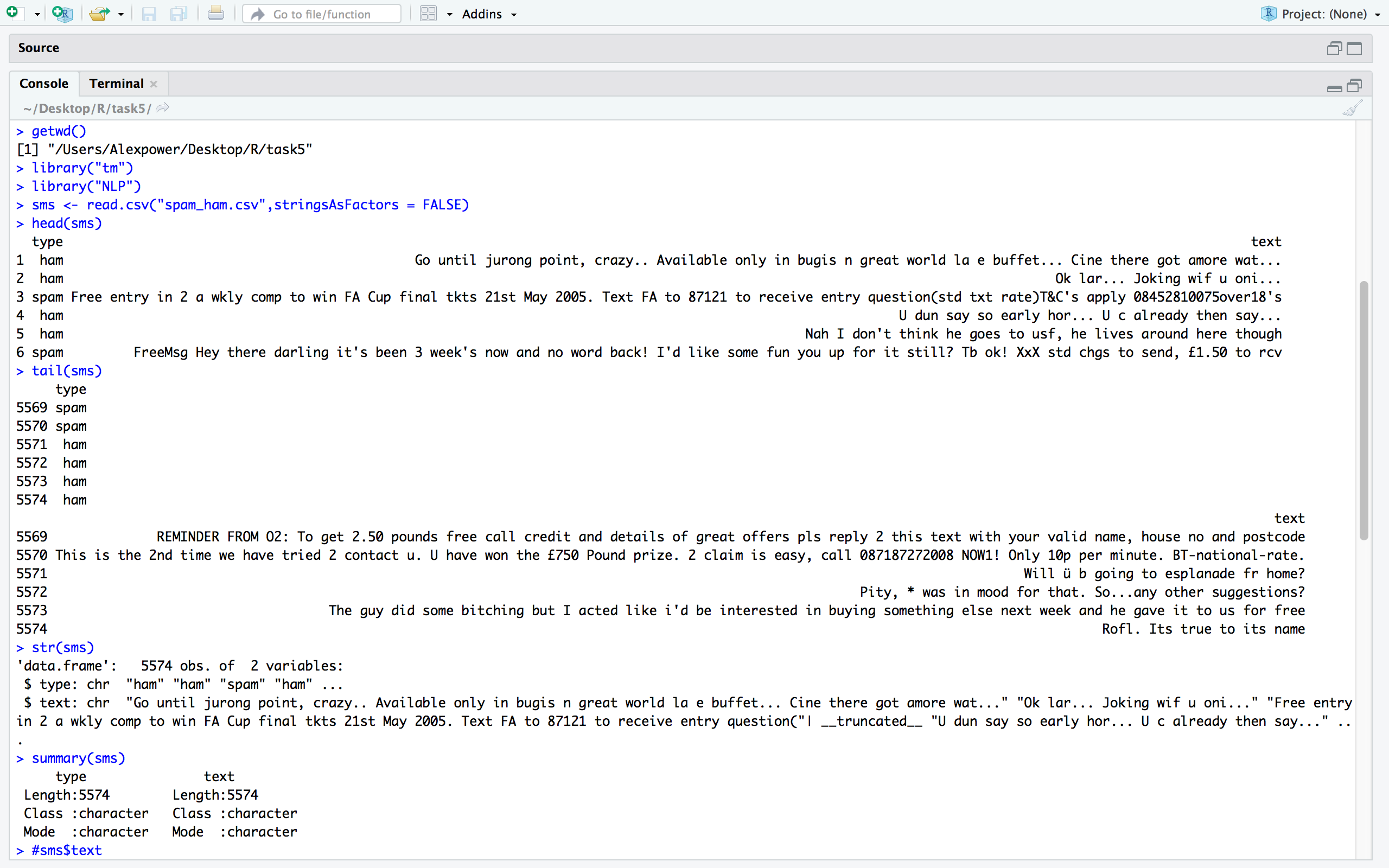
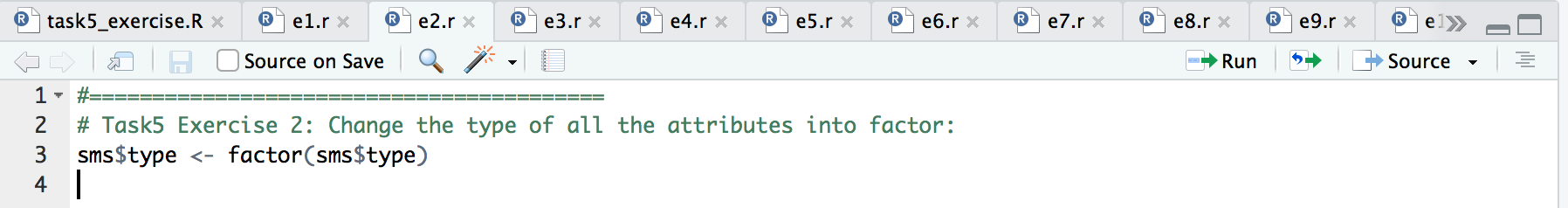
# CS4112/CS54111 INTRODUCTION TO DATA SCIENCE

