Two-Movies

path\_one = "~/OneDrive - MNSCU/myGithub/Unsupervised\_Learning/Movie-Script-Unsupervised-Learning-Methods-Analyses/One\_movie"  
dir\_1 = DirSource(paste(path\_one, sep=""), encoding = "UTF-8")  
corpus\_1 = Corpus(dir\_1)  
head(summary(corpus\_1))

## Length Class Mode  
## A-Few-Good-Men.txt 2 PlainTextDocument list

path\_second = "~/OneDrive - MNSCU/myGithub/Unsupervised\_Learning/Movie-Script-Unsupervised-Learning-Methods-Analyses/Second\_movie"  
dir\_2 = DirSource(paste(path\_second, sep=""), encoding = "UTF-8")  
corpus\_2 = Corpus(dir\_2)  
head(summary(corpus\_2))

## Length Class Mode  
## A-Prayer-Before-Dawn.txt 2 PlainTextDocument list

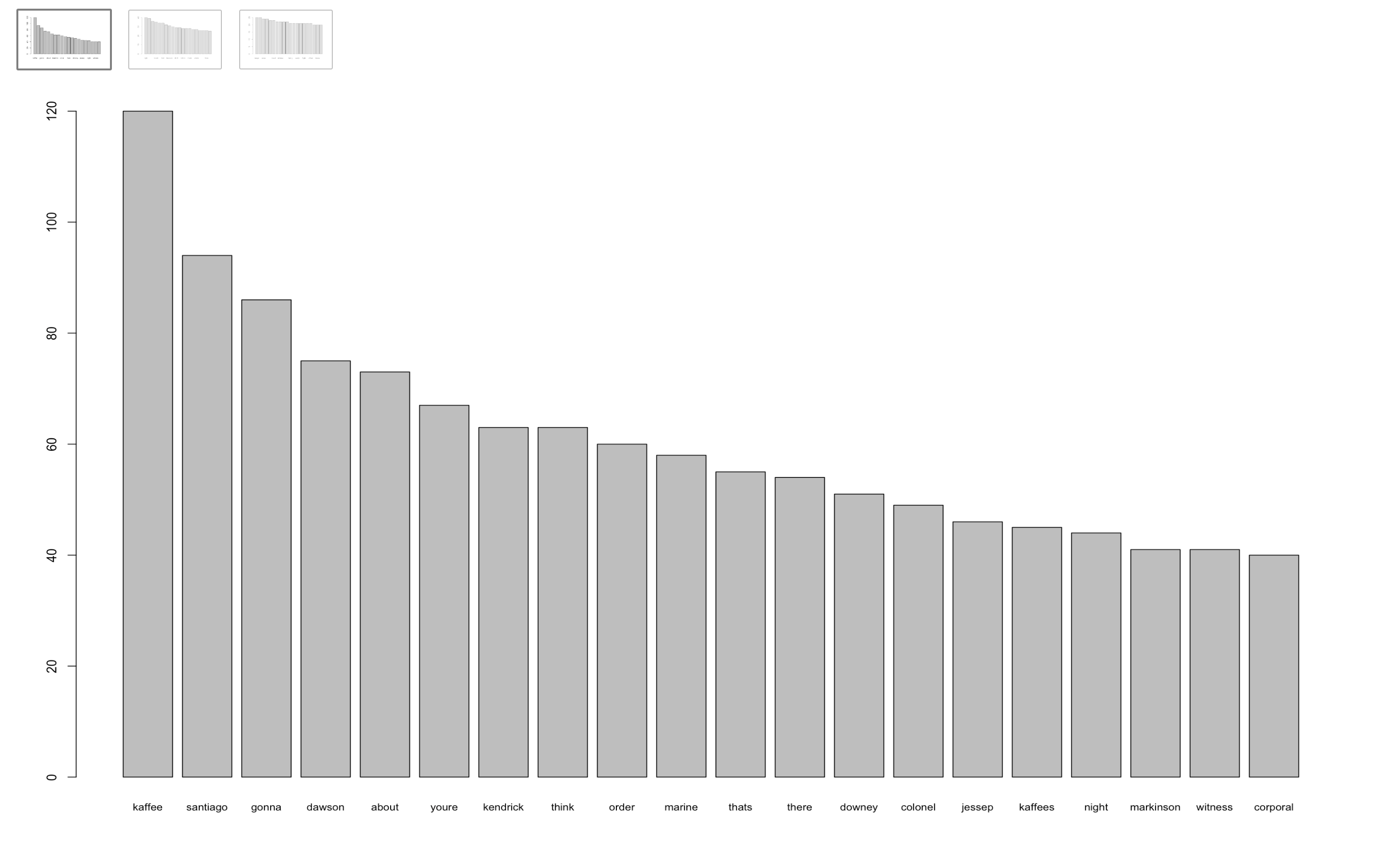
ndocs <- length(corpus\_1)  
# ignore extremely rare words i.e. terms that appear in less then 1% of the documents  
minTermFreq <- ndocs \* 0.05  
# ignore overly common words i.e. terms that appear in more than 50% of the documents  
maxTermFreq <- ndocs \* .5  
dtm\_1 = DocumentTermMatrix(corpus\_1,  
 control = list(  
 stopwords = ("en"),   
 wordLengths=c(5, 15),  
 removePunctuation = T,  
 removeNumbers = T  
 #stemming = T,  
 #removeWords("bateman"),  
 #bounds = list(global = c(minTermFreq, maxTermFreq))  
 ))

