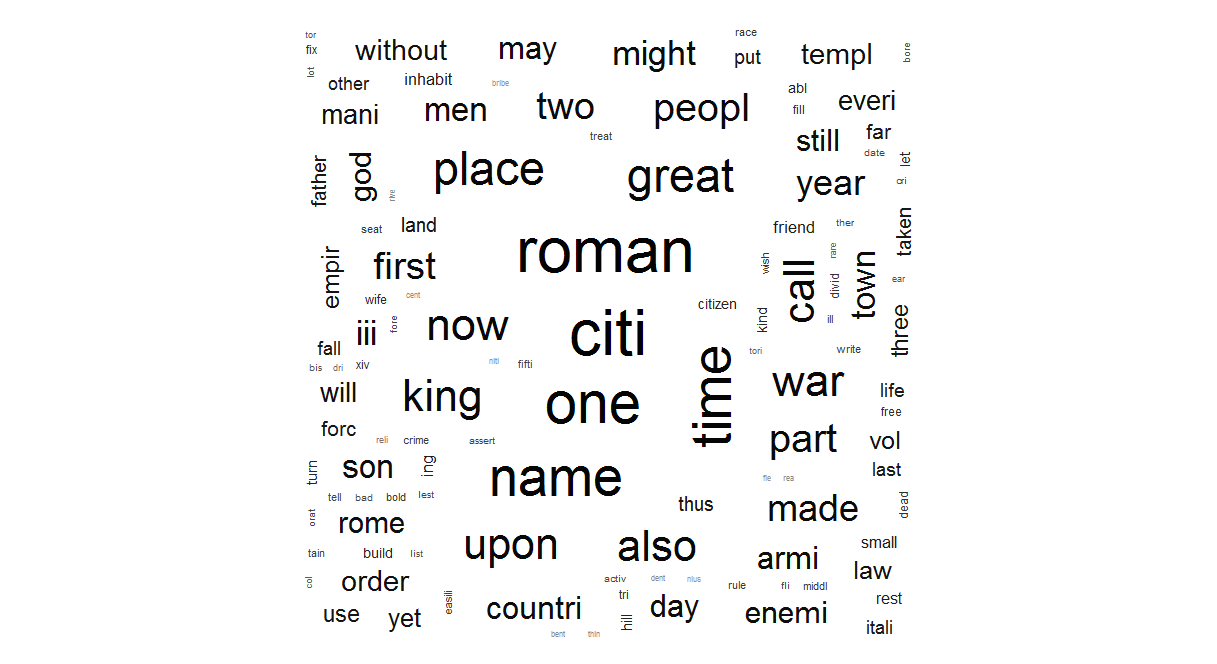
Text Mining

Processing and cleaning:

* Html.raw: extract all text from the documents
* Html.parse: get only words with no html code about the format etc (regular expressions) and create a nested list: each element a document with a list of the words in it.
* Convert to Corpus (to be able to clean with ‘tm’ package: collection of documents
* removePunctuation, removeNumbers, convert to lower case, remove stop words (using an English dictionary in R)
* stem document
* strip white spaces (generated when applying all these cleaning functions)
* convert to text document to be able to perform analysis

Analysis:

* create a document-term matrix: rows document, columns words.
* Study frequency of words and explore ur bag of words: saw that there are many terms scanned in a bad way (some symbols such as squares, €, circles, nonsense words), considering that we can clean this matrix by removing very sparse terms. When doing this is very likely that some real words are going to be deleted, but as they are sparse terms is not going to have a noticeable effect on our analysis (dealing with huge documents). In addition, according to Zipf’s law, to study a document we not interested in rare words or words that appear a lot in every document (verbs, stop words, etc) as they are not going to add value to the relationship between them: Rare words are going to appear on one document so not going to relate to another, the other way very common words are going to appear on every document many times.
* Remove very sparse terms: it fits better: just leave <10% sparse terms
* Name the documents in the matrix
* Tf-idf: weight the words so if a word appears in all documents is not very important in order to differentiate them: explain problems (words with many occurrences… lect9). Maybe using another method: BM25?
* Cloud of words (double check of relationships of documents)