# Exercise.1: Get a ham/spam SMS file :

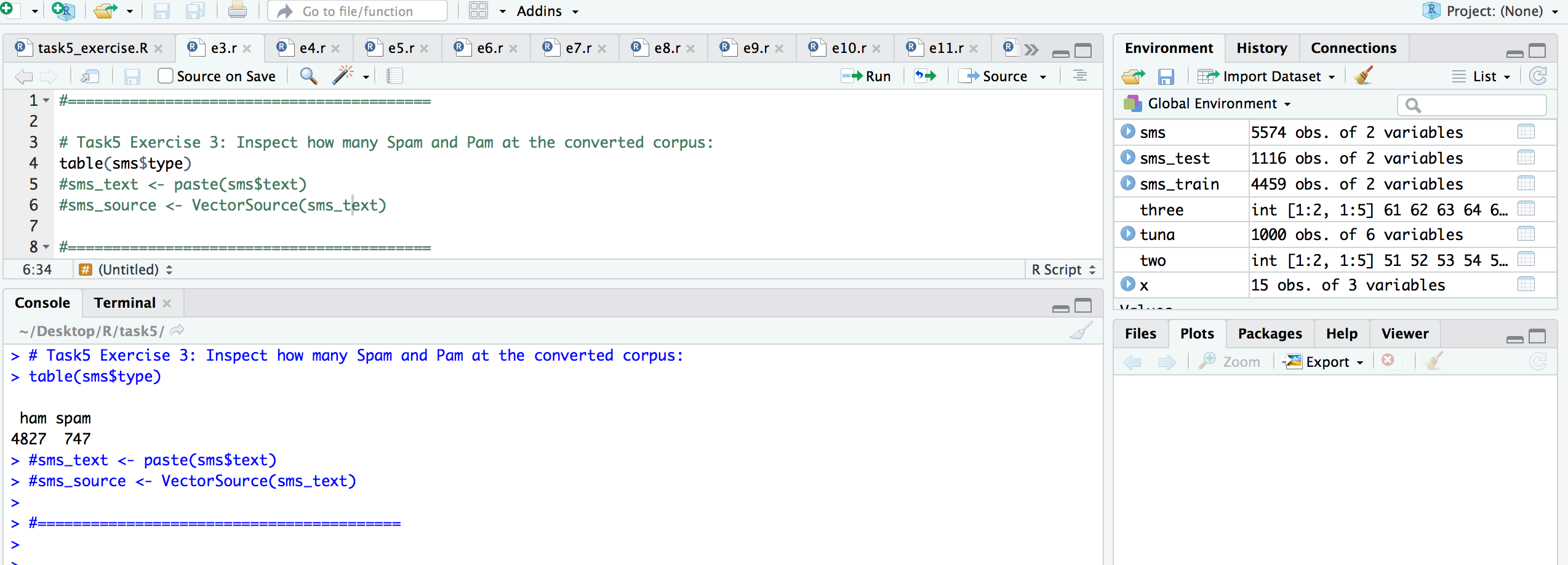




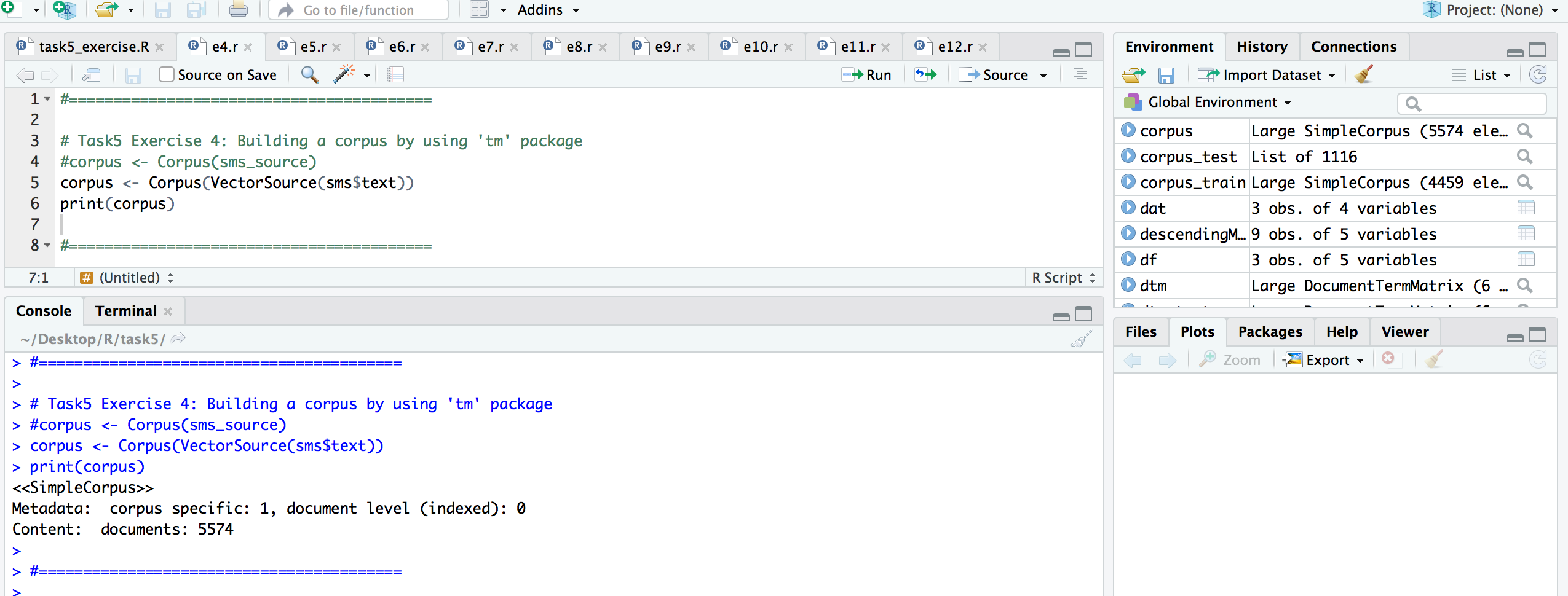
**Exercise.2: Change the type of all the attributes into factor:**