ndocs <- length(corpus\_2)  
# ignore extremely rare words i.e. terms that appear in less then 1% of the documents  
minTermFreq <- ndocs \* 0.05  
# ignore overly common words i.e. terms that appear in more than 50% of the documents  
maxTermFreq <- ndocs \* .5  
dtm\_2 = DocumentTermMatrix(corpus\_2,  
 control = list(  
 stopwords = ("end"),   
 wordLengths=c(5, 15),  
 removePunctuation = T,  
 removeNumbers = T  
 #stemming = T,  
 #removeWords("bateman"),  
 #bounds = list(global = c(minTermFreq, maxTermFreq))  
 ))

### Two movies

tdm\_1 = t(dtm\_1)  
tdm\_2 = t(dtm\_2)  
#Removing sparse terms  
tdm\_no\_sparse\_1 = removeSparseTerms(tdm\_1, sparse = .99)  
tdm.mat\_1 <- as.matrix(tdm\_no\_sparse\_1)  
tdm\_no\_sparse\_2 = removeSparseTerms(tdm\_2, sparse = .99)  
tdm.mat\_2 <- as.matrix(tdm\_no\_sparse\_2)

#First 20 most frequent words for Move#1  
word.freq\_1 = sort(rowSums(tdm.mat\_1), decreasing = T)  
barplot(word.freq\_1[1:20], cex.names = .8)



#Next 20 most frequent words  
barplot(word.freq\_1[21:40], cex.names = .8)

A screenshot of a cell phone

Description automatically generated

#Next 20 most frequent words  
barplot(word.freq\_1[41:60], cex.names = .8)

A screenshot of a cell phone

Description automatically generated

dim(tdm.mat\_1)

## [1] 2376 1

#Word Cloud for Movie#1  
#wordcloud(words = names(word.freq\_1[4:2376]), freq = word.freq\_check, min.freq = 100, col = rainbow(1000), max.words = 50)

Move#1 A Few Good Men

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA close up of a map

Description automatically generated

#First 20 most frequent words for Movie#2  
word.freq\_2 = sort(rowSums(tdm.mat\_2), decreasing = T)  
barplot(word.freq\_2[4:20], cex.names = .8)

