Text Analysis Project

- **Rajat Katyal , MDSI UTS**

Introduction

This report deals with analysing a collection of 42 Documents provided, to gain some meaningful information out of it. In the report a few Text Analytics techniques are used to gain insights. Text Analytics, also known as text mining, is the process of examining large collections of written resources to generate new information, and to transform the unstructured text into structured data for use in further analysis (Linguamatics.com).

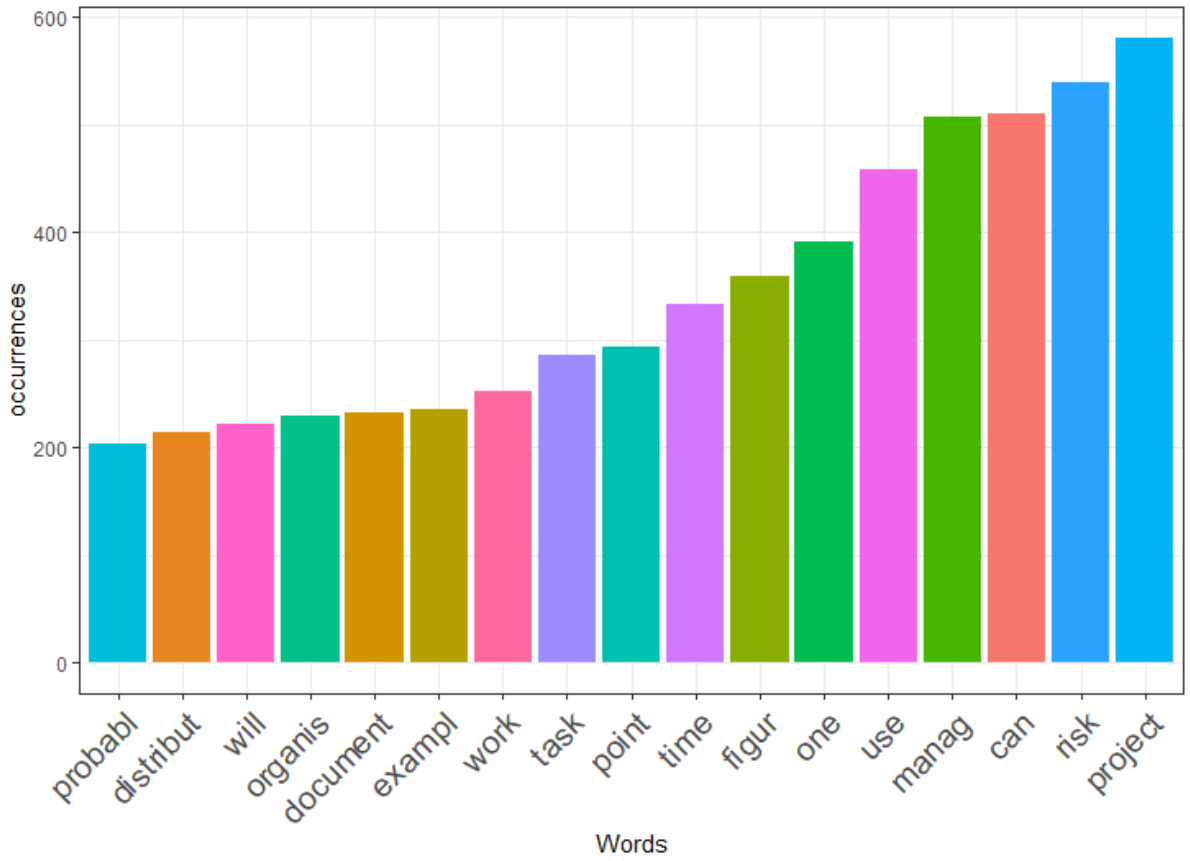
Data Cleaning:

The documents were in text format, so to begin with, the data is cleaned using basic structuring techniques. The following operations are performed on the set of data:

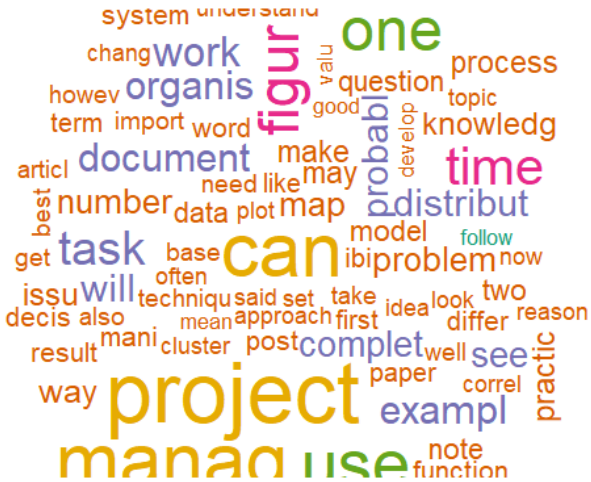
1. Punctuations are removed
2. The words are moved to lower case
3. Numbers are removed
4. Common English words are removed (for example the, for, as, and)
5. Document's blank spaces are removed
6. The words are stemmed of the suffixes (for example 'working' becomes 'work')
7. Special Characters are removed (for example 'â€')

Data Analysis

Now, we look at the most frequently occuring words within the set of documents. Here we can observe the meaningful words with maximun occurance, for example project, risk, manage, task, document. This gives us a fair idea about what the documents must be talking about.