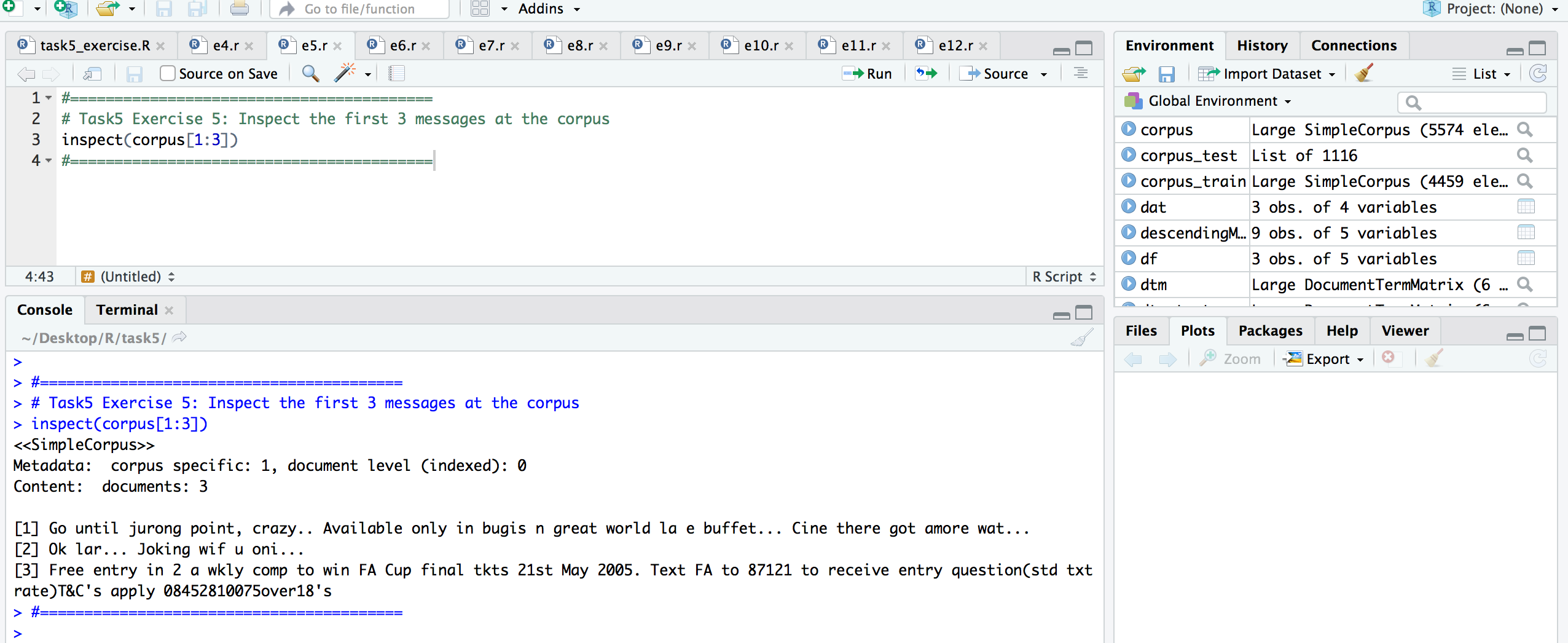
**Exercise.3: Inspect how many Spam and Pam at the converted corpus:**



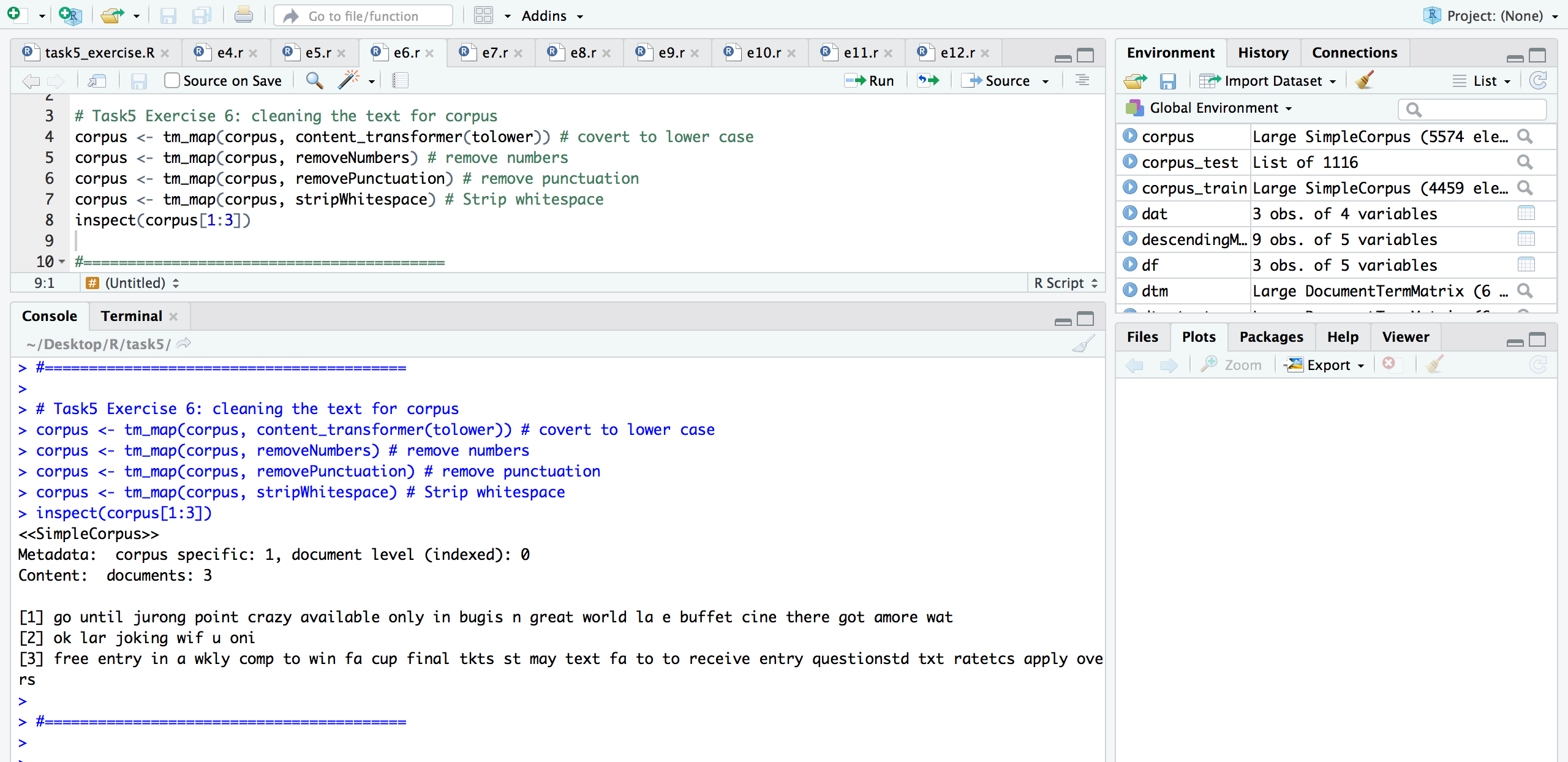
**Exercise.4: Building a corpus by using 'tm' package :**