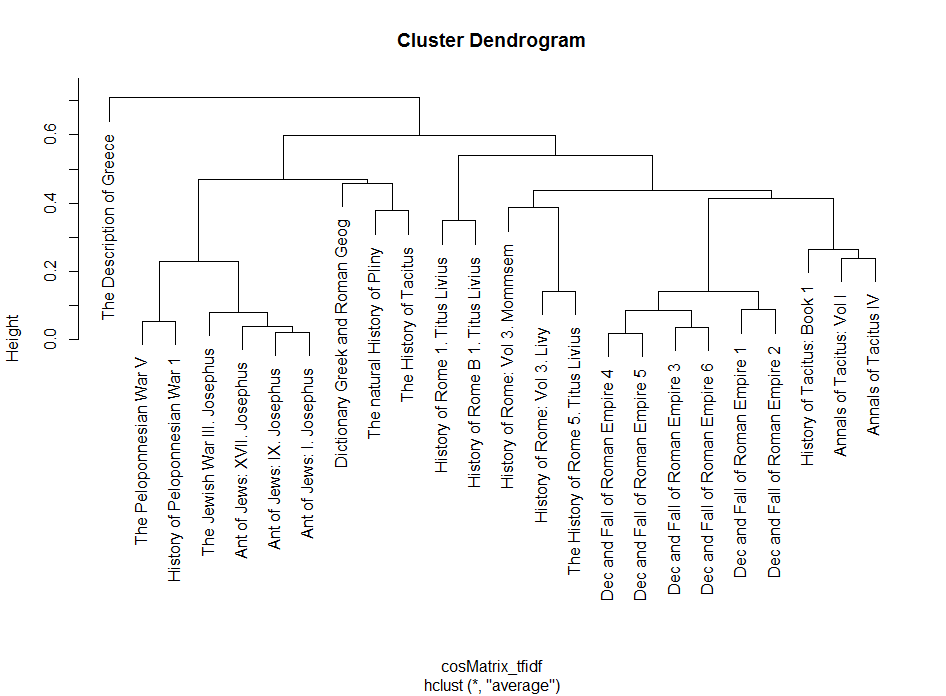


Here we can see that there are many terms that appears lots of time and are meaningless: one, name, place, great… WE can try to remove these terms after all te analysis and see if all the clusters are better organized.

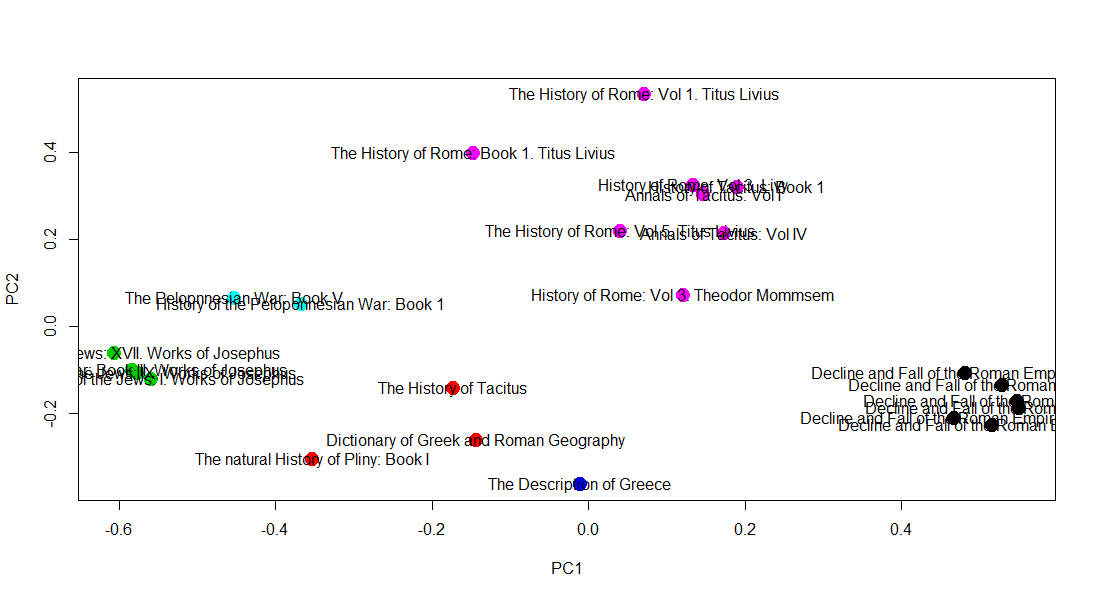
* Hierarquical clustering / other algorithms

