Assignment 5.b

Athul Jose P
11867566

School of Electrical Engineering and Computer Science

Washington State University

CptS 575 Data Science

1.

```r
# Load necessary libraries
library(readr)       # For reading the CSV file
library(tm)          # For text mining and preprocessing
```

```
Loading required package: NLP


Attaching package: 'NLP'


The following object is masked from 'package:ggplot2':

    annotate
```

```r
library(SnowballC)   # For stemming
library(tokenizers)  # For tokenization
library(quanteda)    # For creating Document-Term Matrix
```

```
Package version: 4.1.0
Unicode version: 15.1
ICU version: 74.1

Parallel computing: 32 of 32 threads used.

See https://quanteda.io for tutorials and examples.


Attaching package: 'quanteda'


The following object is masked from 'package:tm':

    stopwords


The following objects are masked from 'package:NLP':

    meta, meta<-
```

```r
# Step 1: Load the Dataset
data <- read.csv("bbc.csv")

# Check the structure of the dataset
str(data)
```

```
'data.frame':   2225 obs. of  2 variables:
 $ category: chr  "business" "business" "business" "business" ...
 $ text    : chr  "Ad sales boost Time Warner profit\n\nQuarterly profits at US media gi
```

```r
# Step 2: Preprocessing the Text
# Convert the text to lowercase, remove punctuation and numbers, and perform stemming
corpus <- Corpus(VectorSource(data$text))  # Create corpus

corpus <- tm_map(corpus, content_transformer(tolower))    # Convert to lowercase
```

```
Warning in tm_map.SimpleCorpus(corpus, content_transformer(tolower)):
transformation drops documents
```

```r
corpus <- tm_map(corpus, removePunctuation)               # Remove punctuation
```

```
Warning in tm_map.SimpleCorpus(corpus, removePunctuation): transformation drops
documents
```

```r
corpus <- tm_map(corpus, removeNumbers)                   # Remove numbers
```

```
Warning in tm_map.SimpleCorpus(corpus, removeNumbers): transformation drops
documents
```

```r
corpus <- tm_map(corpus, removeWords, stopwords("english")) # Remove stop words
```

```
Warning in tm_map.SimpleCorpus(corpus, removeWords, stopwords("english")):
transformation drops documents
```

```r
corpus <- tm_map(corpus, stemDocument)                    # Perform stemming
```

```
Warning in tm_map.SimpleCorpus(corpus, stemDocument): transformation drops
documents
```

```
# Step 3: Create Document-Term Matrix
dtm <- DocumentTermMatrix(corpus)

# Check the dimensions of the matrix
print(dim(dtm))
```

```
[1]  2225 21221
```

```
# Step 4: Remove low-frequency words (15% least frequent terms)
term_frequency <- colSums(as.matrix(dtm))
sorted_terms <- sort(term_frequency, decreasing = TRUE)

# Keep only the top 85% of terms
threshold <- quantile(sorted_terms, 0.85)
dtm_filtered <- dtm[, which(term_frequency >= threshold)]

# Step 5: Display words from the 2205th article with frequency >= 4
article_2205 <- as.matrix(dtm_filtered[2205, ])
feature_vector <- article_2205[article_2205 >= 4]
print(feature_vector)
```

```
 [1] 6 6 4 4 4 4 4 4 4 8 4 5 4 5
```

2.

```
# loading MASS library and Boston dataset
library(MASS)
```

```
Attaching package: 'MASS'

The following object is masked from 'package:dplyr':

    select
```

```
data("Boston")

# predictors
predictors <- setdiff(names(Boston), "crim")
predictors
```

```
 [1] "zn"      "indus"   "chas"    "nox"     "rm"      "age"     "dis"    
 [8] "rad"     "tax"     "ptratio" "black"   "lstat"   "medv"   
```