We use a word-cloud chart to show these highly occurring words. Here we could plot more words and the size of the word is proportionate to it's occurrence. For example, the word "task" occurs more than the word "result" based on their size.

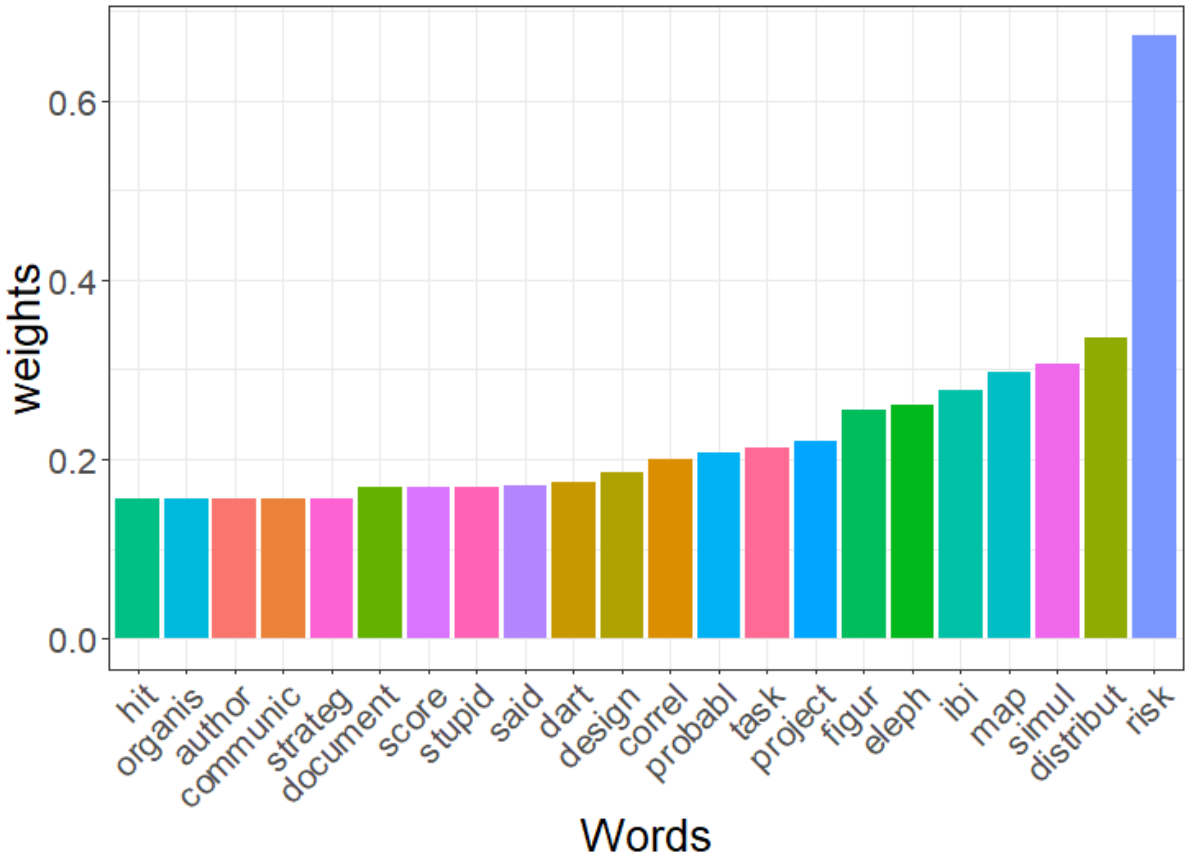


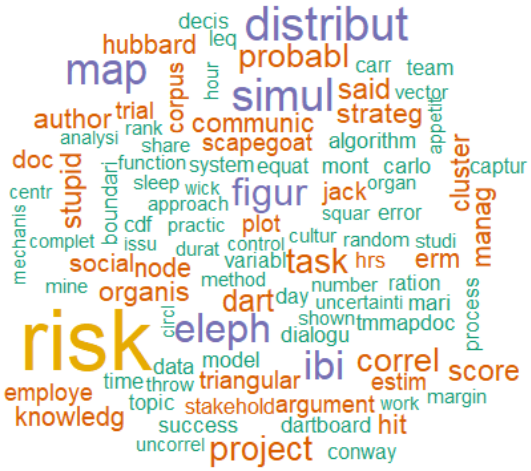
There is a better technique that can be used for analysis, Term Frequency-Inverse Document Frequency. Here a tf-idf weight is calculated, which is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus (tfidf.com).

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).

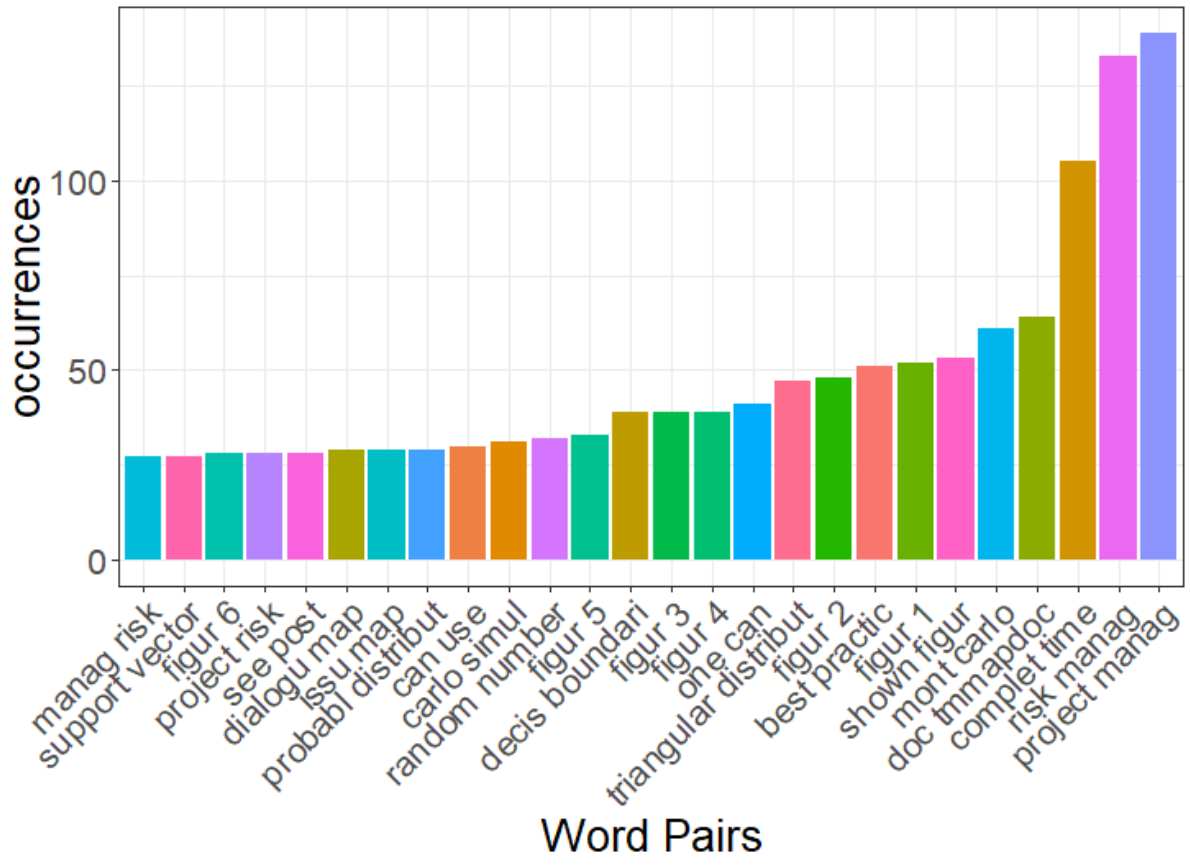
IDF(t) = log_e(Total number of documents / Number of documents with term t in it).

Using this technique, we find the words with higher weights. Below we can observe that “risk” has a higher weight than “project” which was the most frequent word in our documents. Below we can also see the word-cloud for high weighing words in our set of documents with their size proportional to the weights.

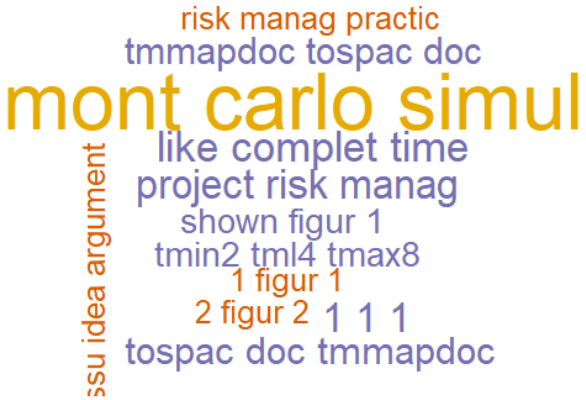




Similarly, we also check which pair of words occur most frequently in the set of documents. Below we have plotted a bar chart and a word cloud for the most commonly occurring pair of words. Here it becomes more clear that document may be talking about “Project management”, “Risk management”, “monte carlo” and “random numbers”.



We further check for a trio of words that occur together most frequently. This gives us some idea about the key topics of these documents.



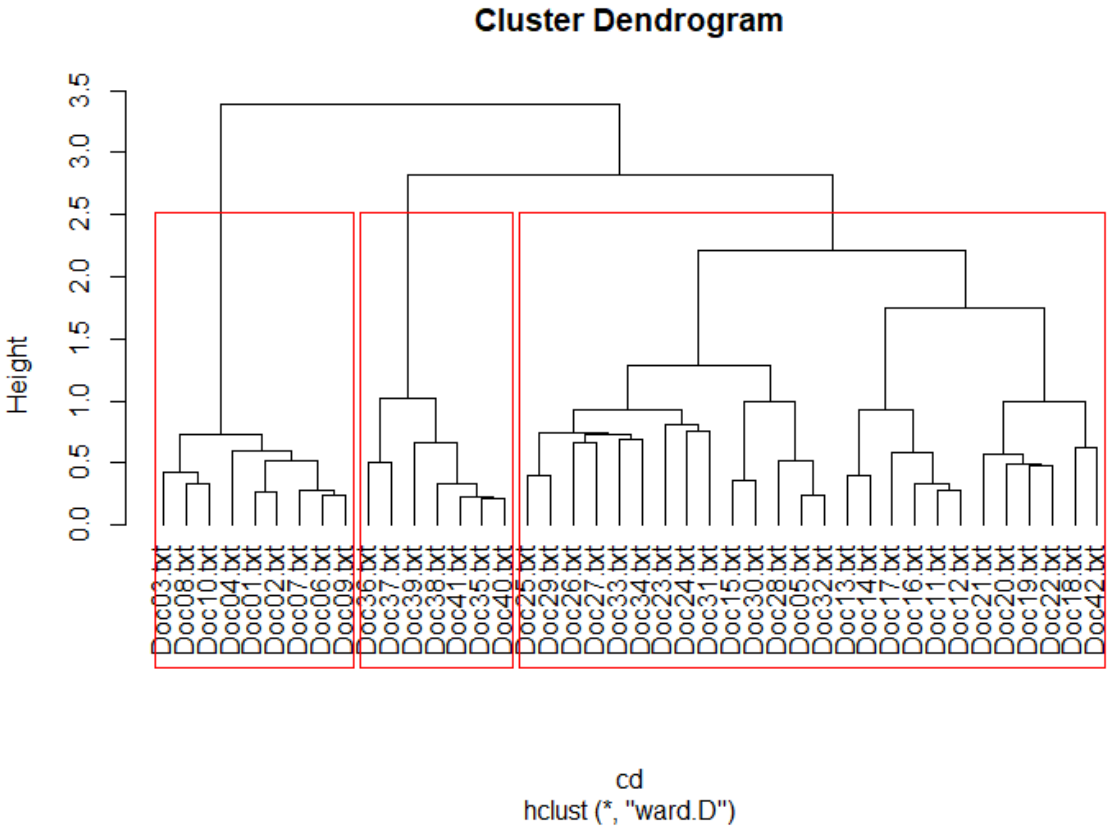
Document Clustering

There are couple of key techniques that are used for document grouping or clustering:

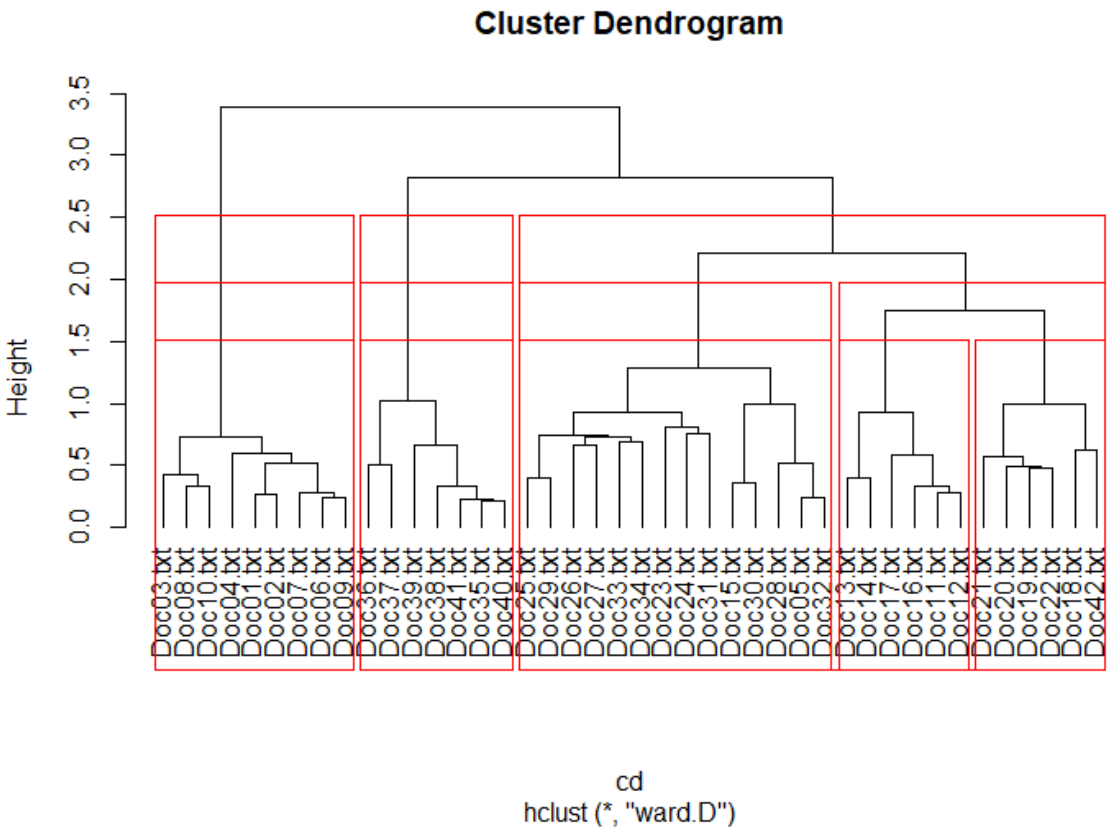
1. Hierarchical Clustering
2. K-Means Clustering

K-Means clustering works by finding cluster centroids of specified K clusters while Hierarchical clustering works by considering each document as a cluster then merging two nearest clusters. K Means clustering requires prior knowledge of K i.e. no. of clusters you want to divide your data into. In hierarchical, you can stop at whatever number of clusters you find appropriate in hierarchical clustering by interpreting the dendrogram.

We have used hierarchical method to find out how many document groups can be best identified. Below we can see that 3 groups as the most ideal way (by comparing largest overlapping height of clusters), of grouping documents in clusters. Groups of documents can be identified by the Red-marking lines.



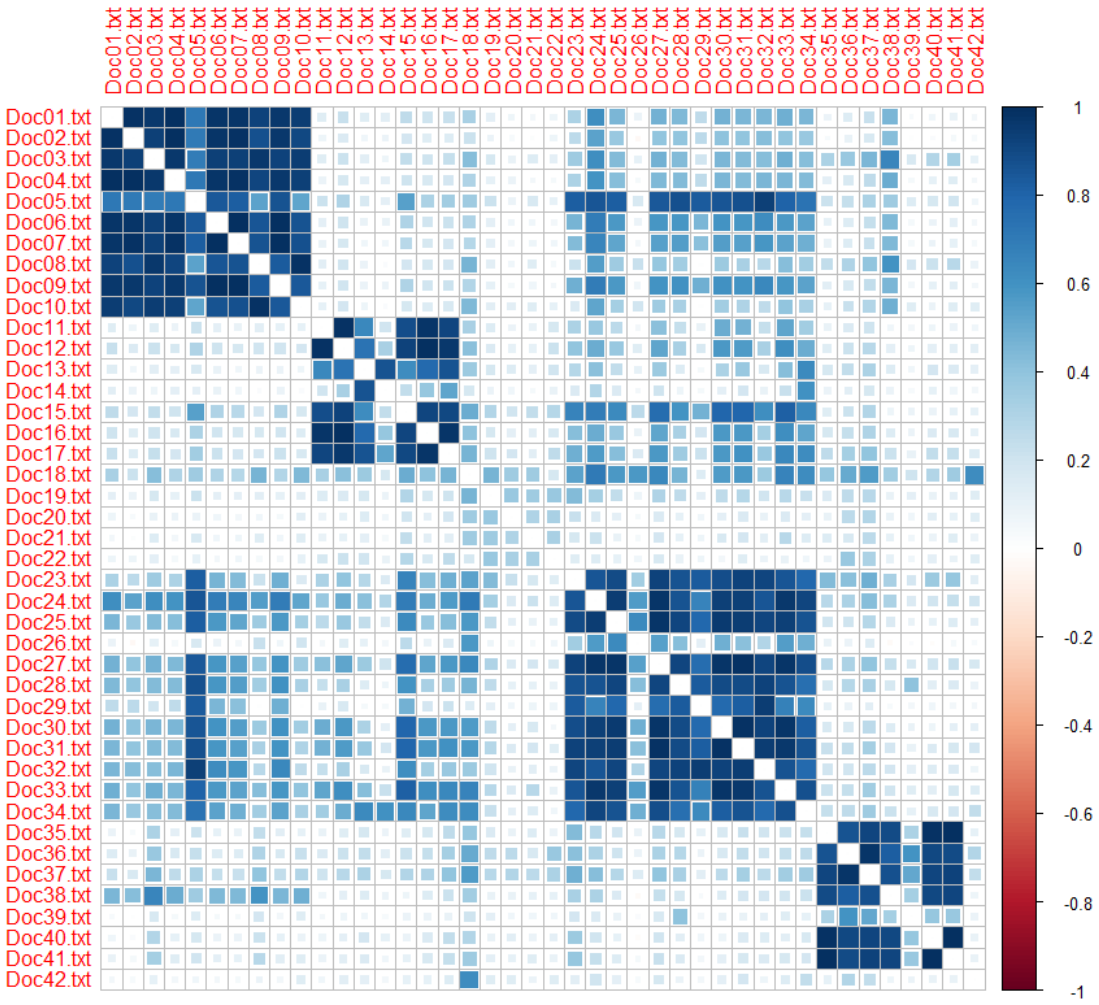
However, as there isn't a significance difference between clustering heights, the documents could be explained by 2-5 clusters. Below we will see how these can be grouped into 5 clusters. We observe be looking at the red lines for clustering, Document 1,2 and 3 are in one group while Document 11, 12, 16 are in another group.



We have seen some basic analysis on the common words and word combination. We have also seen how the documents can be combined into 2-5 groups or clusters. Now we will explore the common themes among the documents and see if our grouping makes sense with the topics.

Topic Modelling and Latent Semantics Analysis

Topic modeling is a method for unsupervised classification of such documents, similar to clustering on numeric data, which finds natural groups of items even when we're not sure what we're looking for. LSA produces measures of word-word, word-passage and passage-passage relations that are reasonably well correlated with human cognitive phenomena involving association or semantic similarity. Here we use Latent Semantic Analysis to calculate the similarity between the documents. The similarity matrix is then plotted using a correlation plot below. Here we can see the darker share of blue means higher similarity between the documents.



From the figure above it is very clear that there are 4 well correlated set of documents or groups.

1. Group 1: Document 1 to Document 10
2. Group 2: Document 11 to Document 17
3. Group 3: Document 23 to Document 34
4. Group 4: Document 35 to Document 41

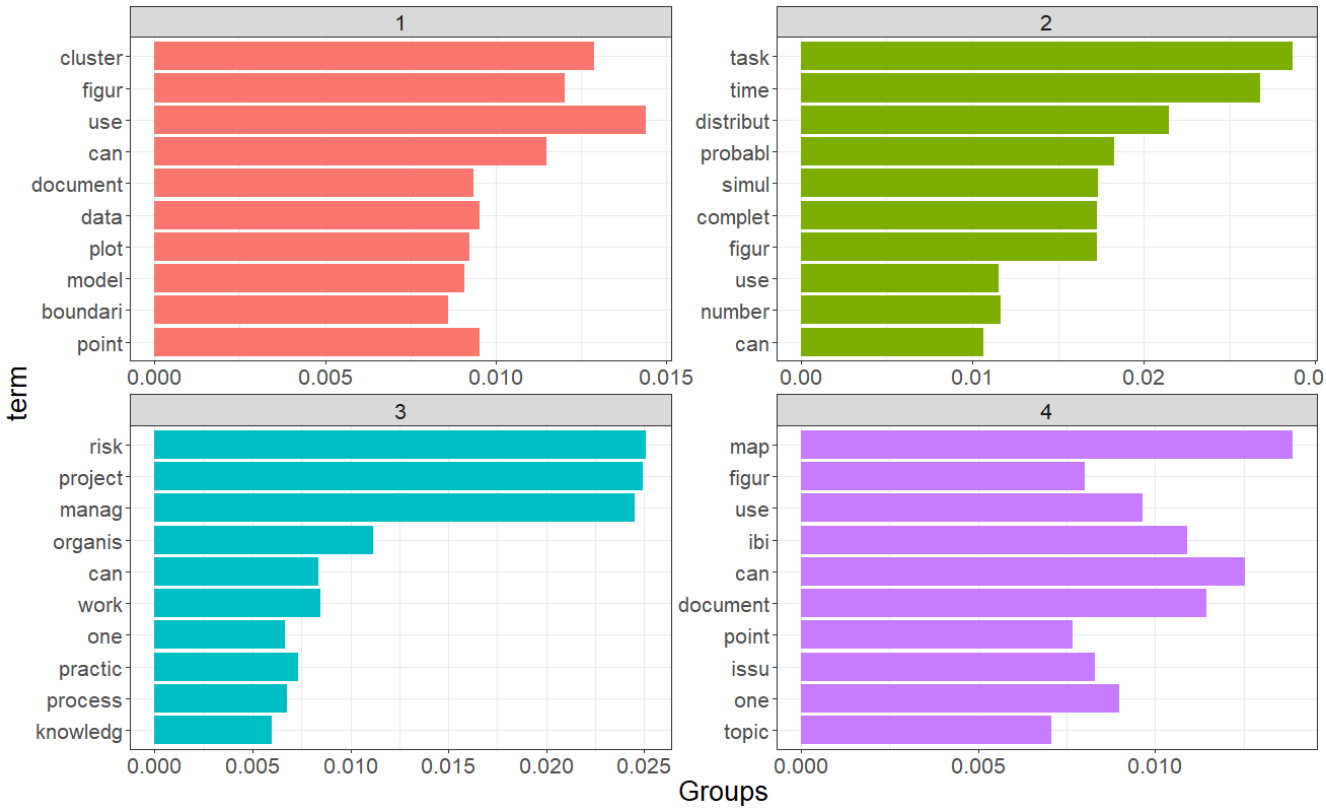
We can also identify a group of documents which are not similar to any other documents

5. Group 5: Document 18 to Document 22

We also observe that Document 5 may identify well with Group 2

Word-Topic Probabilities

Based on our clustering analysis and LSA we could find up to 5 groups of documents, 4 showing similarity while one unsimilar group. Now, we look for Words with high probability as a topic. Since, we have 4 groups we start with a 4 topics approach seen below.



Here we can find four different topics.

1st Topic: It has words like “cluster”, “model”, “boundary”, “plot”, “data”. This tells us that this group maybe talking about data modelling of documents using clusters and boundary points. This points to text-mining material.

2nd Topic: Here the key words are “probability”, “distribution”, “number”, “figures”. The topic seems to be about mathematics and talking numbers and probability distributions.

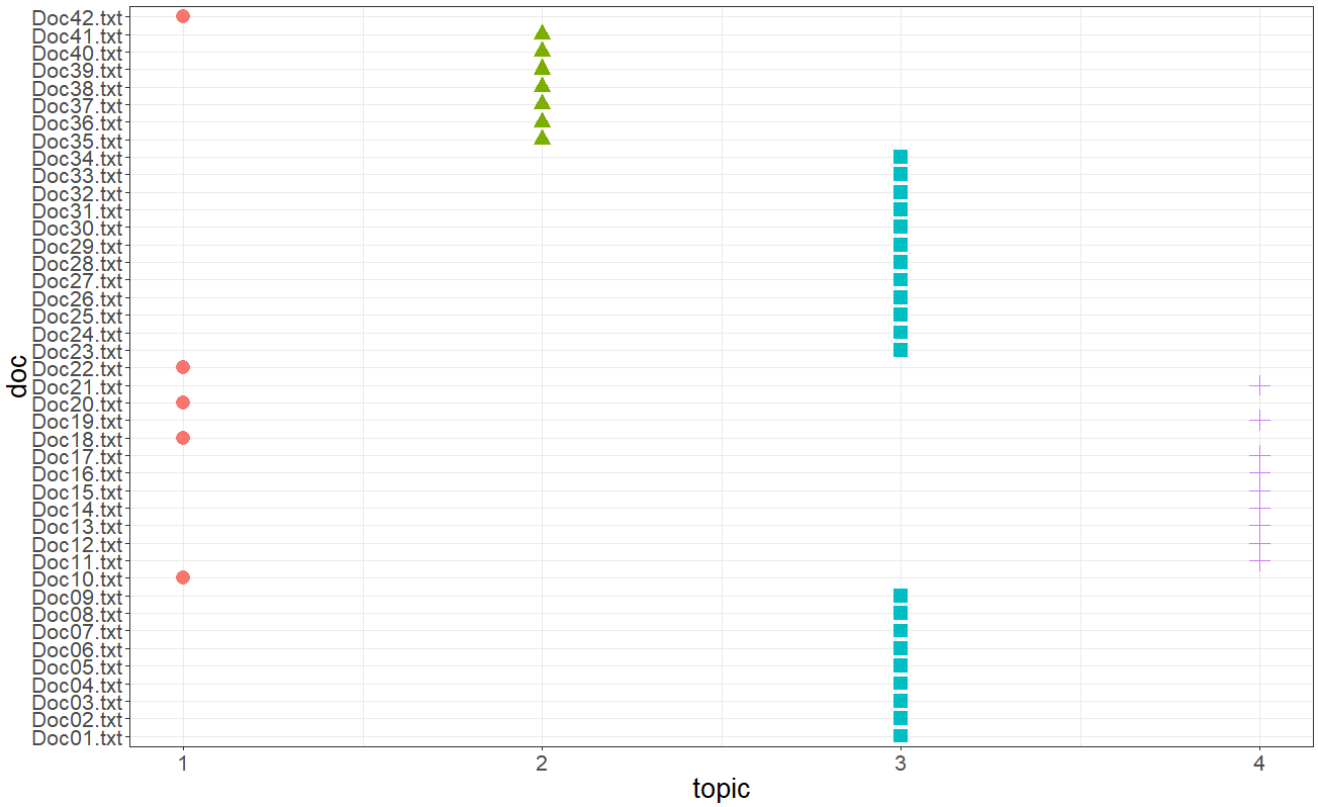
3rd Topic: This topic has some key words as “risk”, “project”, “manage”, “organise”, “process”. The topic seems to be about organisational processes, project risk management and work practices.

4th Topic: This topic has key words like “topic”, “document”, “point”, “map”, “figure”. Here the topic seems to be about a Document Mapping with figures and topics. Though this isn’t very clear.

One thing to note is that Topic 1 and Topic 4 have similar key words like “point”, “document” and it may mean they both point to a same topic. But given the granularity of our clusters, we should try to plot documents onto at least 4 topics to gain some insights.

Document-Topic Probabilities

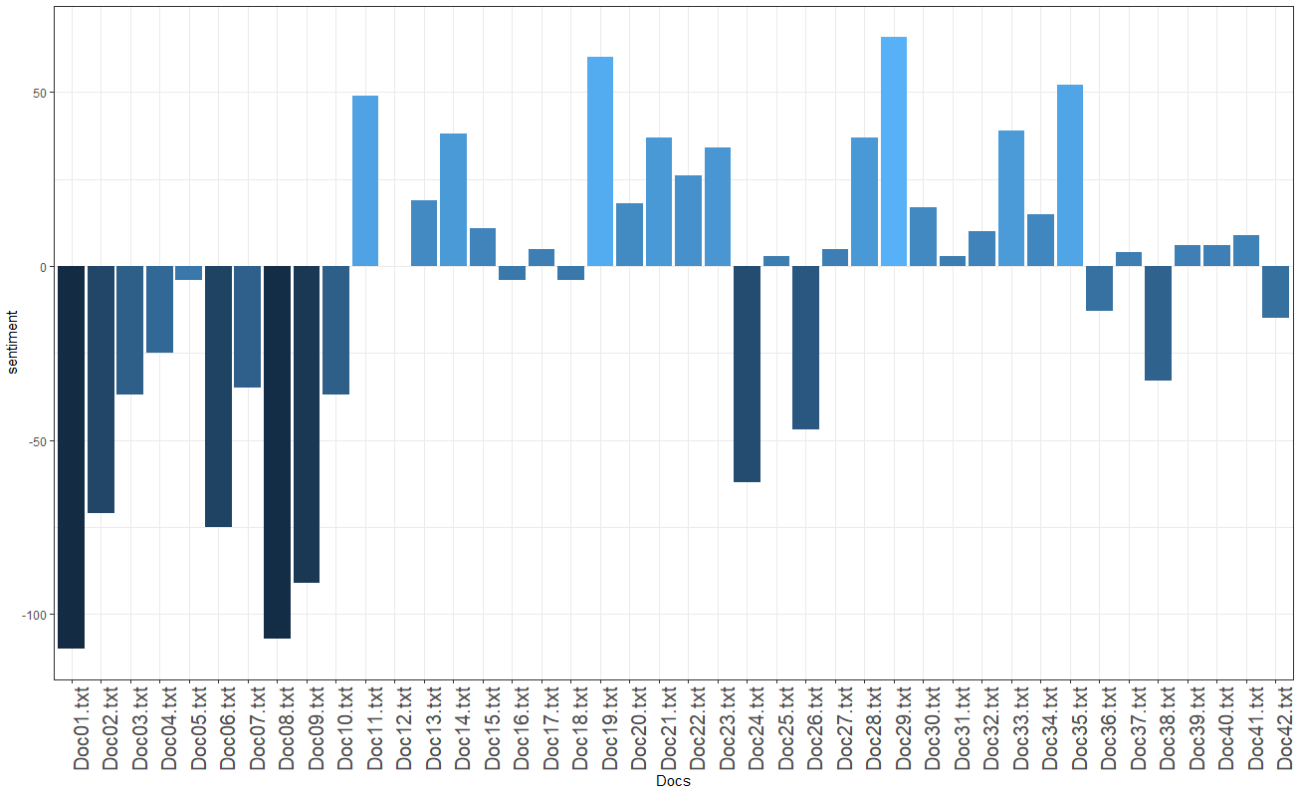
We categorise the Documents as per the word topics to see which Topic they belong to. In the chart below we can see that, Our Group 1 and Group 3 identify with topic 3 while Group 4 identifies with topic 2. We can also see that our Group 2 identifies with topic 4 and Group 5 (non-similar documents) identify between topic 1 and 4. This also highlights that our non-similar Group 5 identifies closely with Group 2 and both have similar topics as seen earlier.



Since, large part of documents Group 1 and 3 are about the same theme, it maybe interesting to note why they are in different groups.

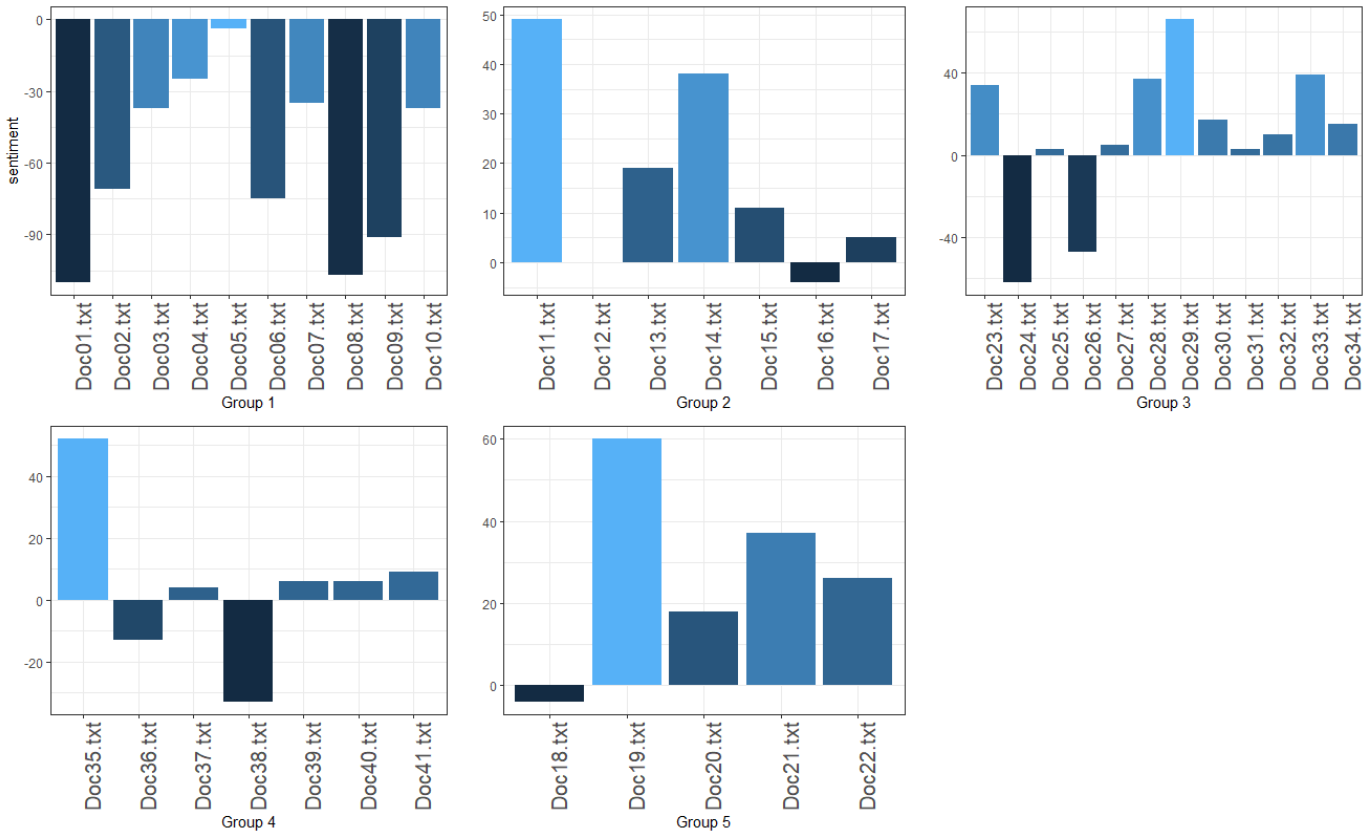
Sentiment Analysis

We have divided the documents based on the pre-defined “nrc” lexicon sentiments. For example, words like “good” and “happy” have positive sentiment while “risk” and “fear” are negative. Now based on the number of positive and negative words we have ranked the documents to see what overall sentiment they relate to. In the below chart we can see different Groups having different sentiments.



Although Group 1 and Group 3 have the same themes, here we can see that they both have completely opposite sentiments. This highlights the possibility that the Group 1(Doc 1 – Doc 10) maybe talking about the negative aspects from our theme like project risk and risk management, while Group 3 (Doc 23 – Doc 34) maybe talking about positive work practices and organisational knowledge management. We can also see that Group 4 (Doc 35 – Doc 41) has negligible sentiment, probably owing to mathematical literature as seen previously.

We can see the sentimental differences among Groups more clearly in chart below.



Conclusion

Using text mining techniques, the documents most frequent and descriptive words were identified. Frequent word pairs and trios were also identified. The 42-document set was divided into 5 groups using clustering and LSA. The document Groups were mapped into 4 topics using high probability words. The text sentiment among document groups was identified and similar themed groups were further categorised with sentimental analysis.