analysis**

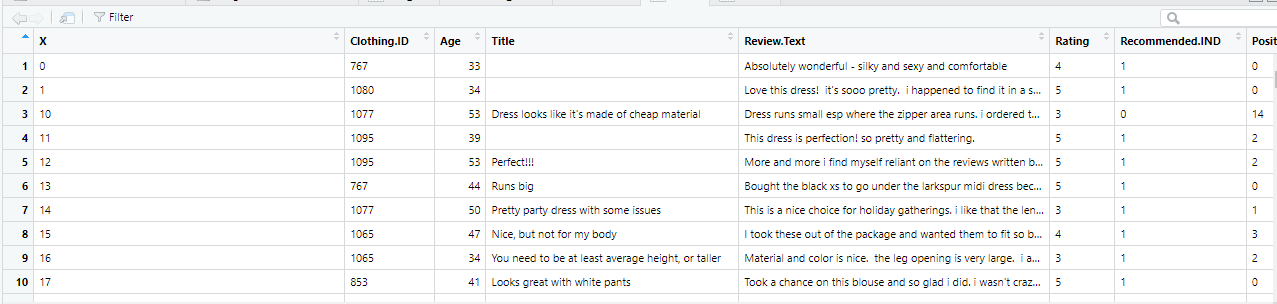
1. **Introduction**

This report outlines the step-by-step approach taken to preprocess and analyze the text data, using techniques such as stopword removal, stemming, and frequency analysis.

**2. Data Loading**

The dataset is read into R using the clipboard:

text <- read.csv("clipboard", sep = "\t", header = TRUE)



**3. Corpus Creation**

TextDoc <- Corpus(VectorSource(text$Review.Text))

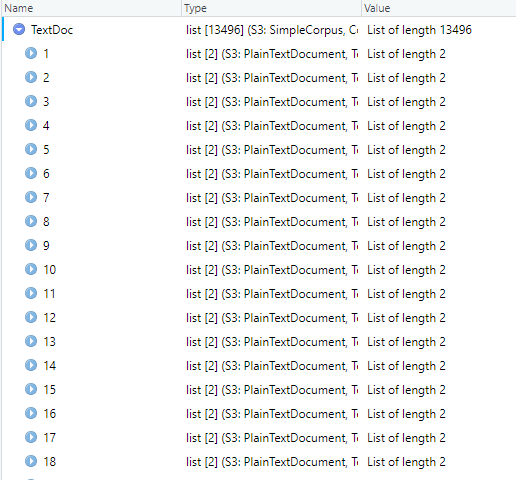
**4. Stopword Removal**

default\_stopwords <- stopwords()

words\_to\_exclude <- c("not", "doesn't")

custom\_stopwords <- setdiff(default\_stopwords, words\_to\_exclude)

TextDoc <- tm\_map(TextDoc, removeWords, custom\_stopwords)



**Replacing special characters with space**

toSpace <- content\_transformer(function (x , pattern ) gsub(pattern, " ", x))

TextDoc <- tm\_map(TextDoc, toSpace, "/")

TextDoc <- tm\_map(TextDoc, toSpace, "@")

TextDoc <- tm\_map(TextDoc, toSpace, "\\|")

TextDoc <- tm\_map(TextDoc, toSpace, "!")

TextDoc <- tm\_map(TextDoc, toSpace, "#")

TextDoc <- tm\_map(TextDoc, toSpace, "%")

TextDoc <- tm\_map(TextDoc, content\_transformer(tolower)) ## lower case removal

TextDoc <- tm\_map(TextDoc, removeNumbers) ## Remove numbers

TextDoc <- tm\_map(TextDoc, removeWords, c("dresses","dress","top", "not", "wear", "general", "petit", "just", "this", "one", "will", "well", "can", "back", "bit", "shirt","sizes")) # Remove your own stop word

TextDoc <- tm\_map(TextDoc, removePunctuation) ## Remove punctuation

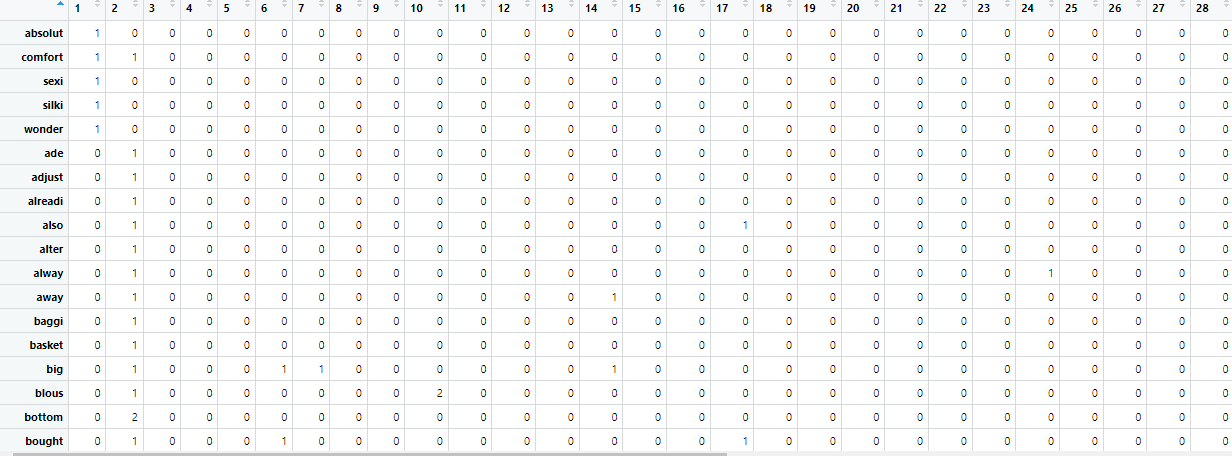
TextDoc <- tm\_map(TextDoc, stripWhitespace) ## Eliminate extra white spaces

