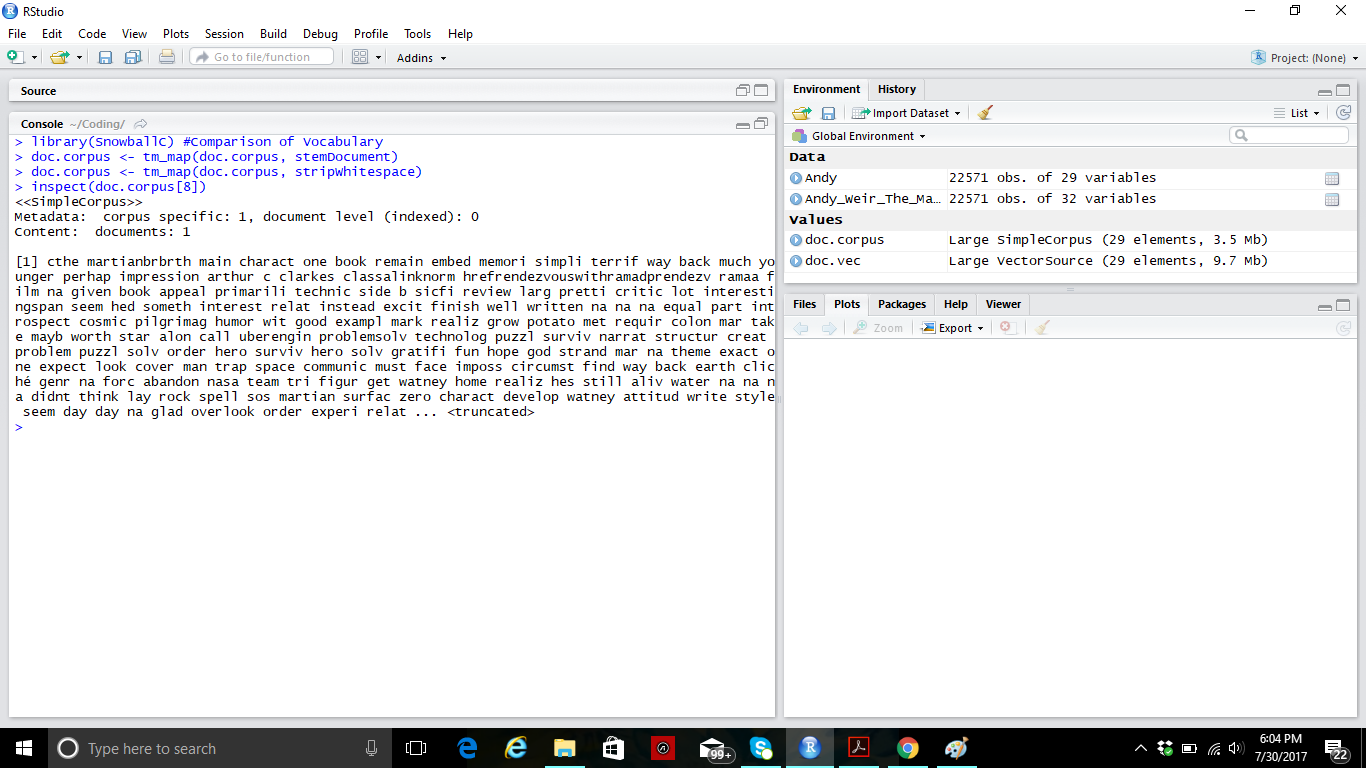
Analysis of Text Mining:

For analyzing the text of the given data sets in R, we use “tm” library. First, we consider the length of the data sets and then perform Natural Language Processing using “tm\_map” function and “Snowball C” library. The TermDocumentMatrix() gives the number of term used per document, further we find the words which occurred frequently and remove the commonly occurring terms and later plot those words using “wordcloud()”.

Here, the dark colors represent the words which occurred more frequently while the light colors stand for less frequently occurring words.

For Amazon Data set1:

Result for the Natural Language Processing using tm\_map():



The result for Text Document Matrix, which gives the sparsity of the document.

*<<TermDocumentMatrix (terms: 17885, documents: 29)>>*

*Non-/sparse entries: 59892/458773*

*Sparsity: 88%*

*Maximal term length: 83*

*Weighting: term frequency (tf)*

*> inspect(TDM[1:10,1:10])*

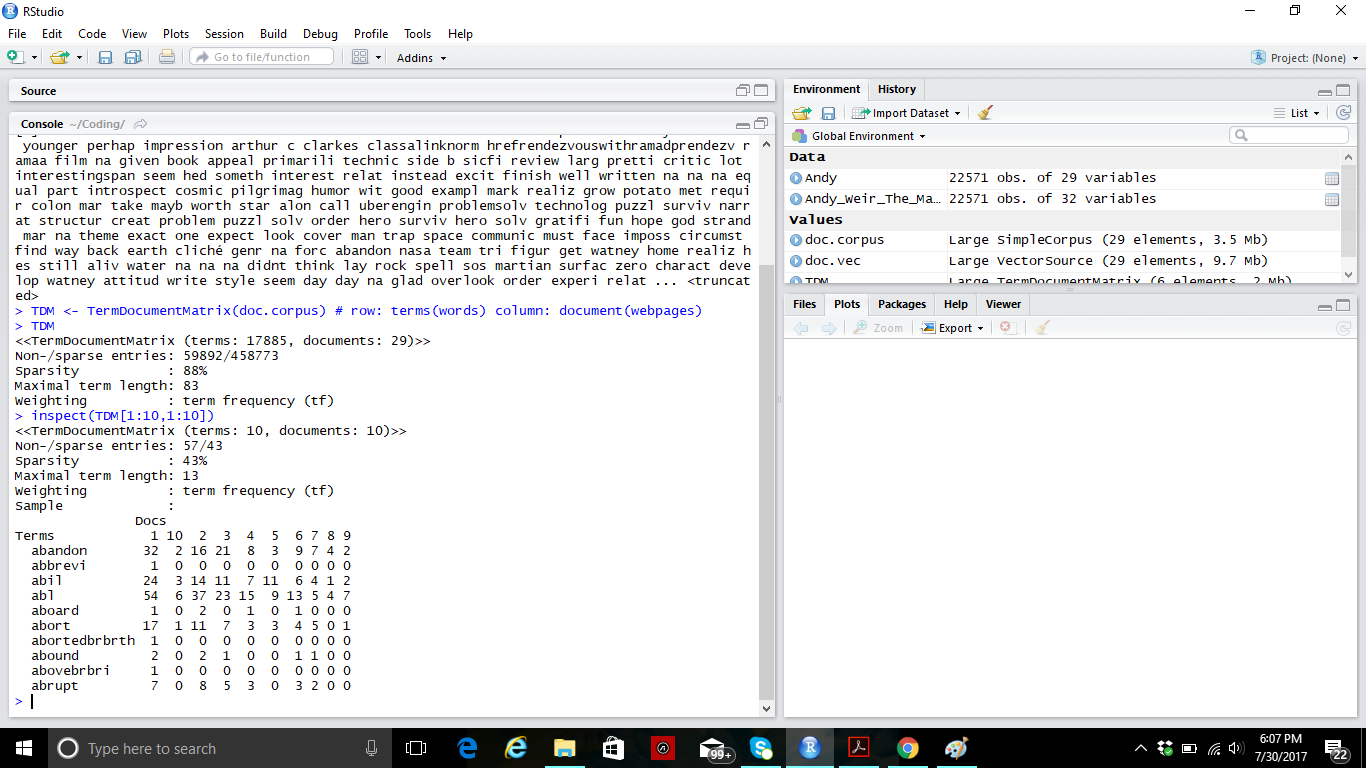
*<<TermDocumentMatrix (terms: 10, documents: 10)>>*

*Non-/sparse entries: 57/43*

*Sparsity: 43%*

*Maximal term length: 13*

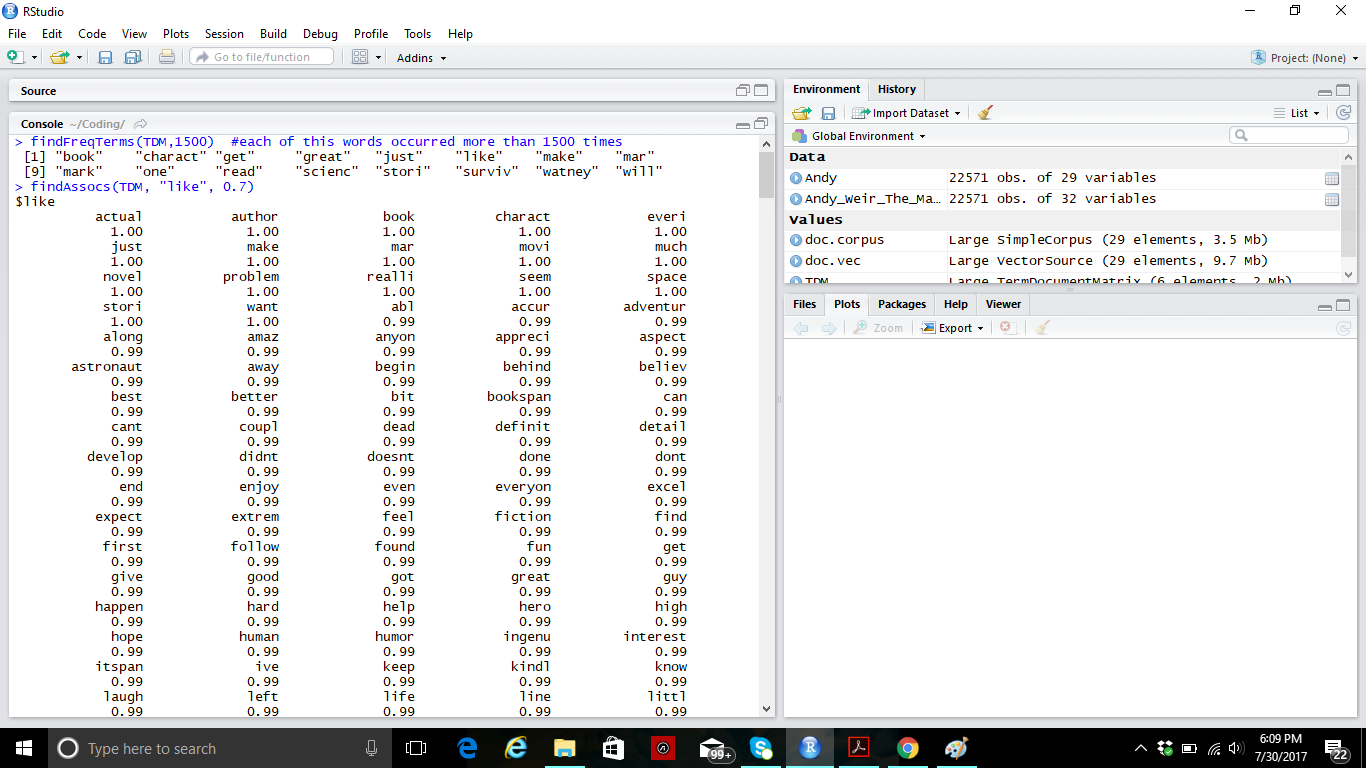
*Weighting: term frequency (tf)*



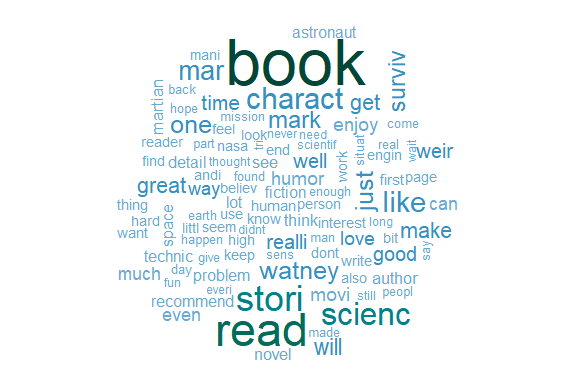
The result of words occurring more than 1500 times is:

*[1] "book" "charact" "get" "great" "just" "like" "make" "mar"*

*[9] "mark" "one" "read" "scienc" "stori" "surviv" "watney" "will"*



The visualization of the 100 frequently used words.



Here, “read” and “book” are the frequently used words. Whereas, “scienc”, “astronaut”, “fiction” depicts that the book is a science fiction novel. The words “like”, “love”, “great” express the sentiments of the readers, which means readers loved the novel.