****

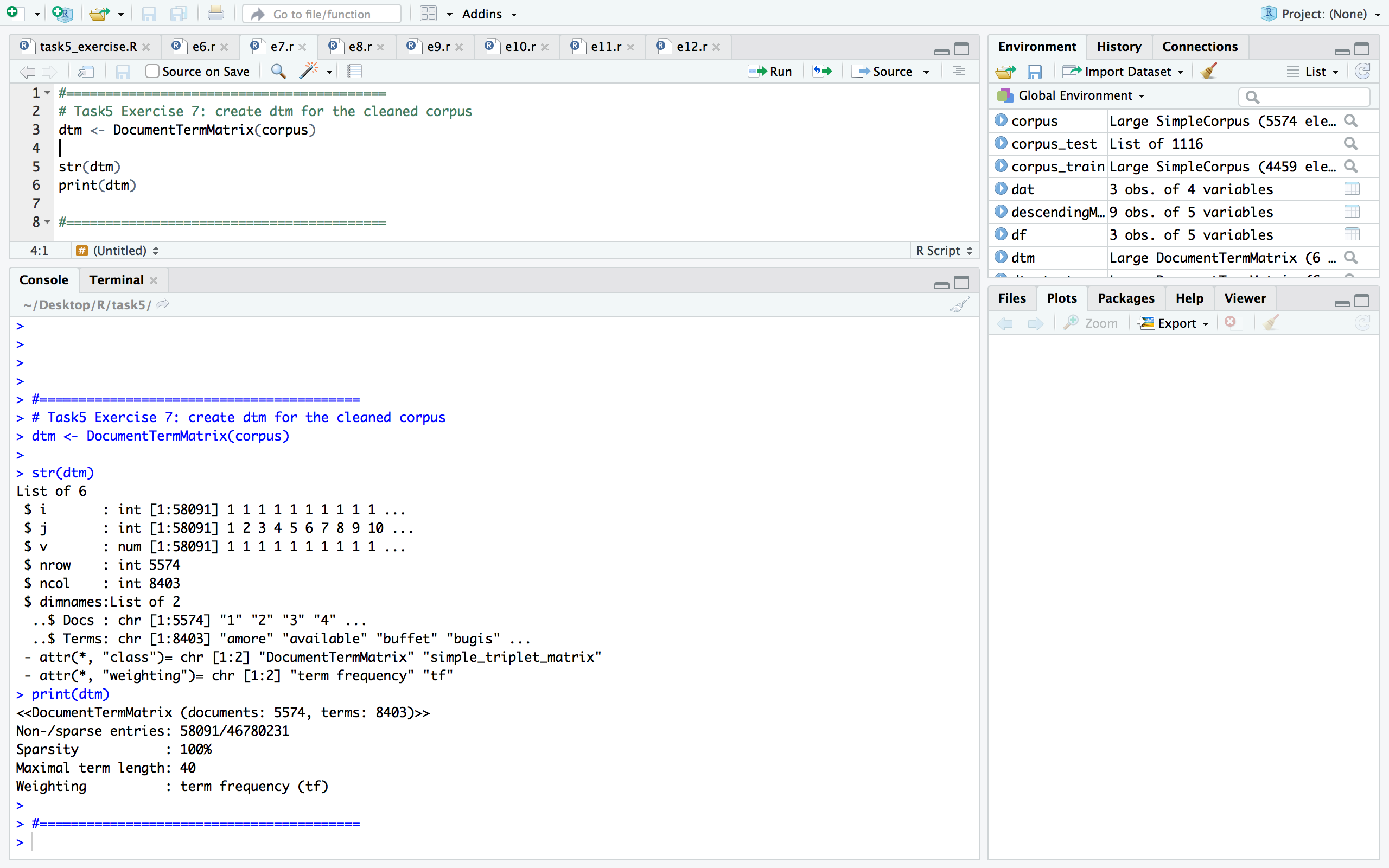
**Exercise.5: Inspect the first 3 messages at the corpus:**

****

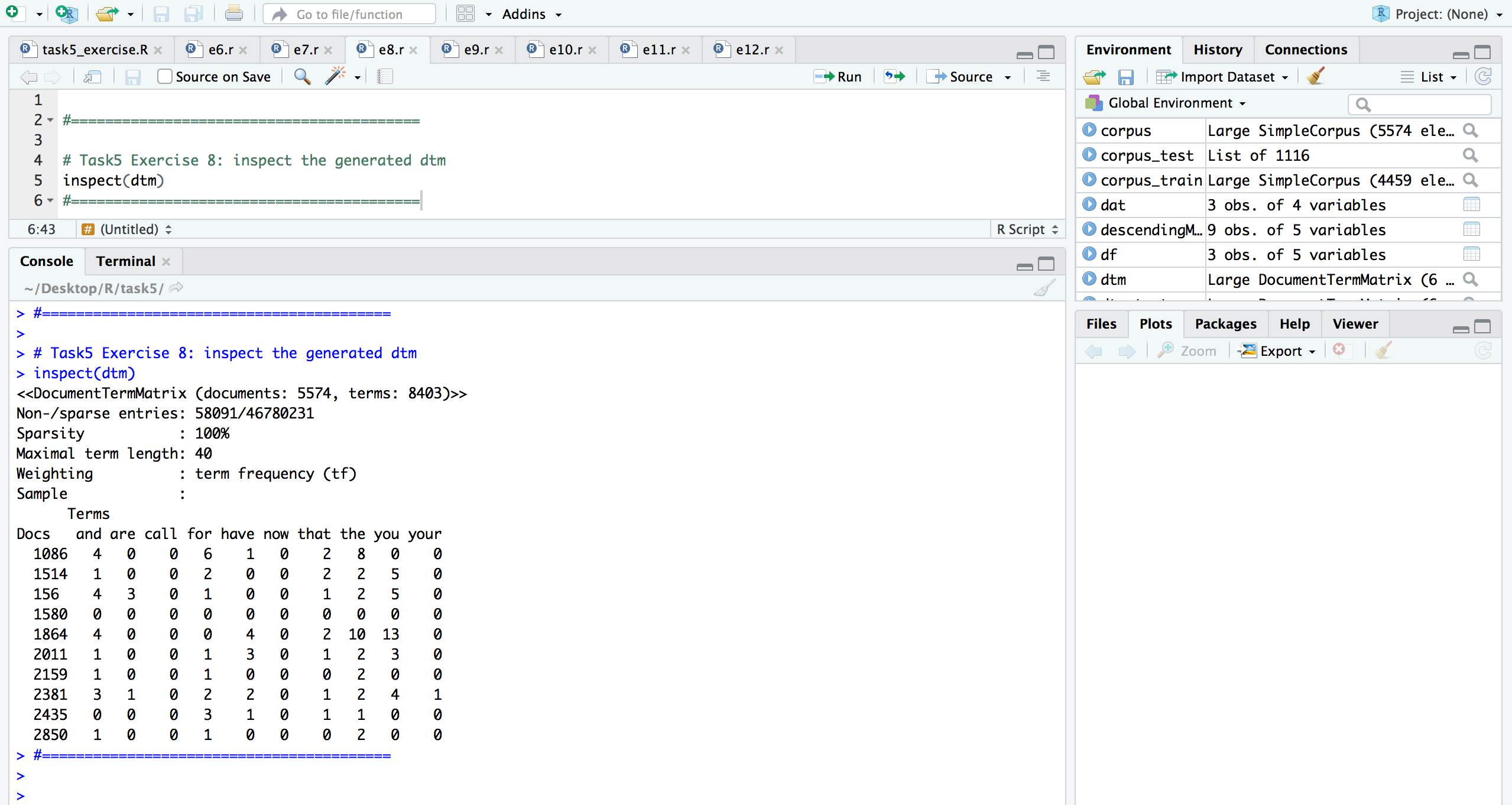
**Exercise.6: cleaning the text for corpus:**