TextDoc <- tm\_map(TextDoc, stemDocument) ##Text stemming

**Building a term-document matrix**

TextDoc\_dtm <- TermDocumentMatrix(TextDoc)

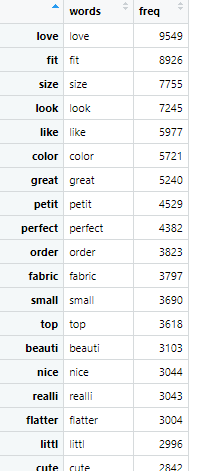
dtm\_m <- as.matrix(TextDoc\_dtm)

****

**Sort by decreasing value of frequency**

dtm\_v <- sort(rowSums(dtm\_m),decreasing=TRUE)

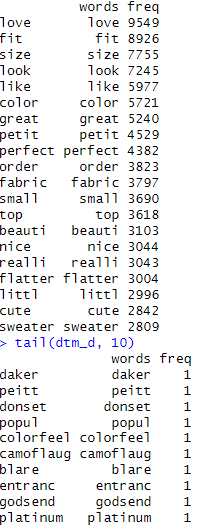
dtm\_d <- data.frame(words = names(dtm\_v),freq = dtm\_v)



**Display the top 10 most frequent words**

head(dtm\_d, 20

tail(dtm\_d, 10)

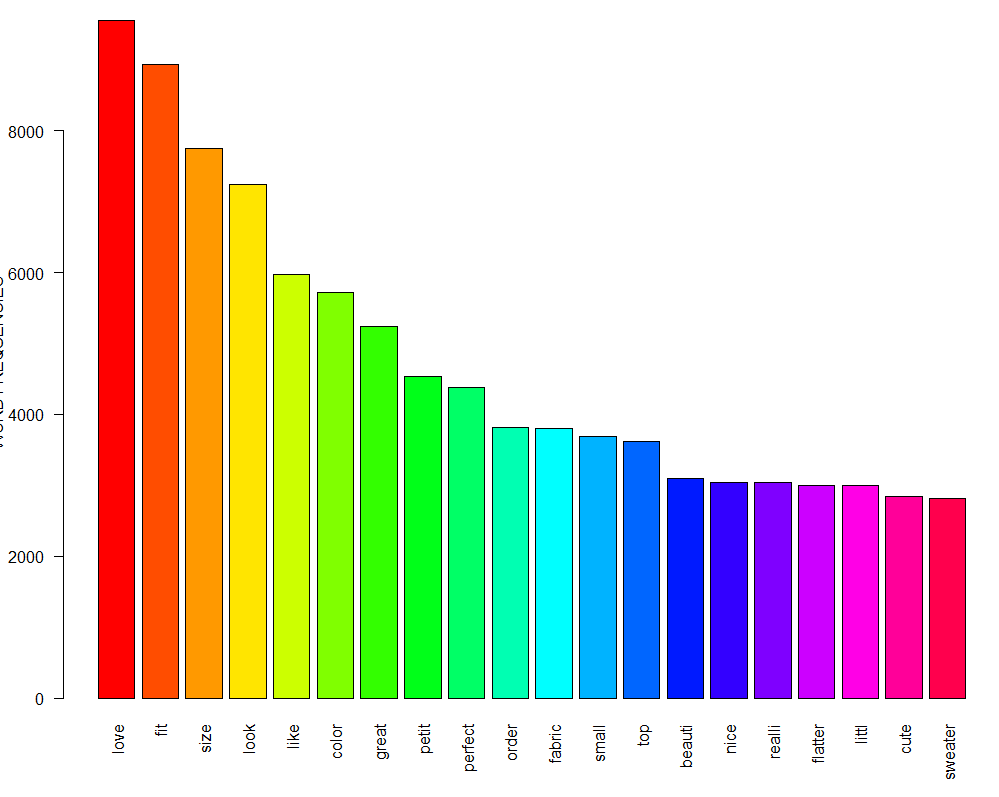


**Barplot**

barplot(dtm\_d[1:20,]$freq, las = 2, names.arg = dtm\_d[1:20,]$words,

col = rainbow(20), main ="Top 20 most frequent words",

ylab = "WORD FREQUENCIES")### las shows legacy axis = 2 vertical



**Wordcloud**

wordcloud2(dtm\_d, size = .4, shape = 'circle', rotateRatio = 0.9, minSize = .5)



**Findings**

The most commonly mentioned words in the reviews include:

* **"Love"** – Indicates customer satisfaction and appreciation for the products.
* **"Fit" & "Size"** – Customers frequently discuss the fit and sizing of the clothing, suggesting that these factors are crucial in purchasing decisions.
* **"Look" & "Color"** – Many customers focus on the aesthetic appeal and color of the garments.
* **"Great" & "Perfect"** – These words reflect positive sentiments, with customers expressing delight in their purchases.

**Conclusion**

**Sizing Consistency:** Since "fit" and "size" appear frequently, the brand could consider providing more detailed size guides or customer-fit reviews to help buyers make informed decisions.

**Aesthetic Appeal Matters:** The prominence of words like "look" and "color" suggests that customers value the appearance of the clothing, making it essential to offer a variety of stylish options.

**Customer Satisfaction is High:** The repeated use of positive words like "love" and "great" indicates overall customer satisfaction with the products.