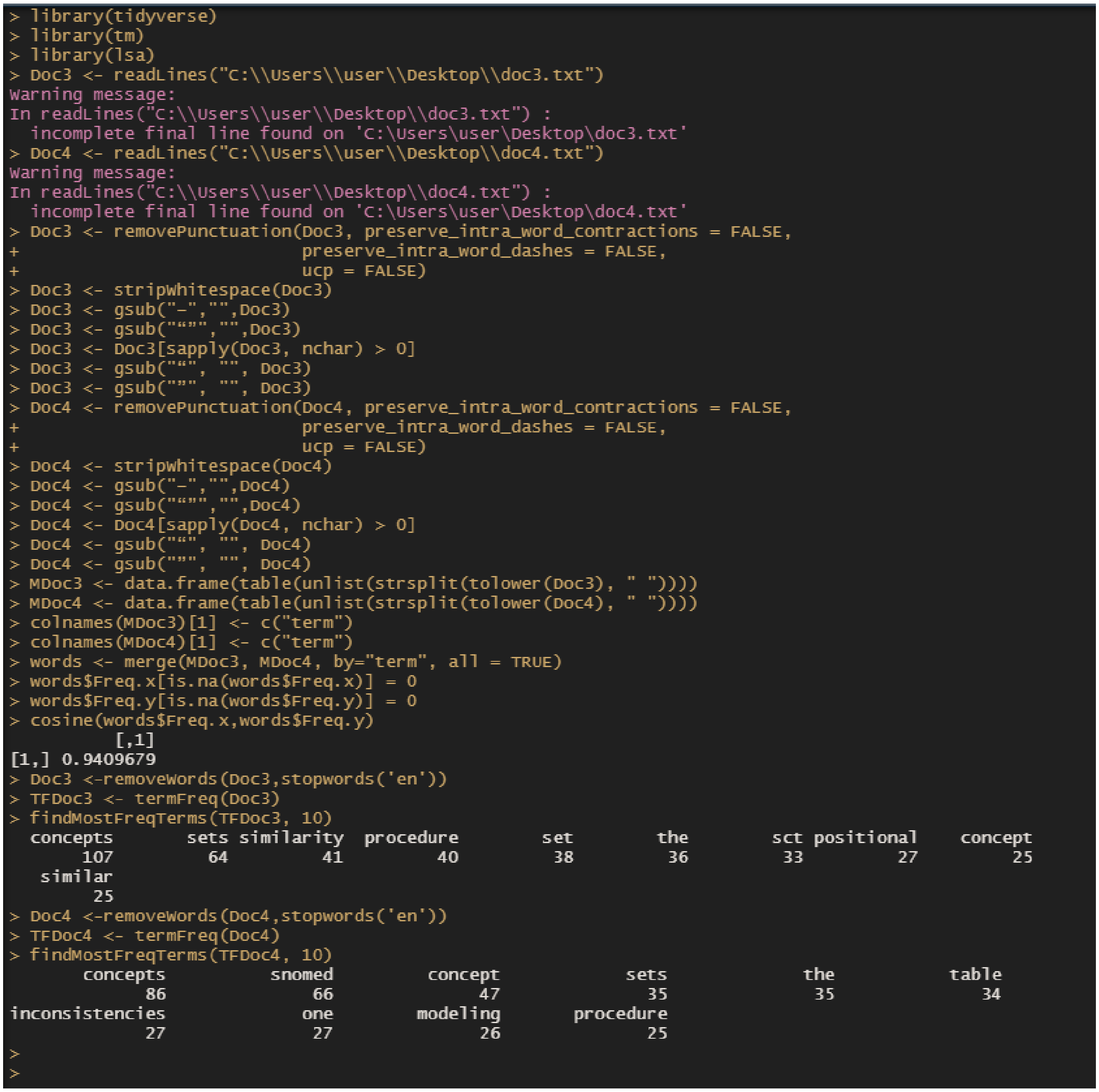
Giorgi Samushia
Andrew Kulp
CMPT 363
Project 01



```
> library(tidyverse)
> library(tm)
> library(lsa)
> Doc3 <- readLines("C:\\Users\\user\\Desktop\\doc3.txt")
Warning message:
In readLines("C:\\Users\\user\\Desktop\\doc3.txt") :
  incomplete final line found on 'C:\Users\user\Desktop\doc3.txt'
> Doc4 <- readLines("C:\\Users\\user\\Desktop\\doc4.txt")
Warning message:
In readLines("C:\\Users\\user\\Desktop\\doc4.txt") :
  incomplete final line found on 'C:\Users\user\Desktop\doc4.txt'
> Doc3 <- removePunctuation(Doc3, preserve_intra_word_contractions = FALSE,
+                           preserve_intra_word_dashes = FALSE,
+                           ucp = FALSE)
> Doc3 <- stripWhitespace(Doc3)
> Doc3 <- gsub("–","",Doc3)
> Doc3 <- gsub("“”","",Doc3)
> Doc3 <- Doc3[sapply(Doc3, nchar) > 0]
> Doc3 <- gsub("“", "", Doc3)
> Doc3 <- gsub("”", "", Doc3)
> Doc4 <- removePunctuation(Doc4, preserve_intra_word_contractions = FALSE,
+                           preserve_intra_word_dashes = FALSE,
+                           ucp = FALSE)
> Doc4 <- stripWhitespace(Doc4)
> Doc4 <- gsub("–","",Doc4)
> Doc4 <- gsub("“”","",Doc4)
> Doc4 <- Doc4[sapply(Doc4, nchar) > 0]
> Doc4 <- gsub("“", "", Doc4)
> Doc4 <- gsub("”", "", Doc4)
> MDoc3 <- data.frame(table(unlist(strsplit(tolower(Doc3), " "))))
> MDoc4 <- data.frame(table(unlist(strsplit(tolower(Doc4), " "))))
> colnames(MDoc3)[1] <- c("term")
> colnames(MDoc4)[1] <- c("term")
> words <- merge(MDoc3, MDoc4, by="term", all = TRUE)
> words$Freq.x[is.na(words$Freq.x)] = 0
> words$Freq.y[is.na(words$Freq.y)] = 0
> cosine(words$Freq.x,words$Freq.y)
          [,1]
[1,] 0.9409679
> Doc3 <-removeWords(Doc3,stopwords('en'))
> TFDoc3 <- termFreq(Doc3)
> findMostFreqTerms(TFDoc3, 10)
  concepts       sets similarity  procedure        set        the        sct positional    concept
       107         64         41         40         38         36         33         27         25
   similar
        25
> Doc4 <-removeWords(Doc4,stopwords('en'))
> TFDoc4 <- termFreq(Doc4)
> findMostFreqTerms(TFDoc4, 10)
       concepts          snomed         concept            sets             the           table
             86              66              47              35              35              34
inconsistencies             one        modeling       procedure
             27              27              26              25
>
>
```