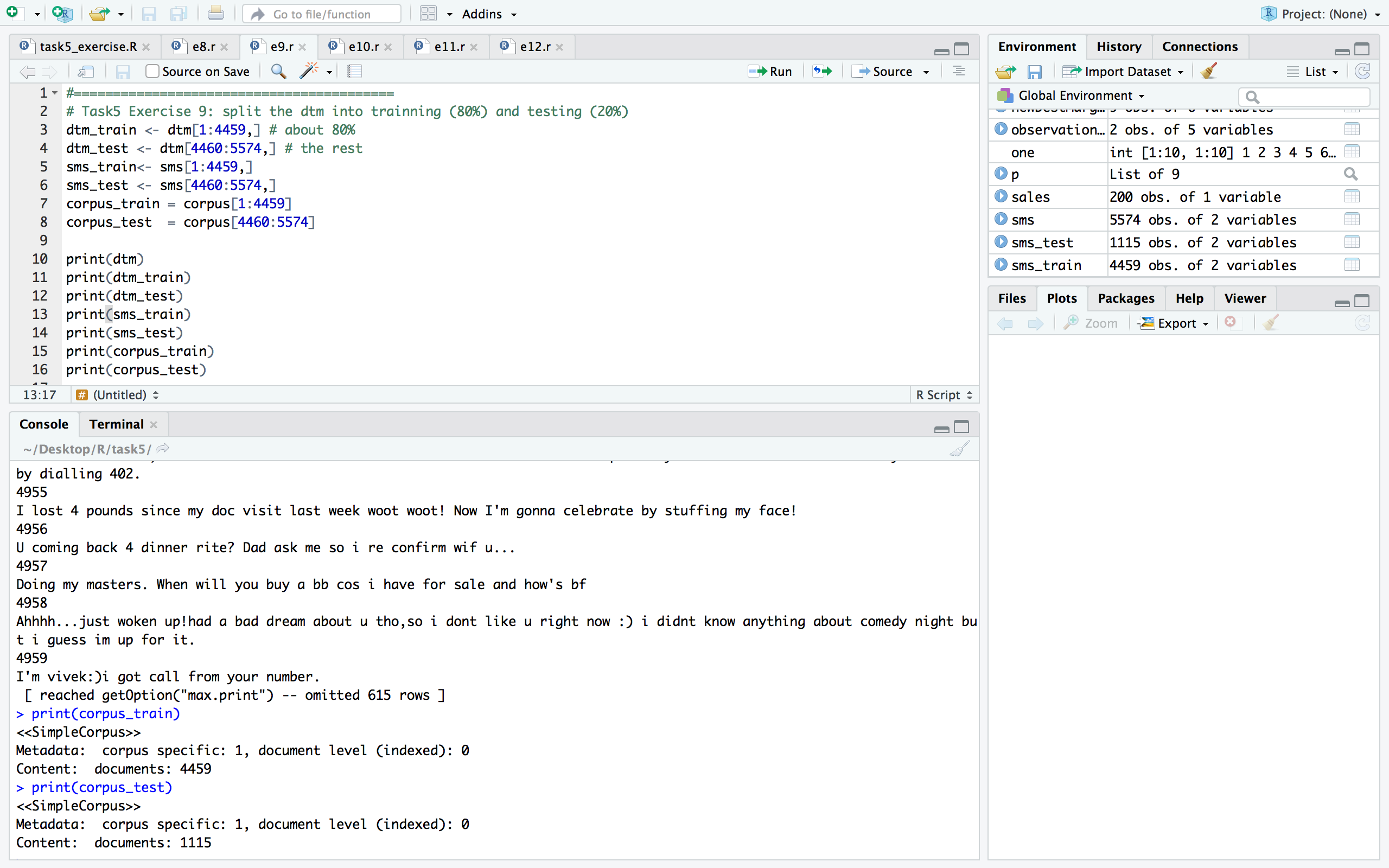
**Exercise.7: Create dtm for the cleaned corpus:**