A close up of a logo

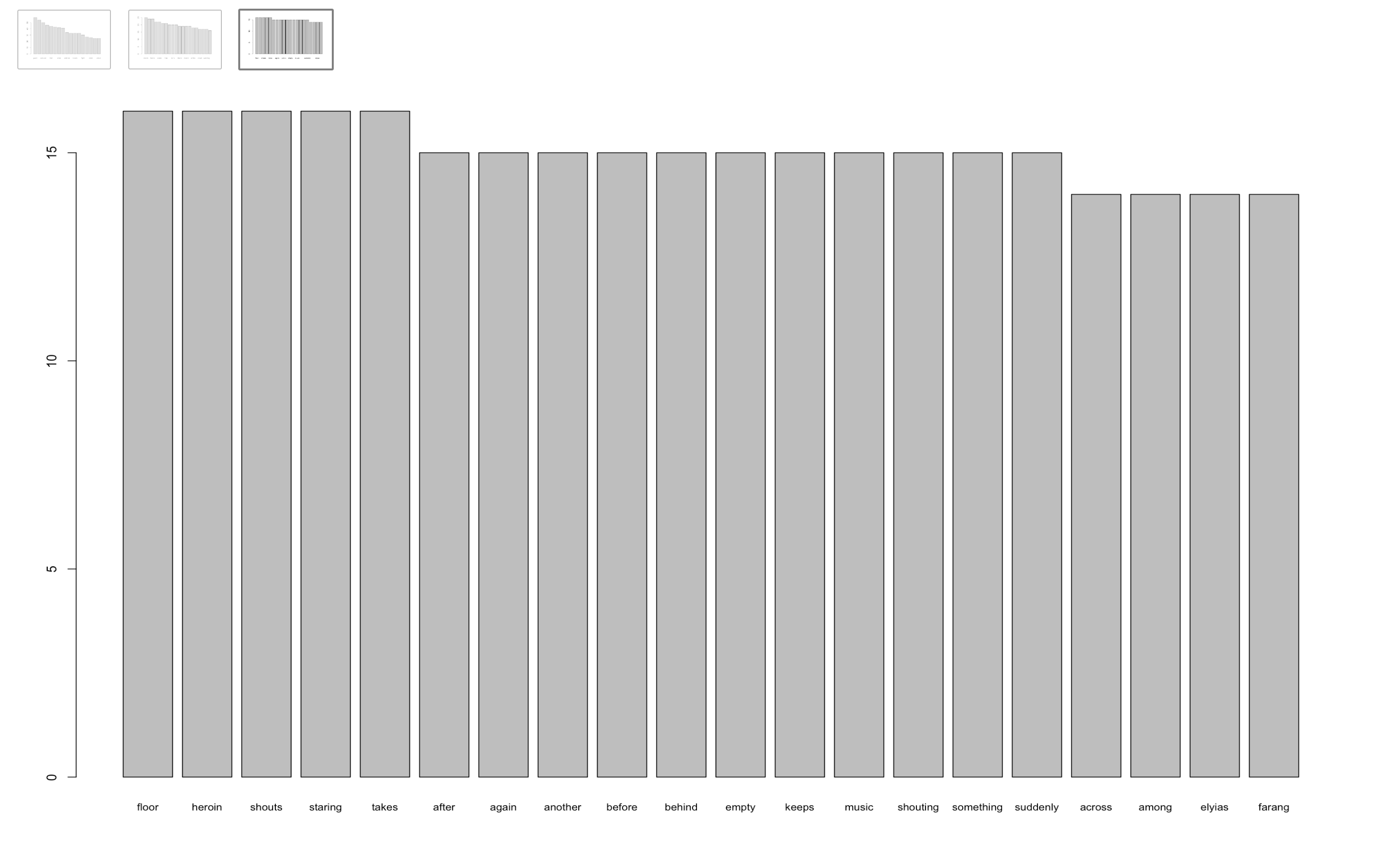
Description automatically generated

#Next 20 most frequent words  
barplot(word.freq\_2[21:40], cex.names = .8)

A screenshot of a cell phone

Description automatically generated

#Next 20 most frequent words  
barplot(word.freq\_2[41:60], cex.names = .8)



dim(tdm.mat\_2)

## [1] 2544 1

#Word Cloud for Movie#2  
#wordcloud(words = names(word.freq\_2[4:2544]), freq = word.freq\_check, min.freq = 100, col = rainbow(1000), max.words = 50)

Movie #2 A Prayer Before Dawn

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generatedA close up of a map

Description automatically generated