Using function hclust: agglomerative clustering, so where we start out with each document in its own cluster. The algorithm iteratively merges documents or clusters that are closest to each other until the entire corpus forms a single cluster. Each merge happens at a different (increasing) distance. So backwards clustering, starting from treating each document as a cluster itself.

Seems it works good, still there is one of the Josephus books (jews) that is clustered in decline and fall of rome. Dictionary is apart from everyone else, that makes sense as it is going to have a lot of sparse terms that in the other books are not going to even appear.

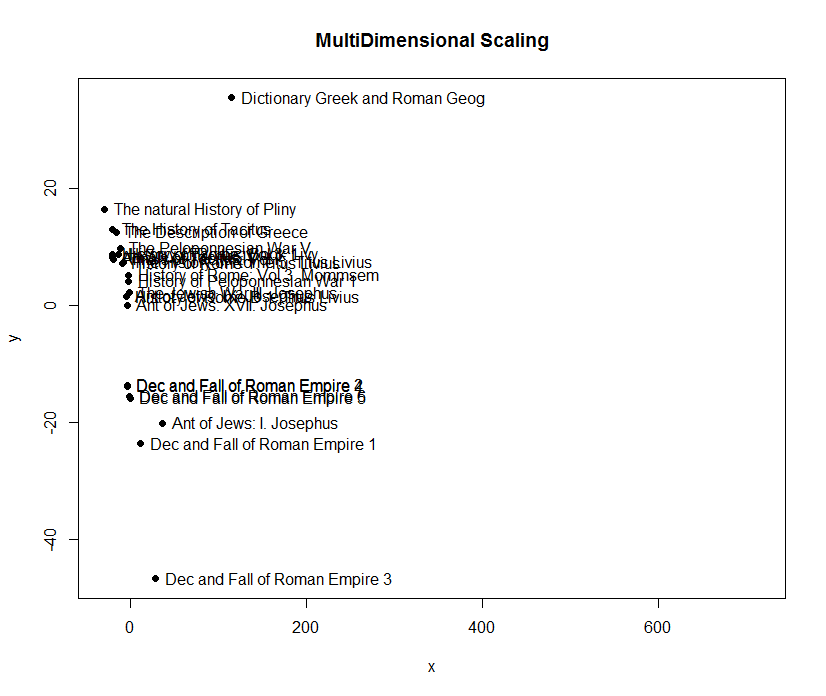


* K-means clustering



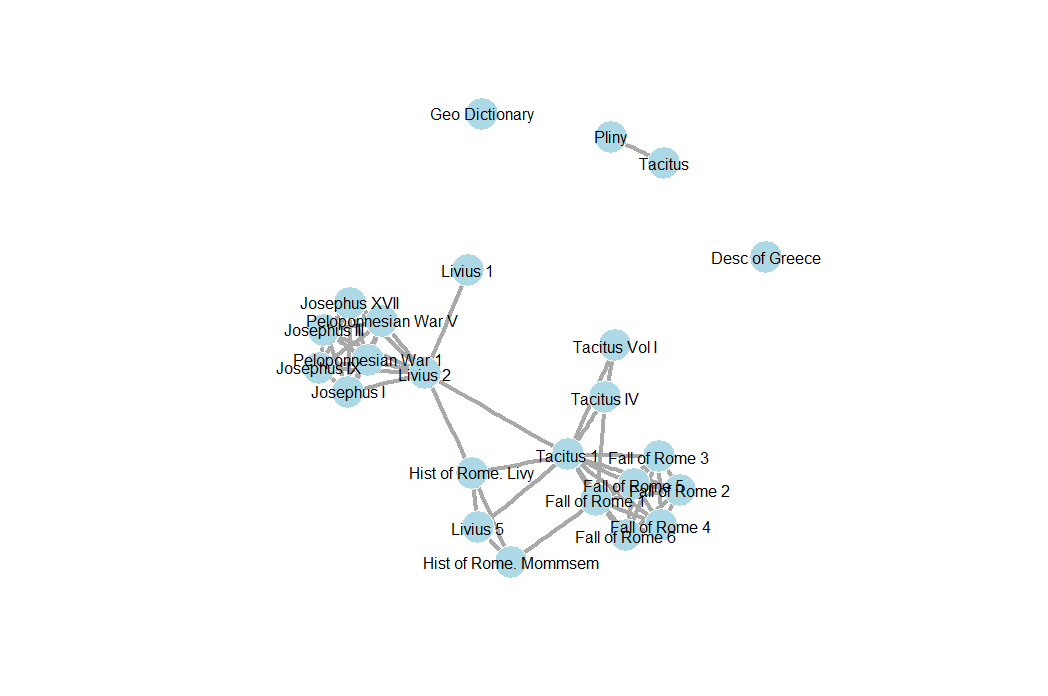
K-means clustering (K=6) using Euclidean distance. very dense matrix (>10% sparse terms) and tf-idf weighting.

* MultiDimensional Scaling



distMatrix, very dense matrix

* Network Map



Plotting correlation matrix: (1 – cosineSimilarity), very dense matrix, tf-idf

* Proximity search?? 🡪 this will refine our algorithm for sure
* SVD (singular value decomposition), semantic analysis?

After some exploratory analysis and first hierarchical algorithm we can spot expected clusters (books from the same author and/or same title but different volumes).

* Decline and Fall of the Roman Empire by Edward Gibbon (6 Volumes): 98 to 1590 AD
* Tacitus: Annals(14 AD - 68 AD), 1 book and History (69 AD – 96 AD), 2 books.

Roman senator and historian, makes sense that all these previous books are in the same big cluster as they are ordered chronologically, if we have a closer look the smallest distance between these groups is between the 1st and 2nd books of decline and fall and Tacitus, as the first ones are following Tacitus History.

The only ‘mistake’ classifying here is one of the books classified nearer to the dictionary and the natural history book.

* History of Rome: Titus Livius (2 Books) 752 BC – around 15AD.

This author wrote about ancient Rome: the Roman Republic most of it (just 40 years of Empire). Then it makes sense that this cluster appears far from the previous ones as they are about the Empire.

* History of Rome: Mommsem. It also deals with the Roman Republic. That is the reason it appears clustered with Livius works and further from Gibbon’s and Tacitus works.
* Natural History of Plinius (about multiple scientific and technological subjects). This Book 1 is the preface of all the natural history, probably speaking about the context of all the subsequent books. If it is like this then it speaks about his historical context (around 50 AD) and table of contents. Its character of natural history book makes it to be close to the dictionary.
* Josephus books are about the war between Jews and Romans and the history of the world from a Jewish perspective. Then they treat basically the same topics than the previous ones but from a different perspective, this is the main reason that they are far from the cluster of Roman History.
* Peloponnesian war was between Greece and mainly Sparta, that’s why it is far from Roman History and close to Greek books such as Description of Greece, that is isolated.