****

**Exercise.8: Inspect the generated dtm**

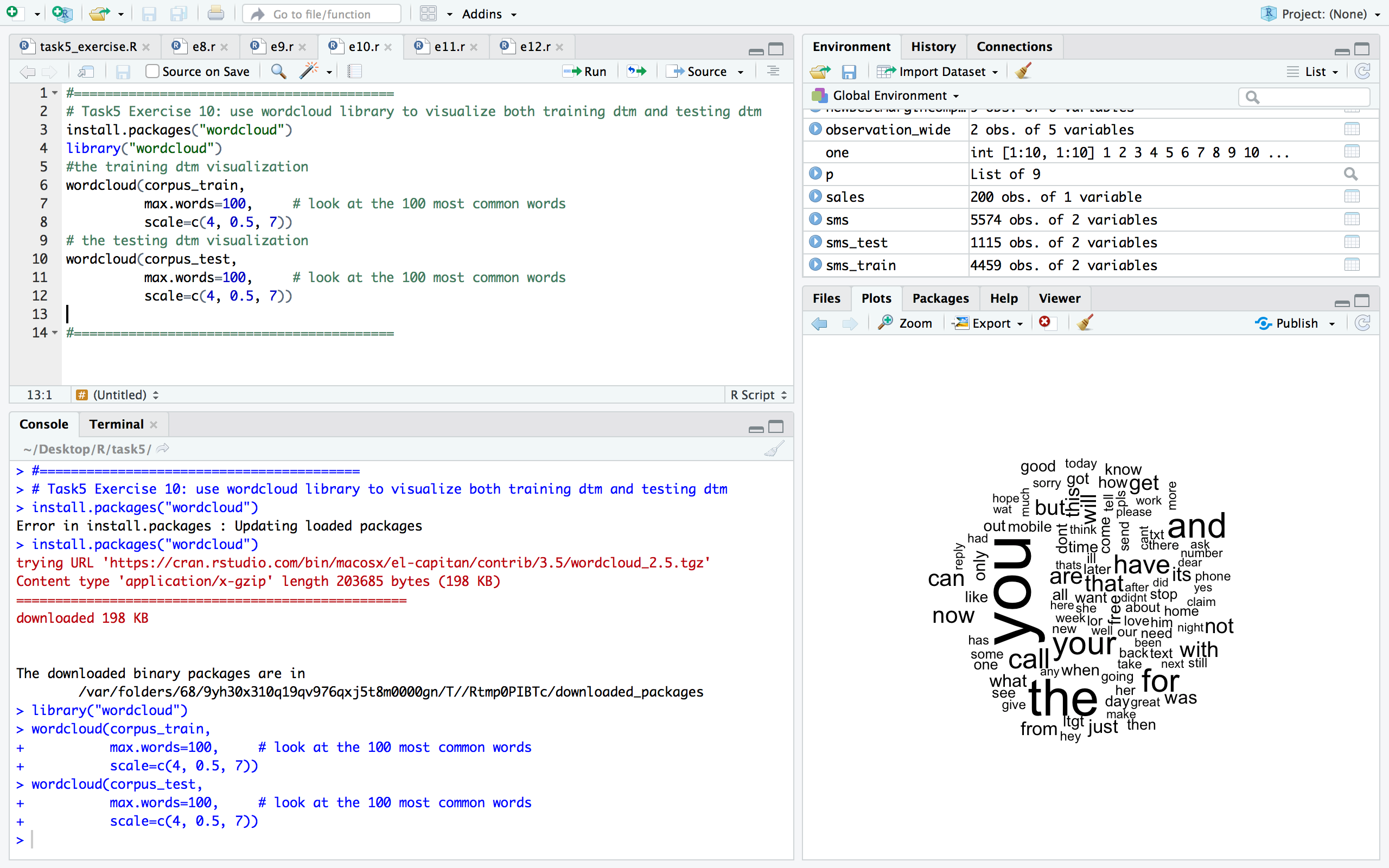


**Exercise.9: Split the dtm into trainning (80%) and testing (20%)**

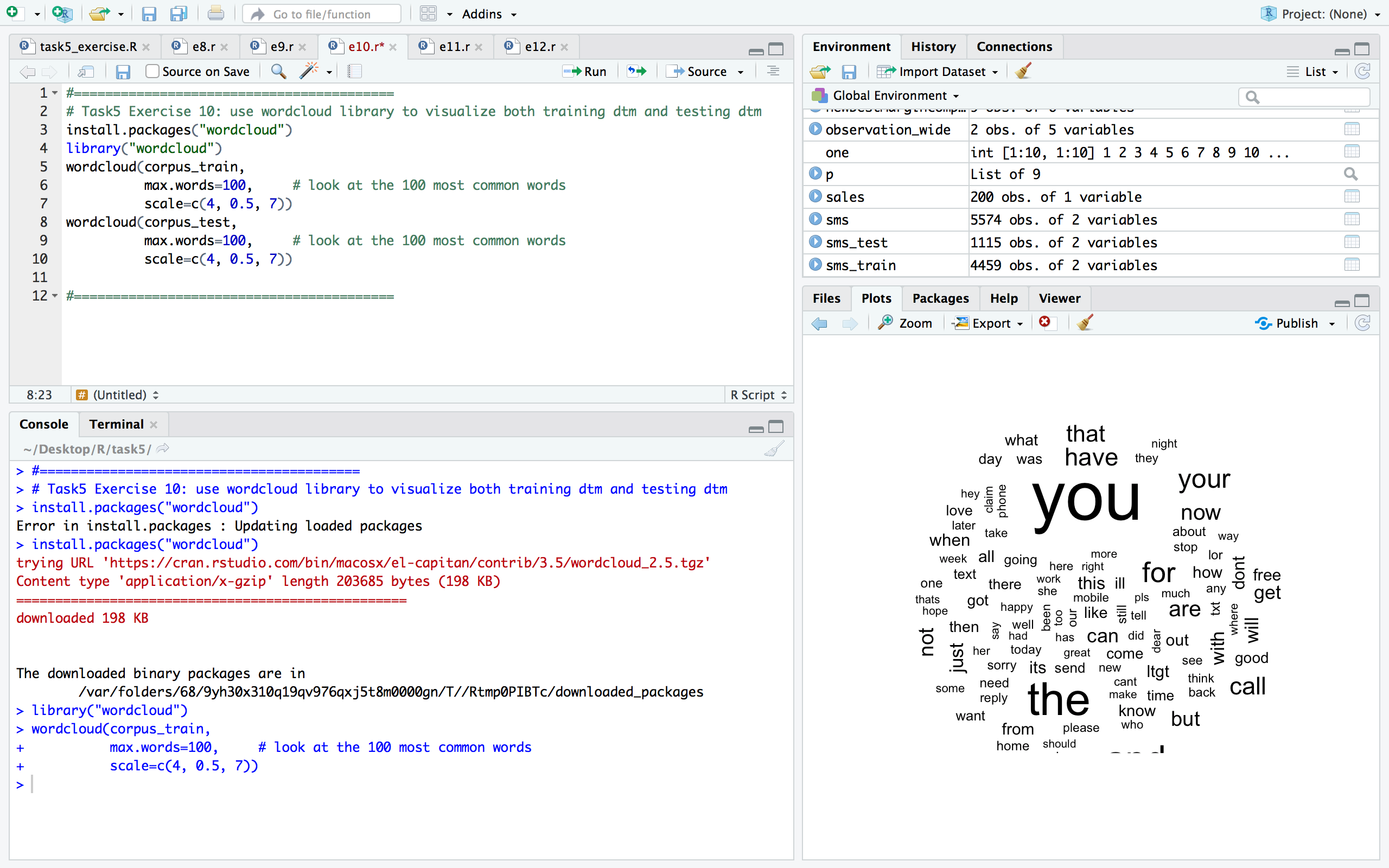
