**Project 3 – Text Analysis**

By

Abhinav Verma

Aditya Dharne

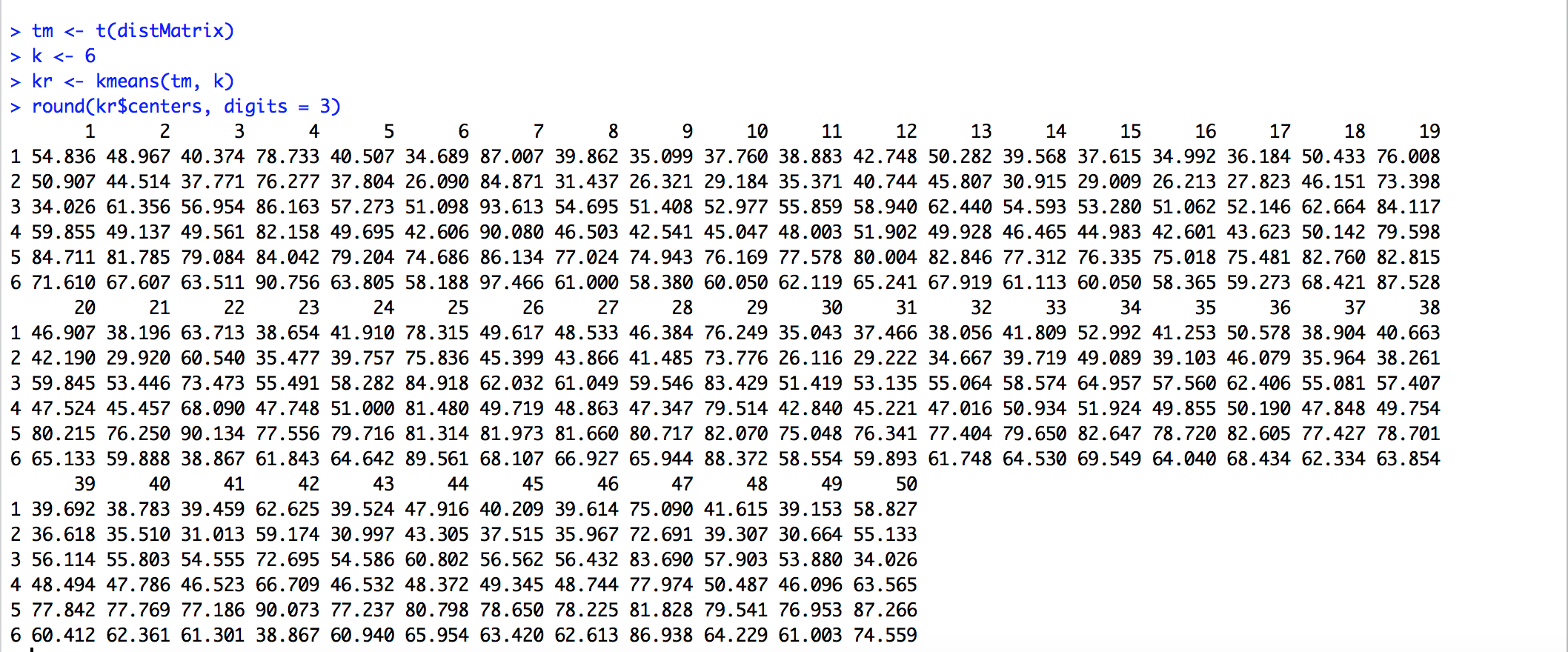
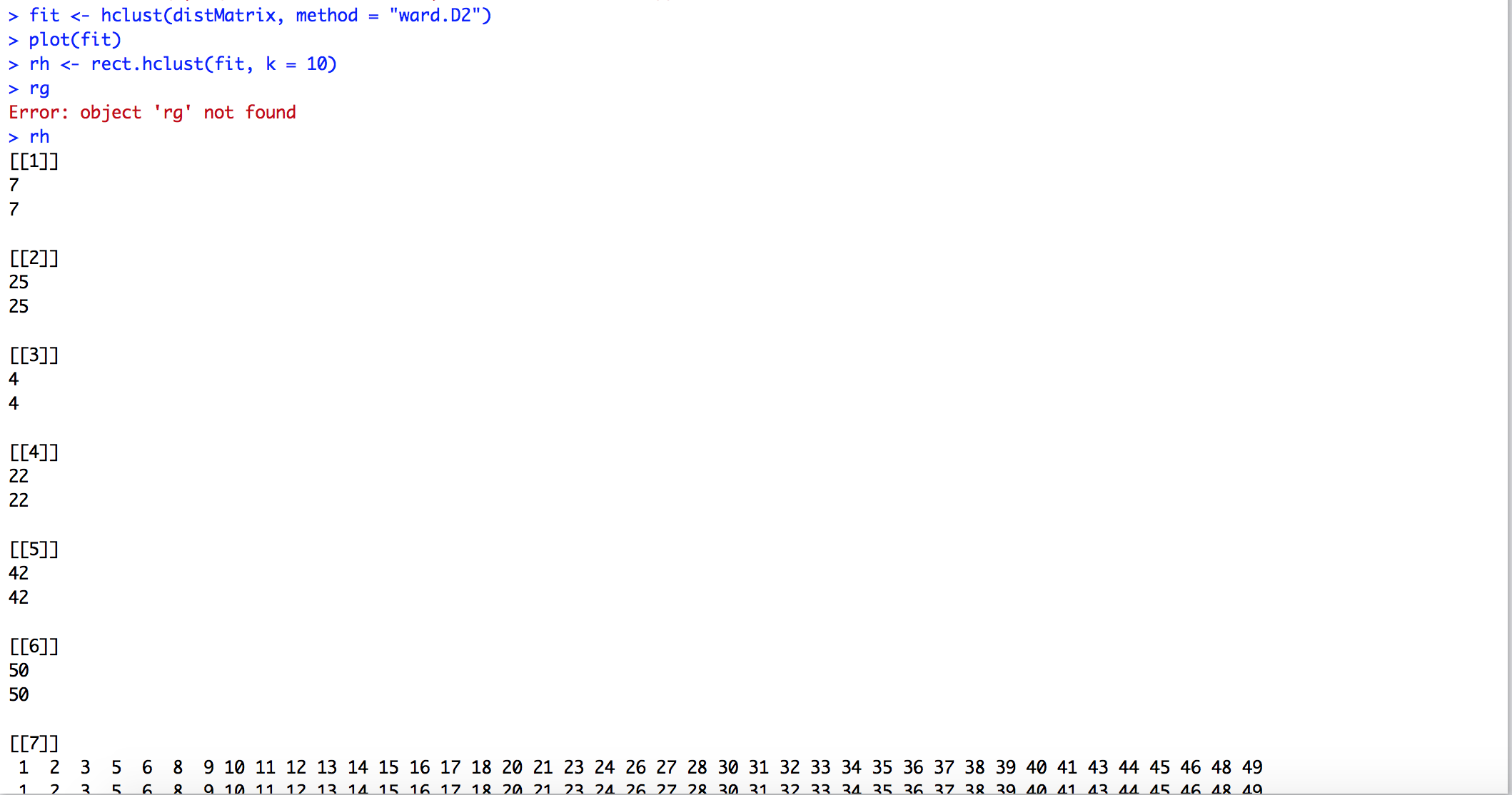
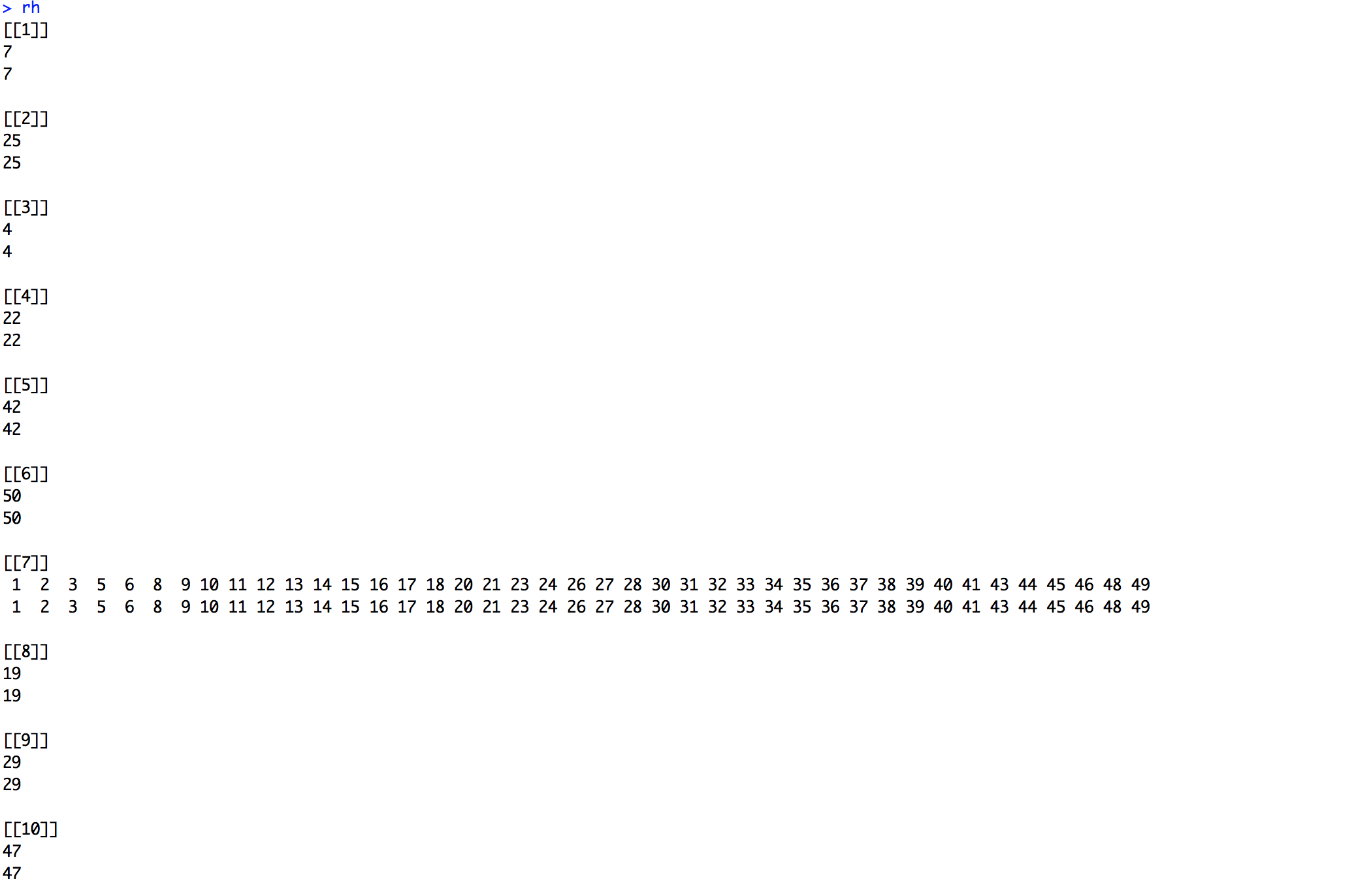
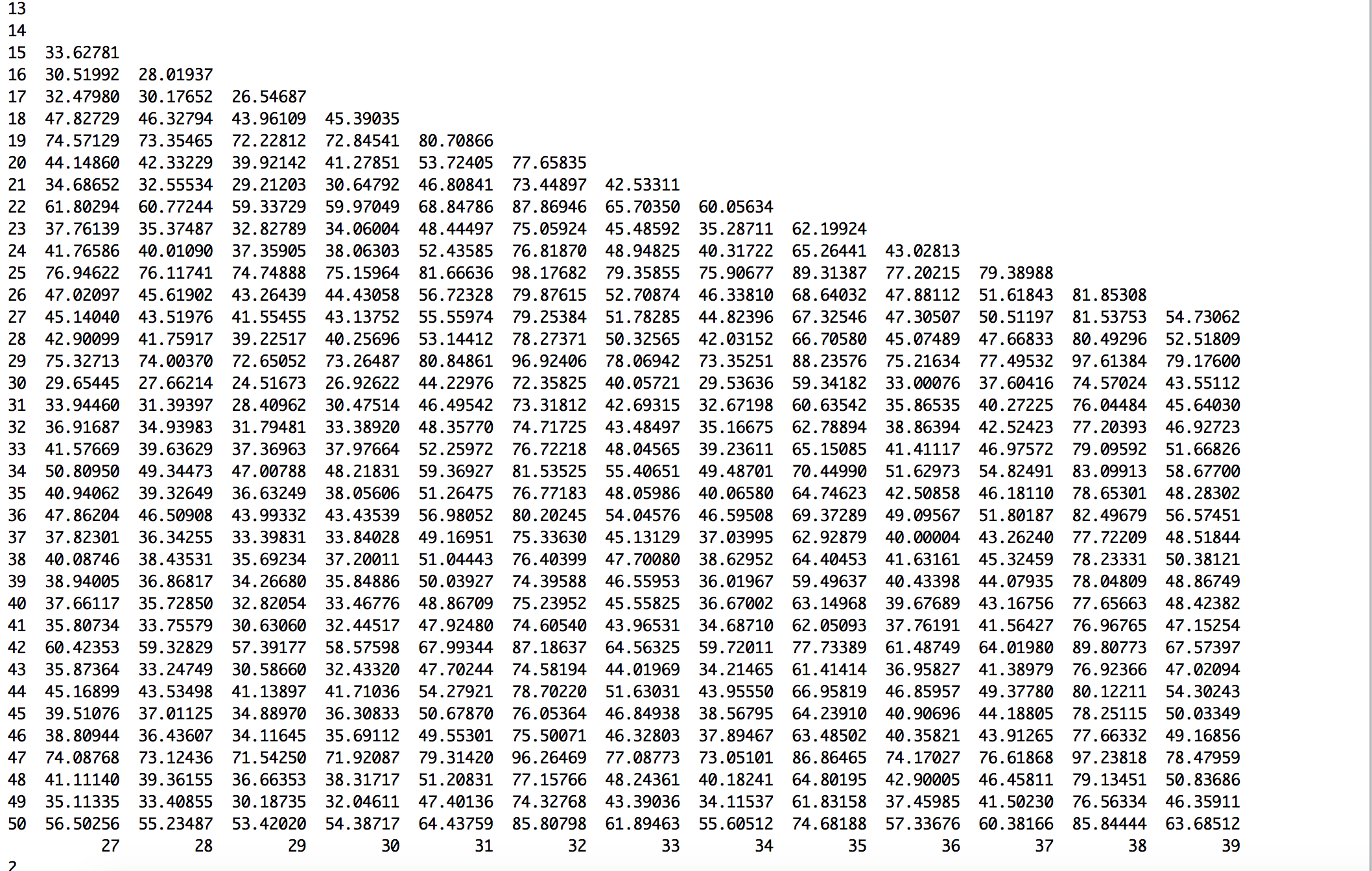
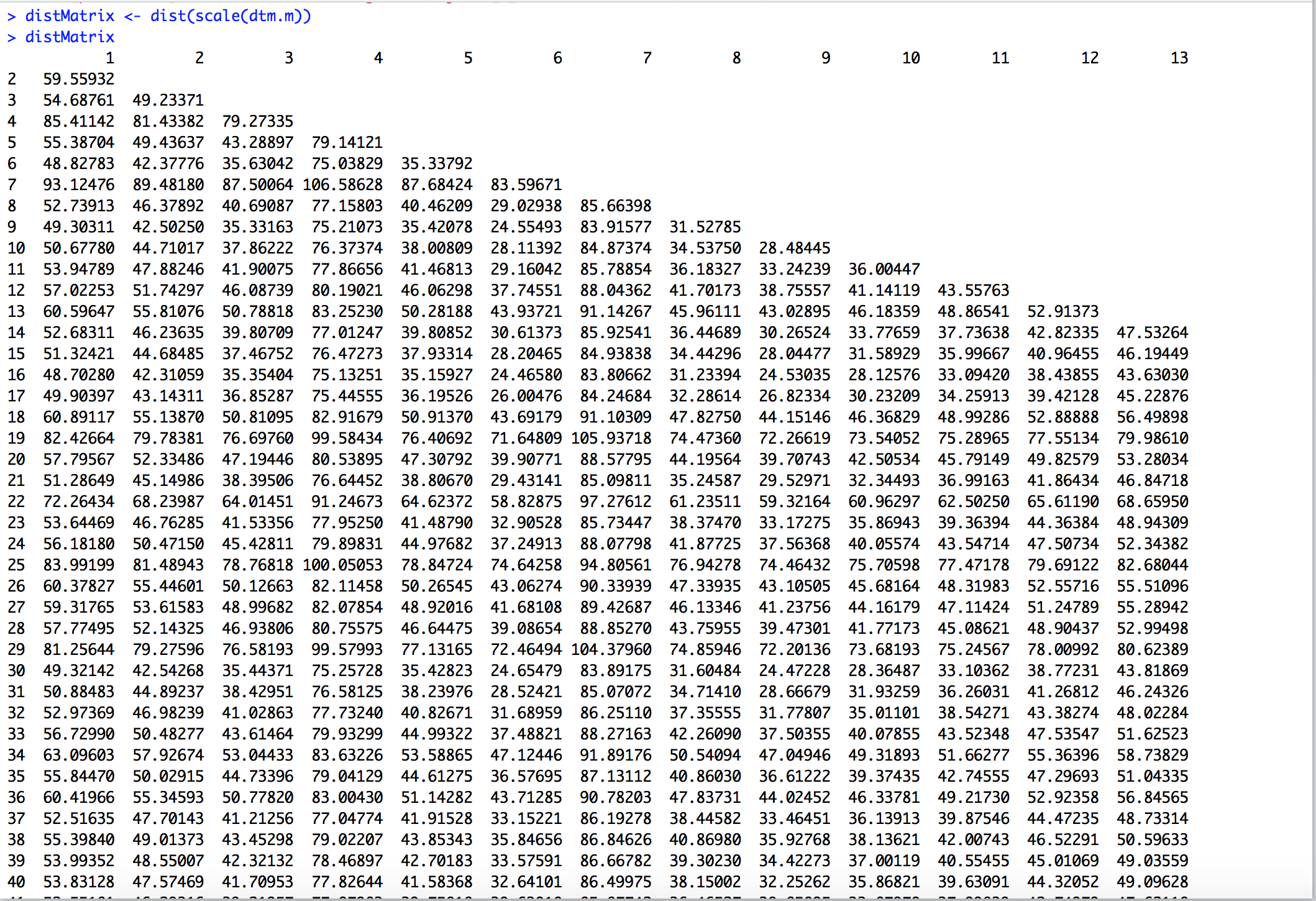
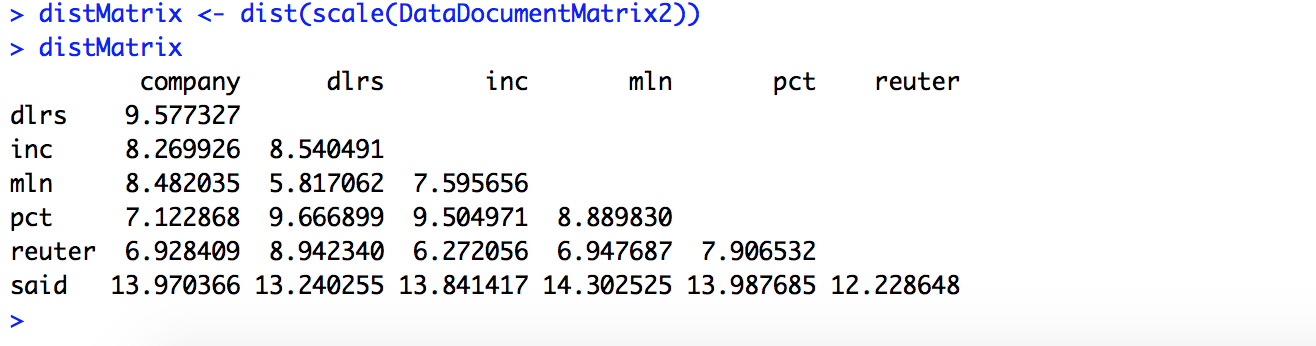
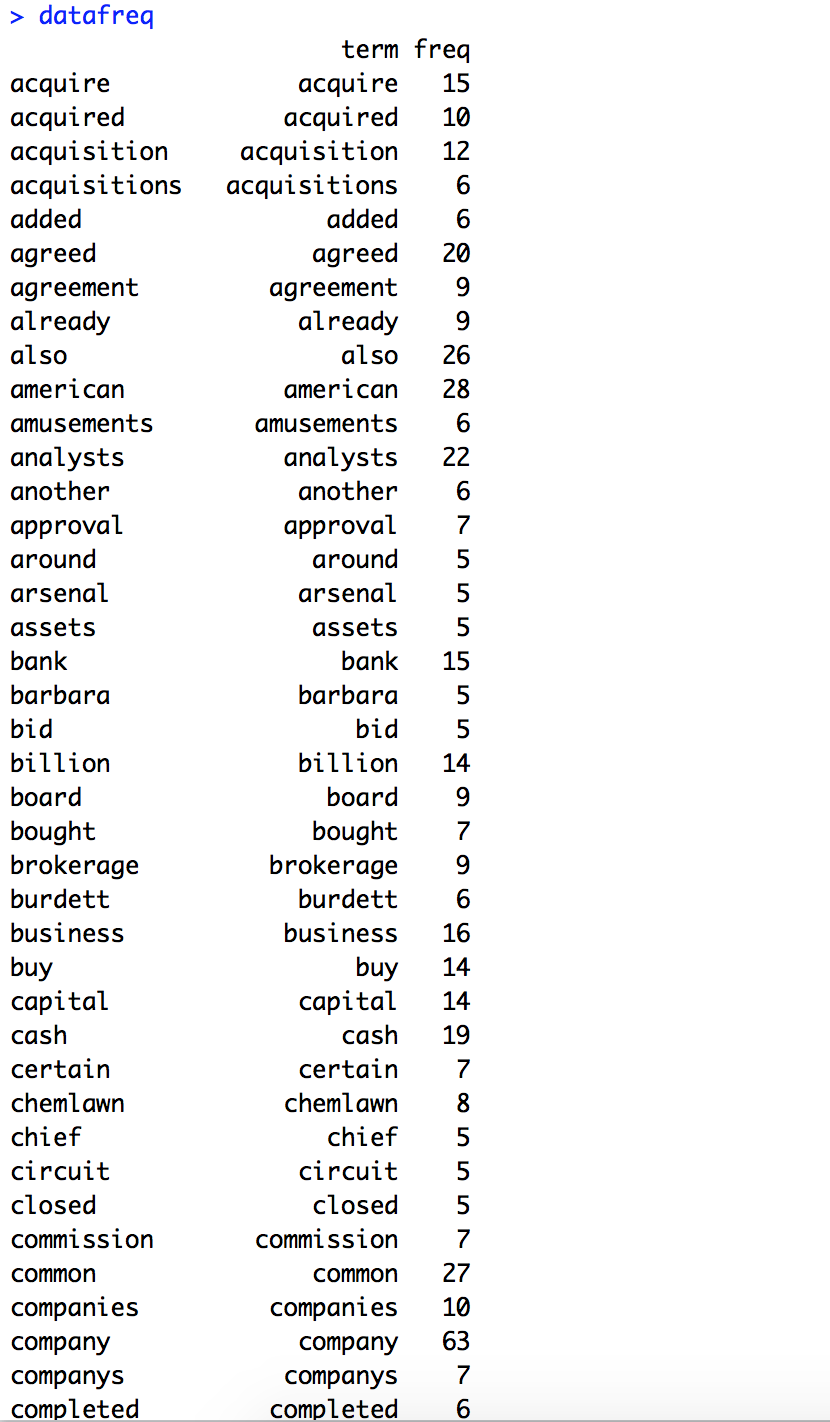
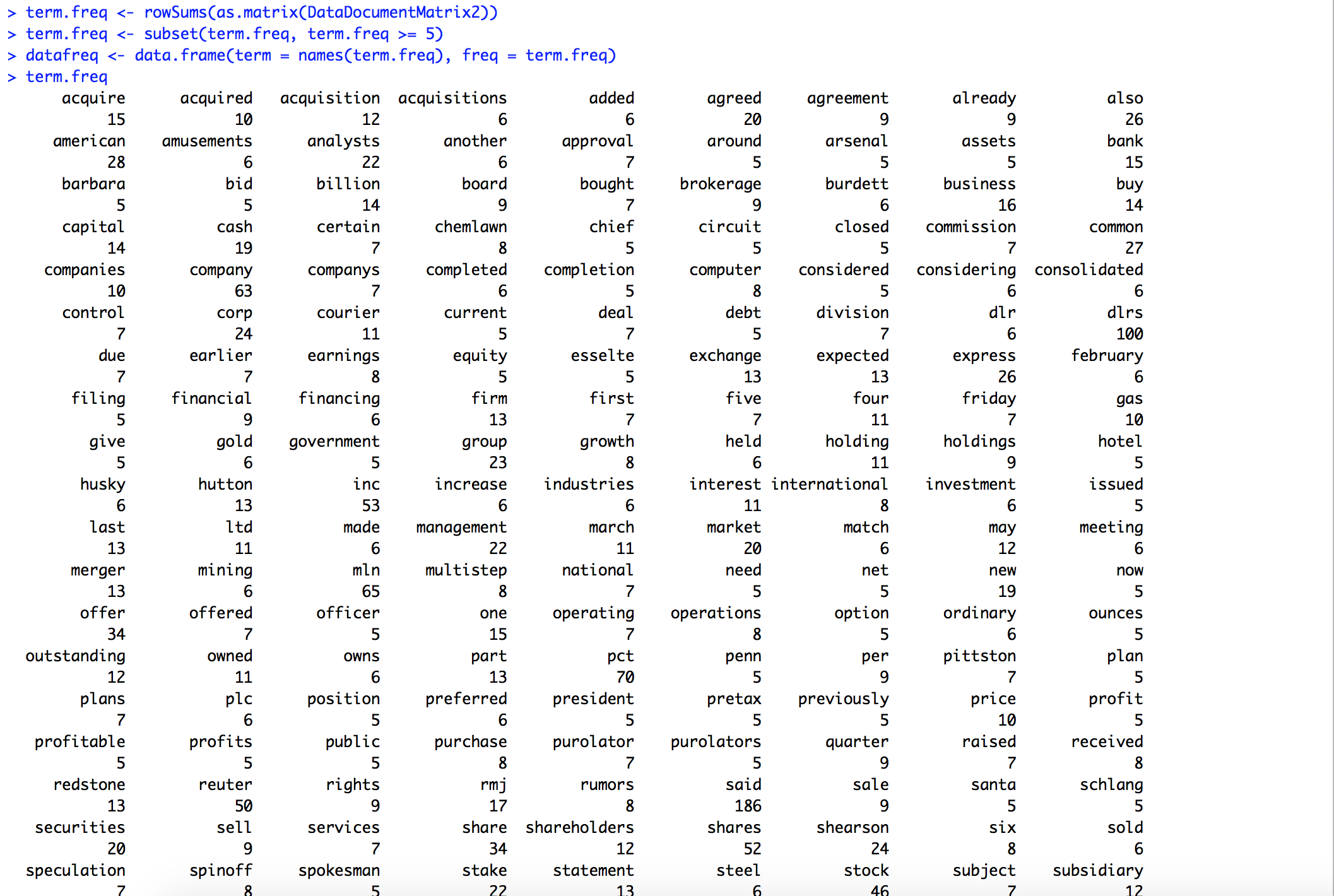
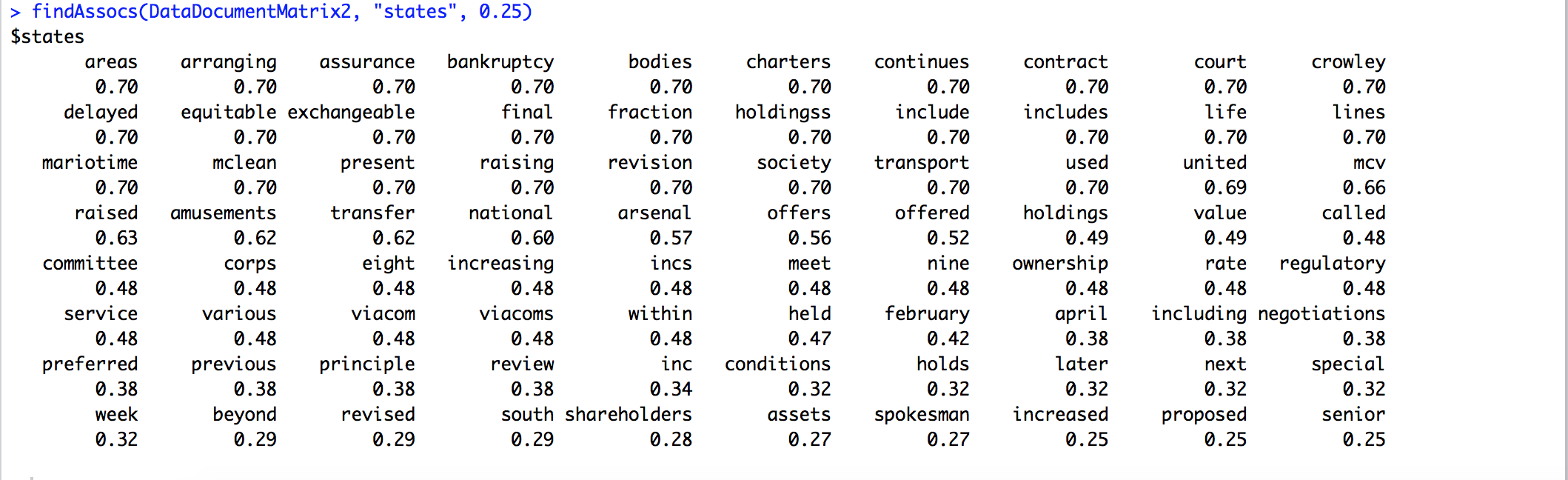
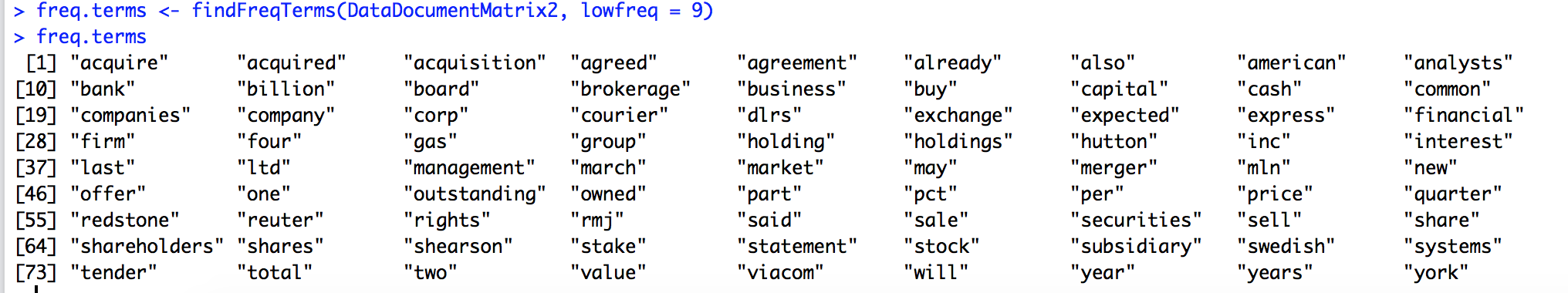
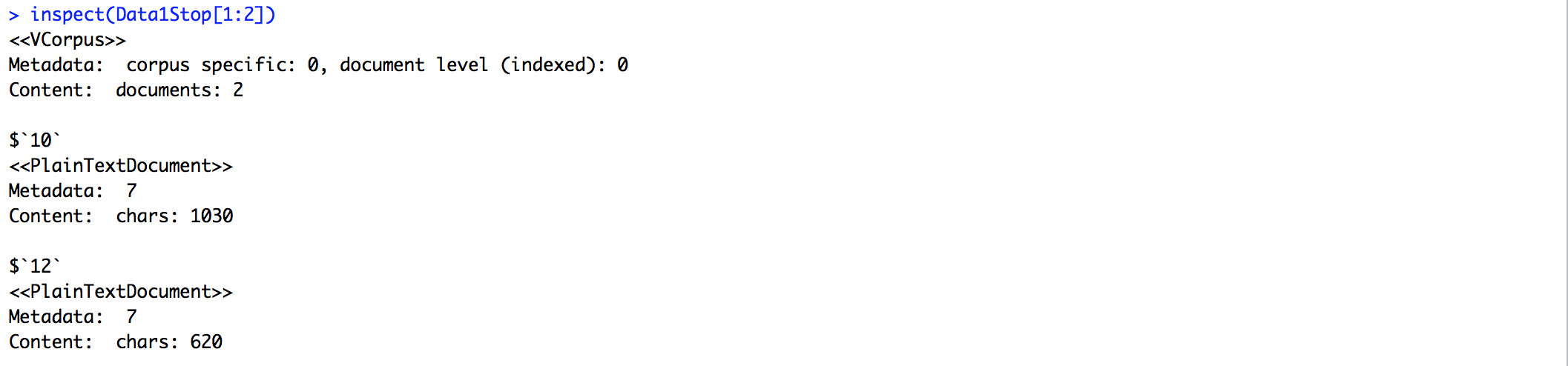
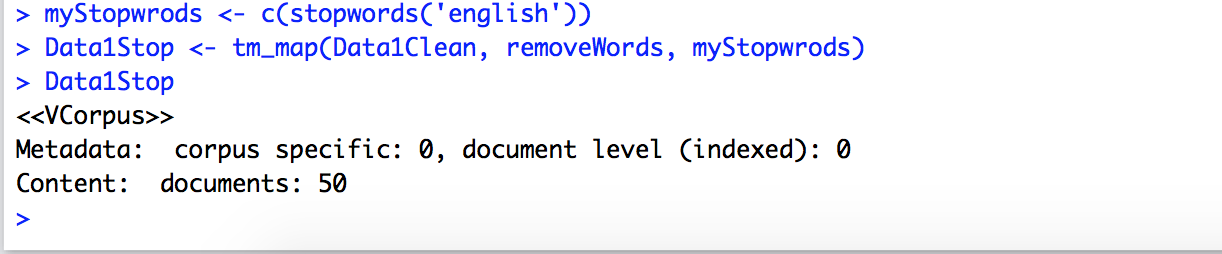
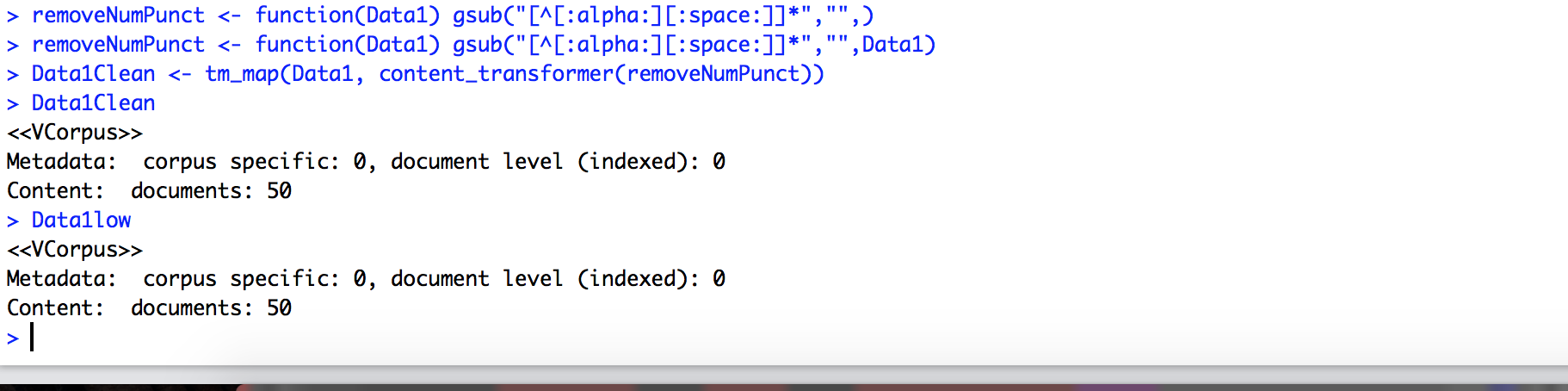
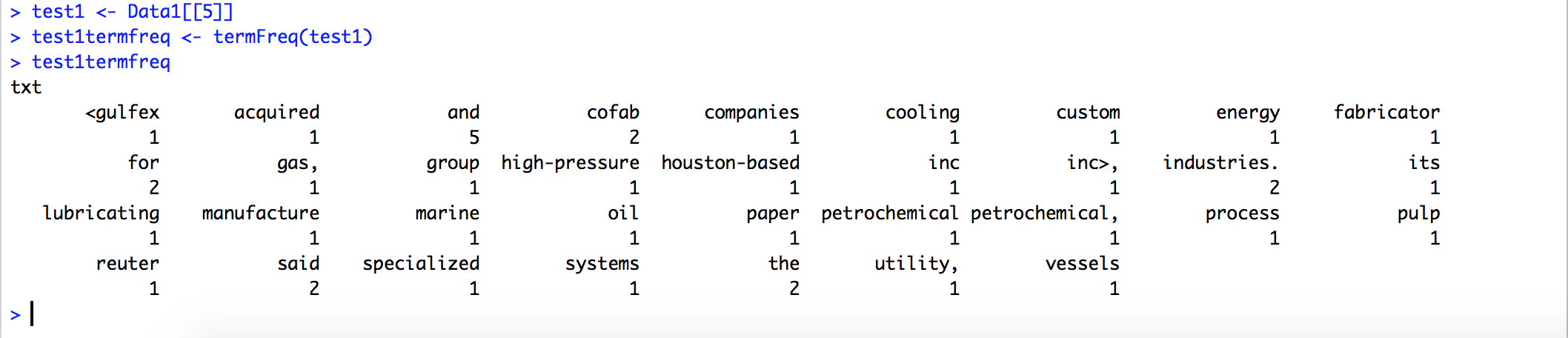
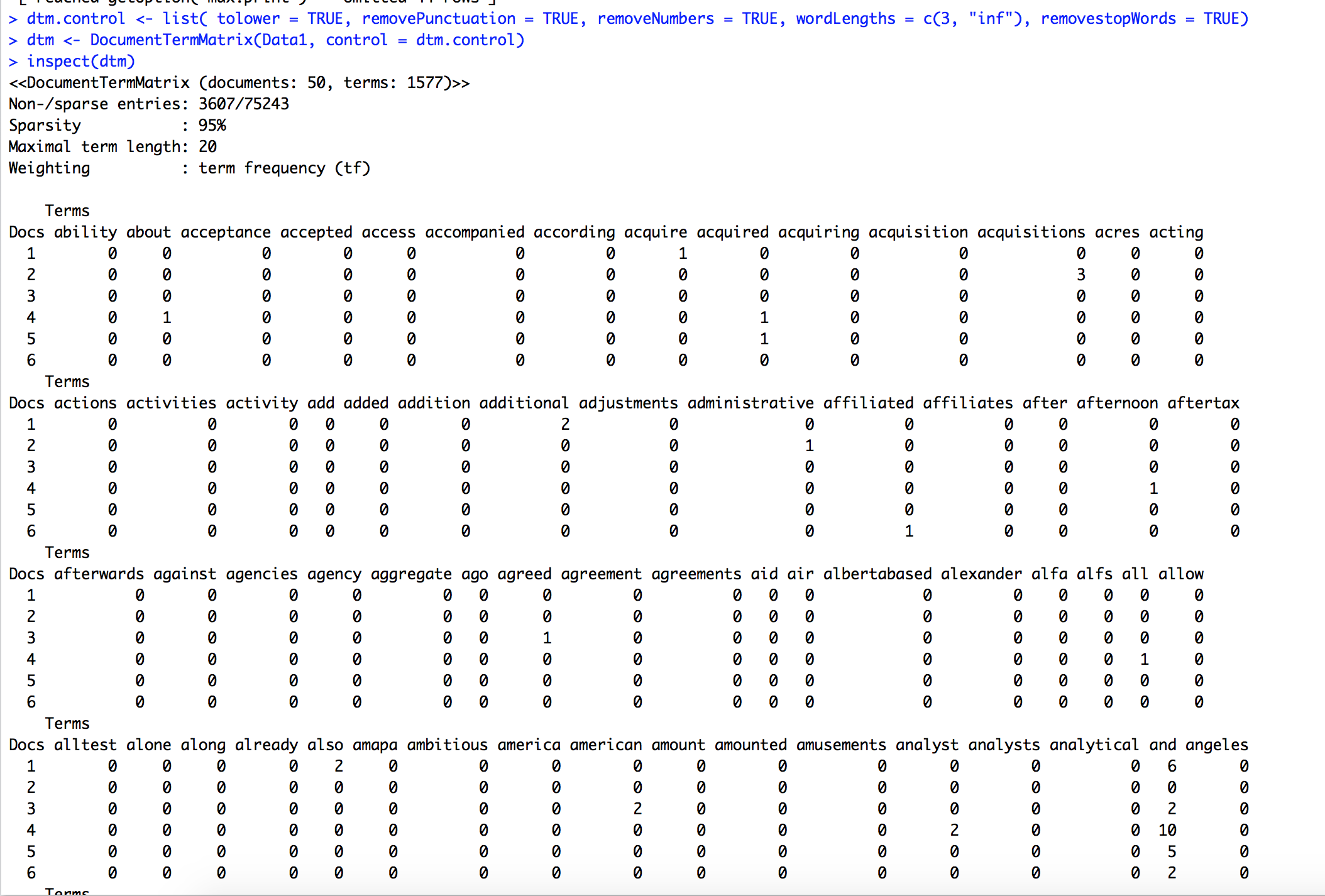
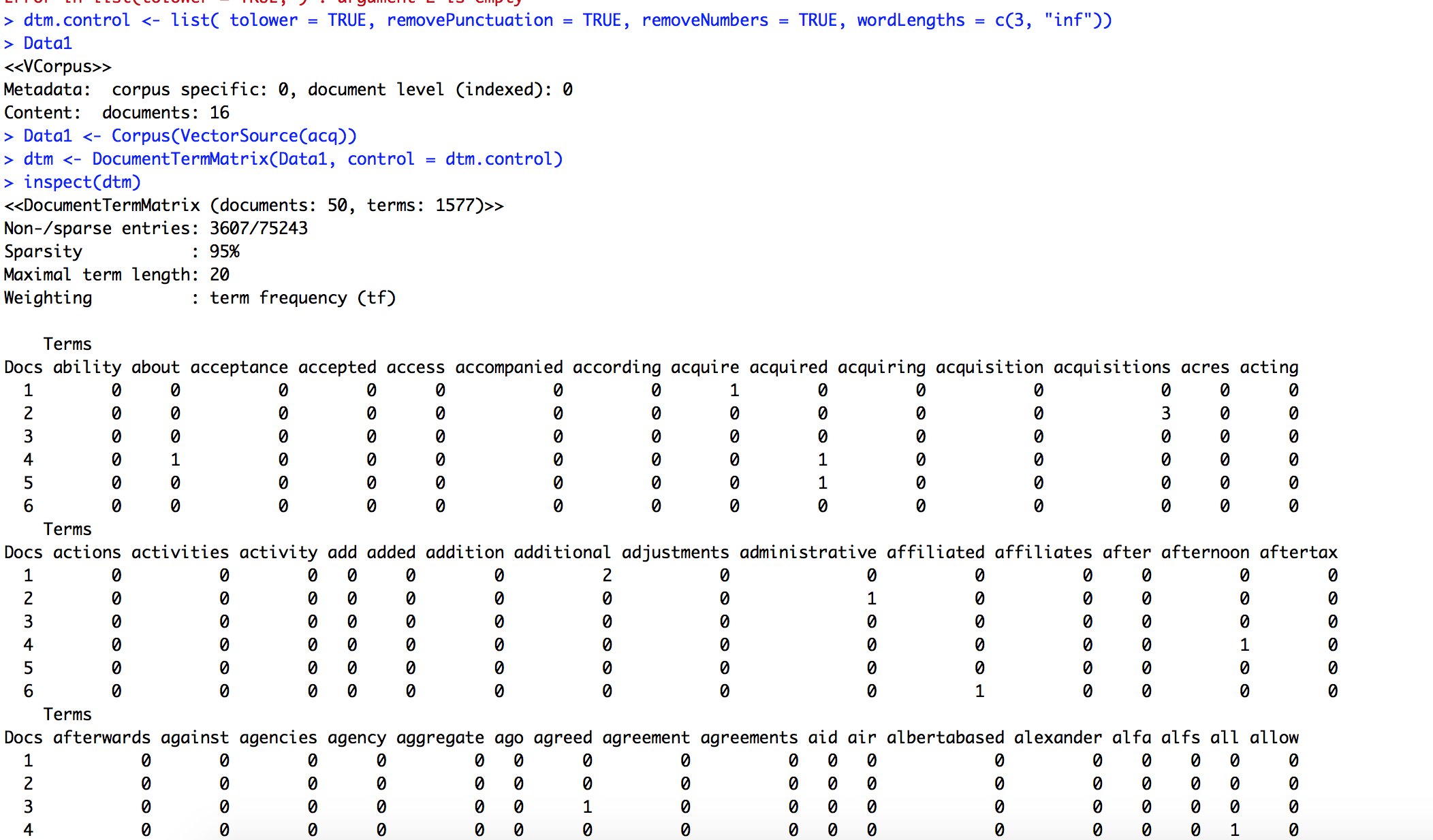
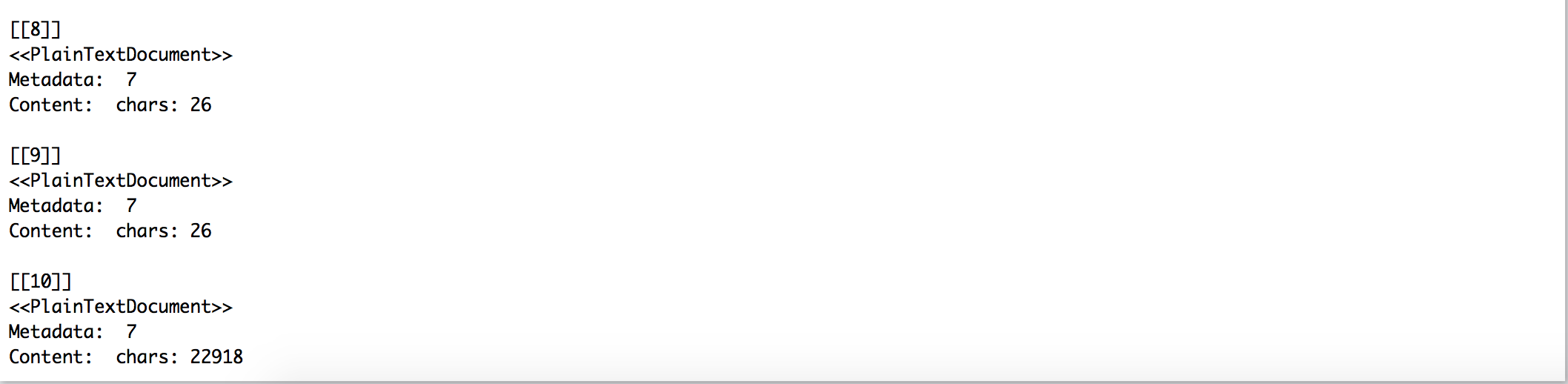
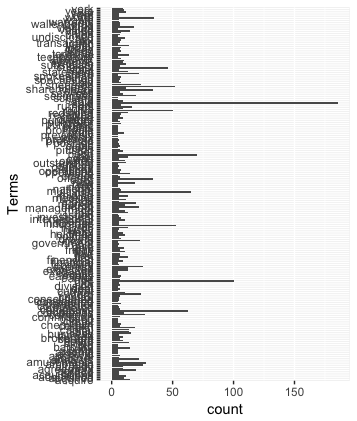
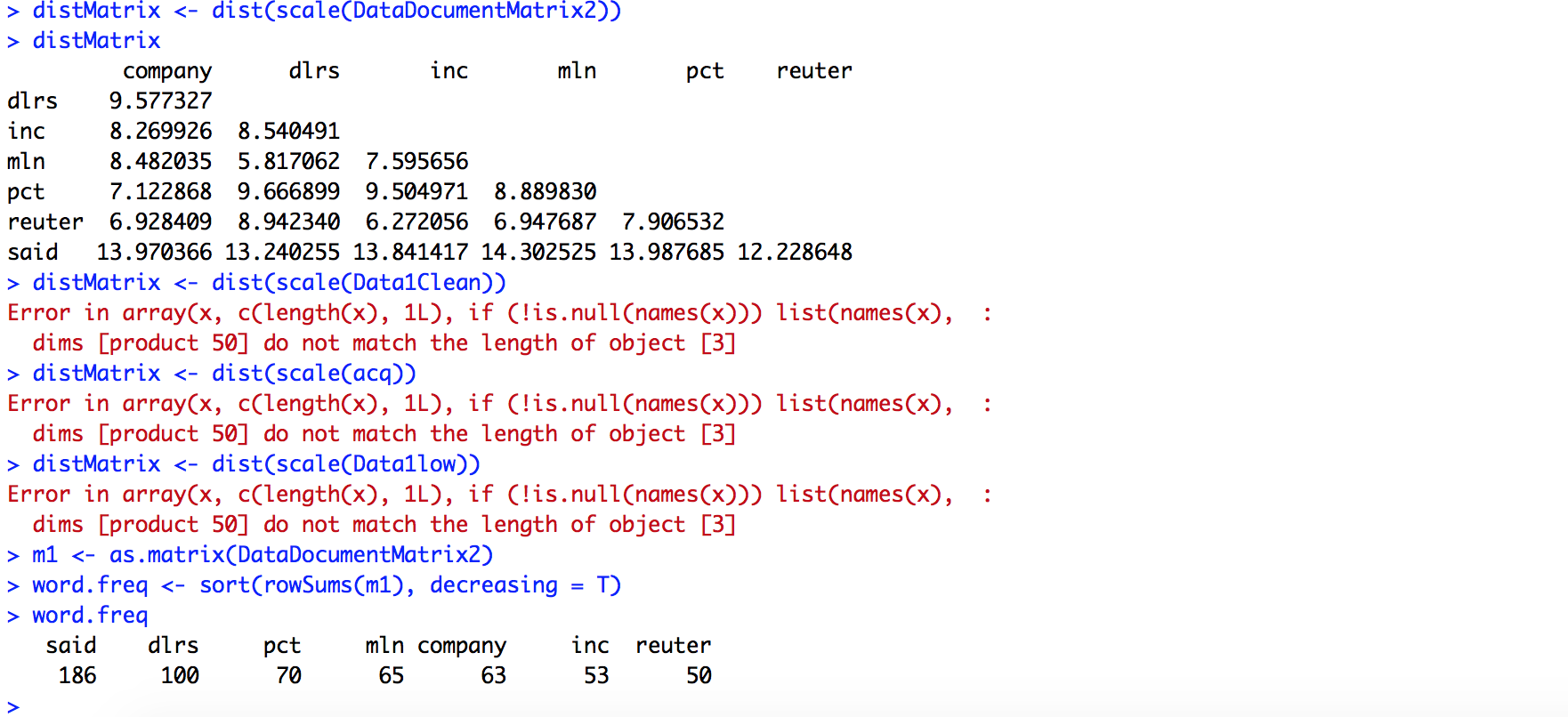
Shrey Arora

Vishnu Sarvagna Gollamudi

**Data Set**: acq in package tm

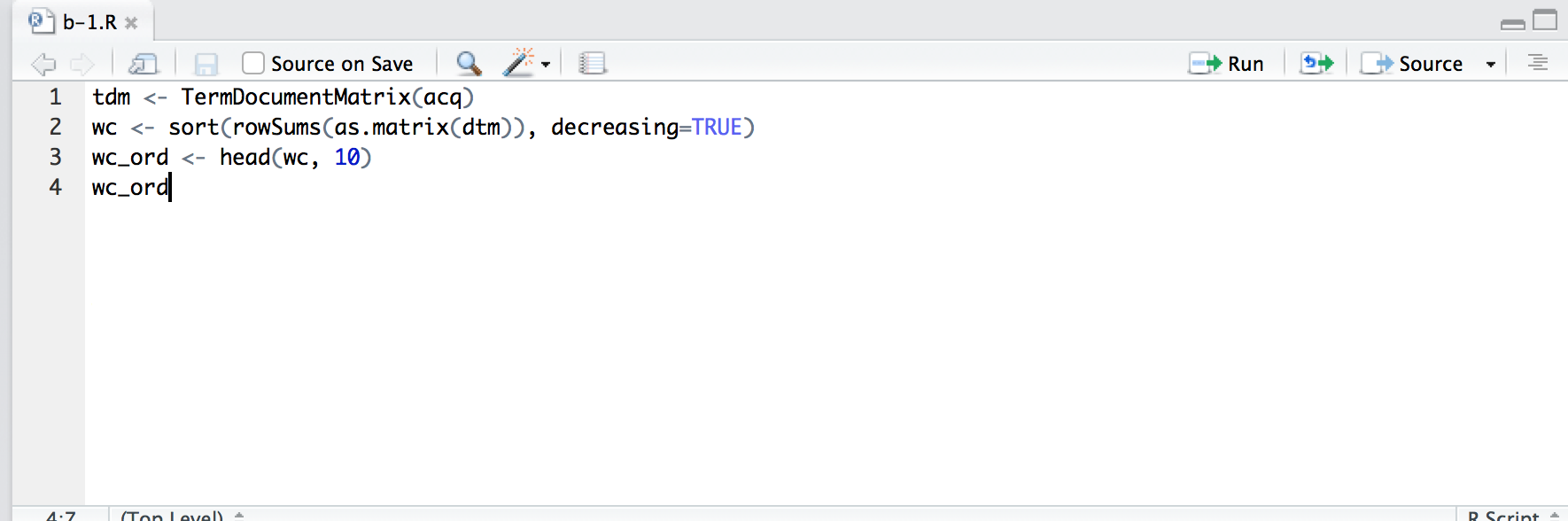
**Deliverables**

1. Functions from Lecture 7 slides:



1. Find the 10 longest documents (in number of words).

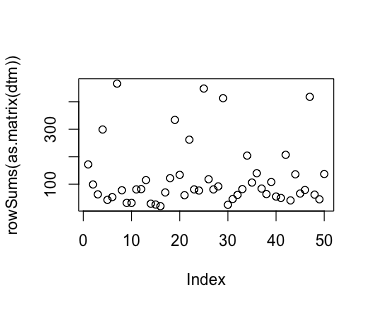
* Creating a TermDocumentMatrix tdm using acq dataset
* wc is word count per document
* wc\_ord is word count sorted in decreasing order to get 10 largest documents



* The output of the script above is as follows:
* wc is the undordered word count in the 50 documents
* wc\_ord is the top 10 largest documents

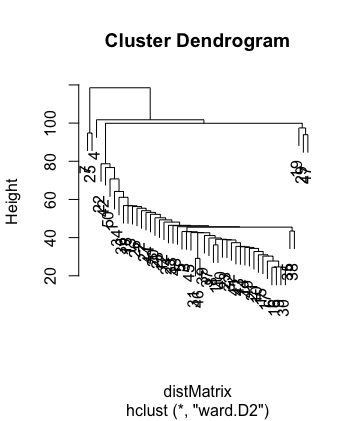


* This plot is the of words on Y and document on X



1. For each document work through the examples given in Lecture 7 to display the dendrogram and the WordCloud

Dendrogram is as follows:

****

Word Cloud is as follows:

****



1. Prior to removing the punctuation, find the longest word and longest sentence in each document from the 10 largest document

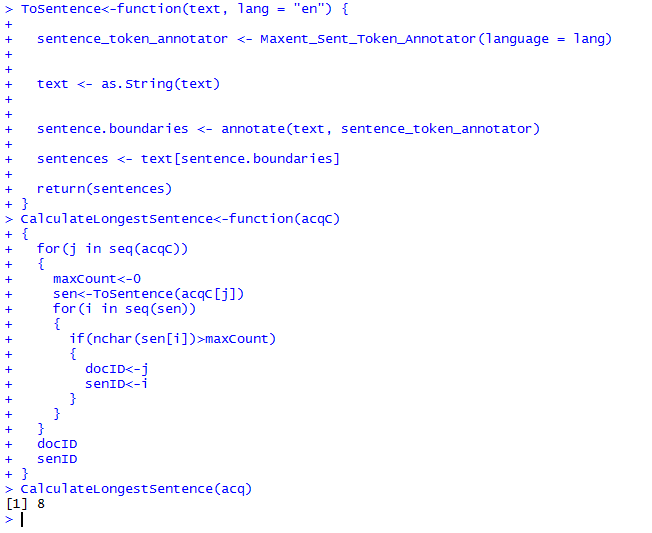
To do this task we have created two functions:

1. ToSentence
2. CalculateLongestSentence

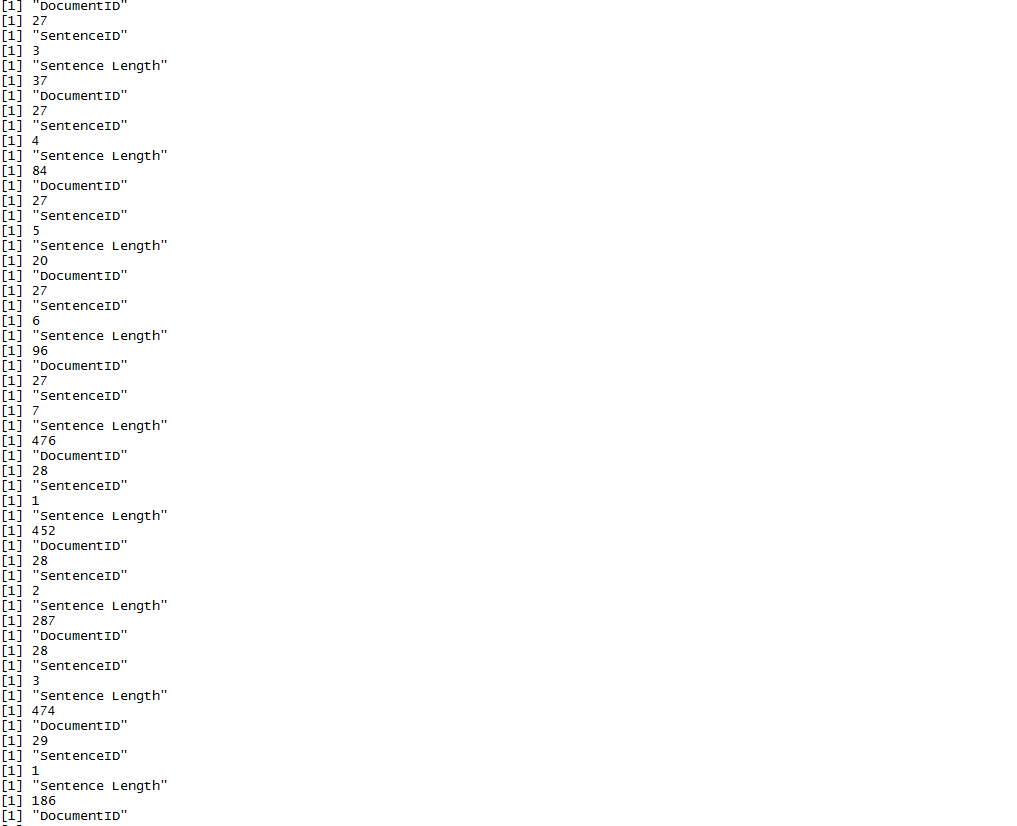
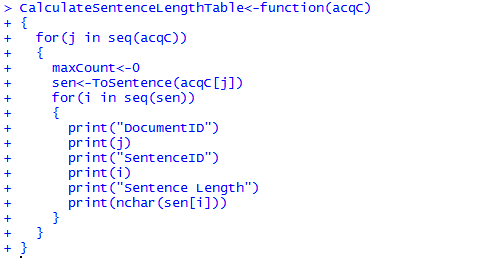
ToSentence: This function converts the text in the argument to sentences

CalculateLongestSentence: This function iterates into each of the 50 documents and then identifies the longest sentence in the top 10 largest documents which is then displayed in terms of document ID and sentence ID

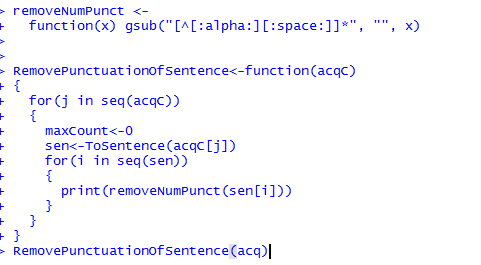
The document Id in our data set came out to be 1 and the 8th sentence is the longest in out dataset.



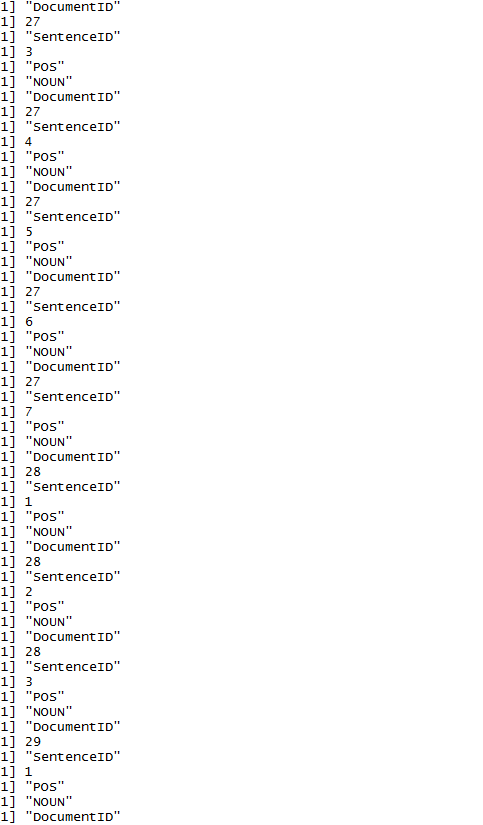
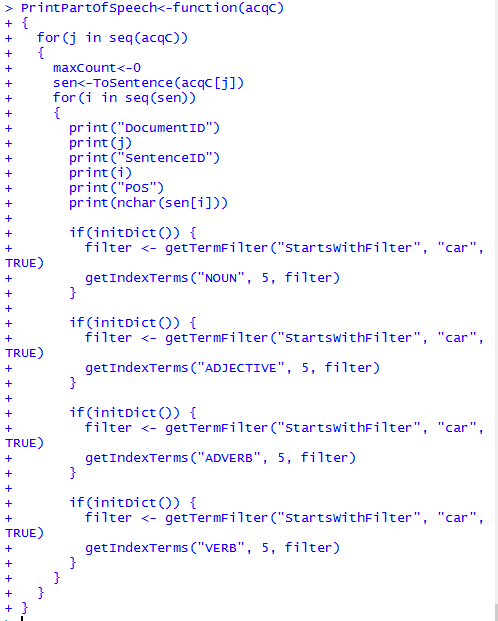
1. Print a table of the length of each sentence in each of the 10 documents.



1. For each sentence of each document, remove the punctuation. Display the sentences.



1. For each word print its part of speech using the Wordnet package.



1. Analyze word frequency using functions from package zipfR.

