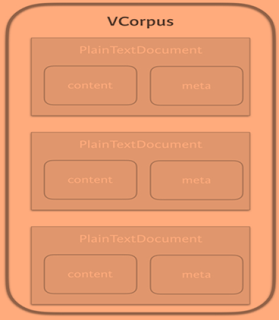
**Text Mining**

The first thing in text mining is to employ common data cleaning functions in R to get the data prepared for text mining (see Table 1). Next, we extract different corpus for each specific question, with the term corpus defined as a collection of written text.

The Corpus object is a nested list, or list of lists. At each index of the Corpus object, there is a plain text document object, which is essentially a list that contains the actual text data (content), as well as some corresponding metadata. For more information please click this [link](https://eight2late.wordpress.com/2015/05/27/a-gentle-introduction-to-text-mining-using-r/).

[](https://1.bp.blogspot.com/-ffjMnr3w18c/WKJBVZ08gfI/AAAAAAAABKM/4eRhxCJO2lIm9t7IkqPsbjzkR7IIy44UgCLcB/s1600/amit.png)

**Table 1. Common pre-processing functions used in R to clean the data.**

|  |  |
| --- | --- |
| tolower(): | Make all characters lowercase |
| removePunctuation(): | Remove all punctuation marks |
| removeNumbers(): | Remove numbers |
| stripWhitespace(): | Remove excess whitespace |

We create a list of words that are frequent, but provided little information (stop words). Next, we remove these stop words from the text and then use the tm\_map() function from the tm package in R to apply cleaning functions to a corpus. We proceed to create a term document and using the as.matrix() function in R create a matrix from the Term Document Matrix.

To extract useful phrases, which lead to some additional insights or provide improved predictive attributes for a machine learning algorithm, we may use the RWeka package to create bigram. Bigram’s provide some additional insight and improve predictive attributes and then we apply rowSums() on newly made matrix to aggregate all the terms used in a passage. Once we had the rowSums(), we sorted them to focus on the most frequent terms.

**Term document matrix**

A term document matrix (see Table 1) is a list of all word occurrences in the corpus, by document. The term-document matrix is a two-dimensional matrix whose rows are terms (i.e., each subject response to that question) and columns are responses (i.e., whether or not that word/term was used). The term document matrix describes the frequency of terms that occur in the collection of text. The plot of term document matrix gives the frequency of words that appear in the text and its importance.

Term document matrix also give the association of words with each other. The word association plot explains how the words are inter connected in the comments. The line joining them give the strength of the association between terms. Thicker line shows the strong association between those words. It is useful in text mining to visualize the relation between words in the text.

**Table 2. Example of Term Response Matrix**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Response 1** | **Response 2** | **Response 3** | **…** | **Response n** |  | **Sum** |
| **Term1** | 0 | 0 | 1 | 0 | 0 |  | 1 |
| **Term2** | 1 | 1 | 0 | 2 | 0 |  | 2 |
| **…** | 1 | 0 | 3 | 0 | 0 |  | 4 |
| **Term n** | 0 | 1 | 0 | 0 | 1 |  | 2 |

**Bigram, word cloud, and dendrogram analyses**

Word cloud (for more information see Appendix B) is the common technique in text mining to extract the prominent words from the text, with the bigram bar charts providing the frequency to which those terms/words occur. The cloud gives the prominence to the words that frequently appear in the text. Larger word font size indicates higher word occurrences in that open-ended question, which correspond to greater importance.

There are various methods of constructing word clouds, as it can be done using a single word selection method, two words at a time, phrases and so on. However, frequent word counts can mask useful insights of the text at times, thus Bigram analyses can give better understanding of how the words used in context over single term analysis. More specifically, the Bigram analysis will help us to extract useful phrases which offer additional insights and provide improved predictive attributes for a machine learning algorithm. The dendrogram, which is a branching diagram, can then be used to show the connections/relationships between terms/words.