****

**Exercise.10: Use wordcloud library to visualize both training dtm and testing dtm**

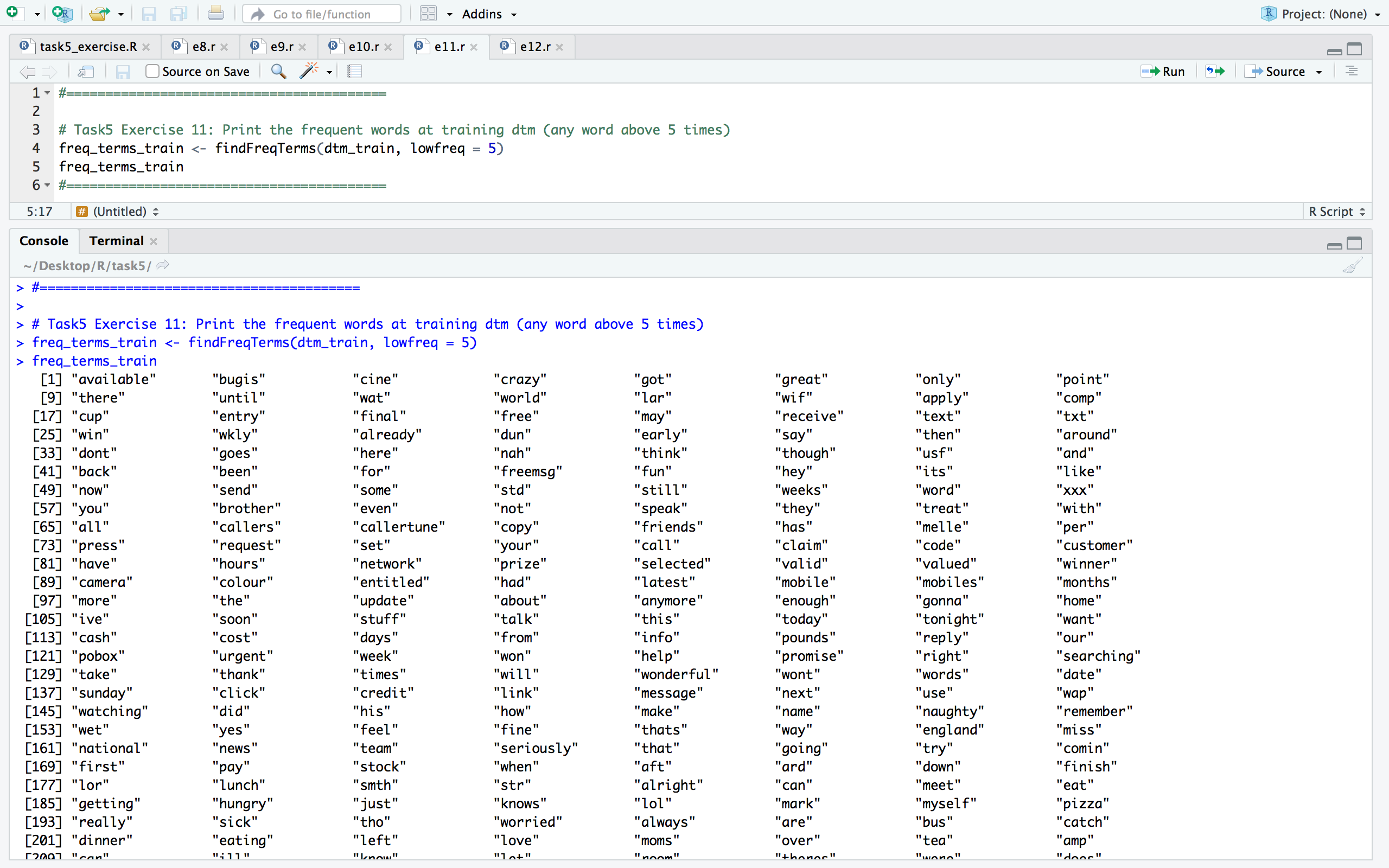
The Testing dtm visualization:

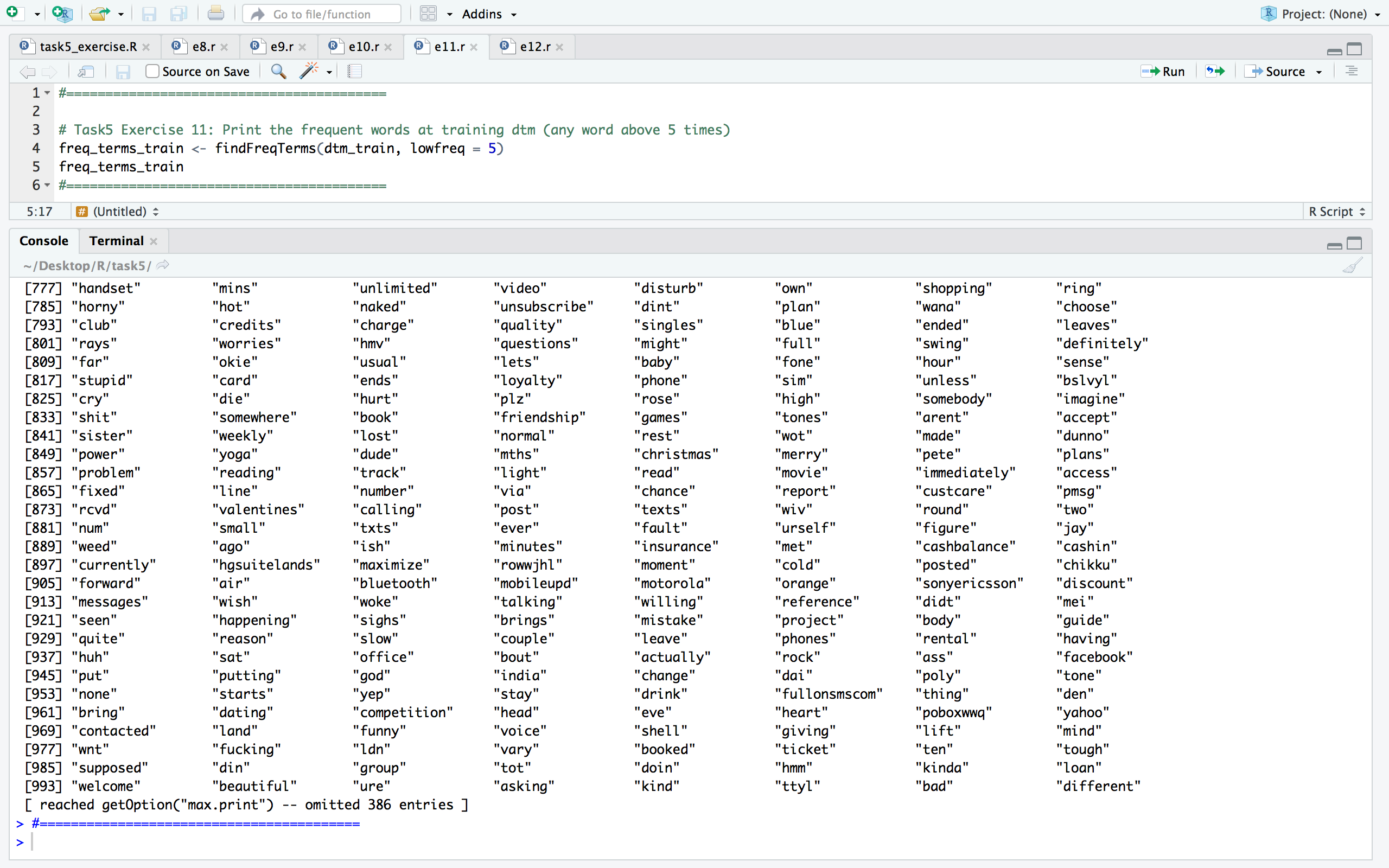
****

The Training dtm visualization:

****

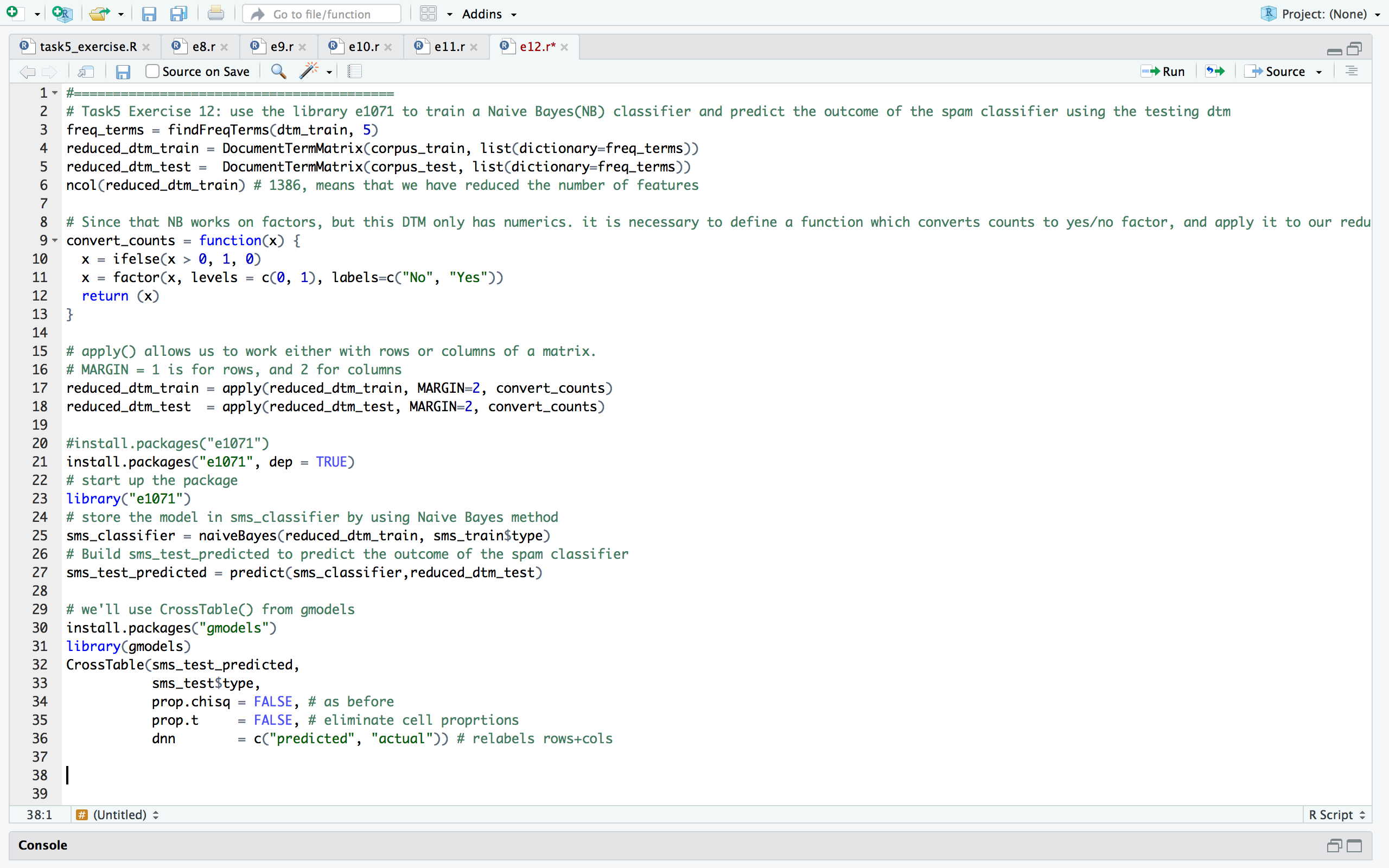
**Exercise.11:** **Print the frequent words at training dtm (any word above 5 times)**

****

****

**Exercise.12: Use the library e1071 to train a Naive Bayes(NB) classifier and predict the outcome of the spam classifier using the testing dtm.**

The code:



The predicted result:

