2018

DATA, ALGORITHMS AND MEANING

ANALYSIS OF UNSTRUCTURED DATA

LHOGESHWARAN PURUSHOTHAMAN (13313491)

UNIVERSITY OF TECHNOLOGY | FTDI | MDSI

Data, Algorithms and Meaning – Spring 2018

Assignment 3 - Part A | Analysis of Unstructured Data

Author: Lhogeshwaran Purushothaman | 30 – October - 2018

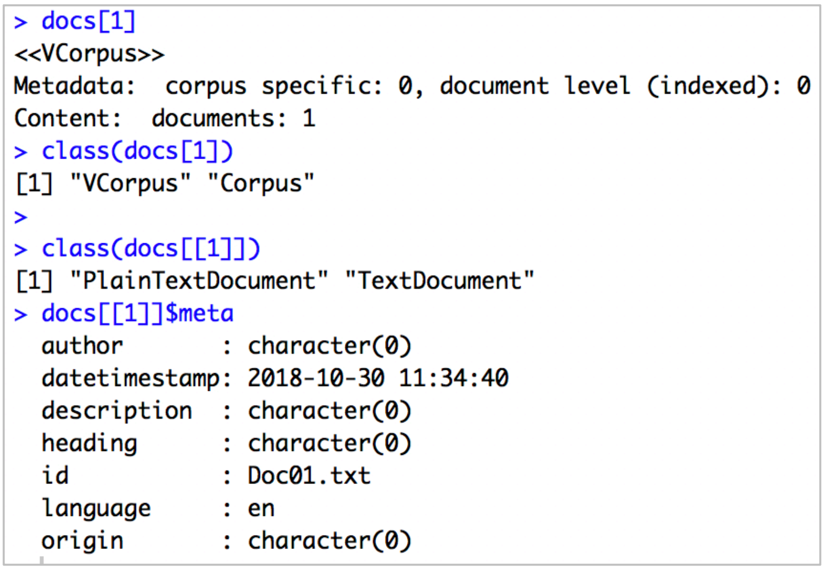
Understanding of the task

The task is to think from the shoes of an aspiring Data Scientist, to help my Manager get insights on a huge collection of his old files. Taking on the challenge from my Manager – '*provide insights into the contents and themes of the documents in the directory* ', I have to leverage the new techniques on text analysis I learned over the recent past. While I am at it, I plan to take my time to understand the data, then begin the process by preparing the data for further analysis and finally deploy the techniques of clustering and topic modelling to dig out the hiding wealth of knowledge from the corpus. I have to be conscious to take a professional approach and present the insights in a manner a professional Data Scientist would, which I plan to do following the CRISP-DM methodology, as this task could be the key for me to secure the new Data Science role in my organisation.

The Given Dataset

The Raw Data

The file shared has 41 text documents. The names of the document do not provide any clue as to what might be the content of the file but appear only to follow a standard naming convention like 'Doc##.txt' with ## ranging from 01 to 41. For further analysis using the statistical text analysis methods, the files are loaded into R-Studio to create a corpus, which is nothing but a term for collection of articles or documents usually written by the same author.



```
> docs[1]
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 1
> class(docs[1])
[1] "VCorpus" "Corpus"
>
> class(docs[[1]])
[1] "PlainTextDocument" "TextDocument"
> docs[[1]]$meta
  author       : character(0)
  datetimestamp: 2018-10-30 11:34:40
  description  : character(0)
  heading      : character(0)
  id           : Doc01.txt
  language     : en
  origin       : character(0)
```

Preparing the data

Dirty data is potential to cause misconception of information, therefore the first and foremost step for any analysis would be cleaning of data. Preparing of this text dataset is carried out in the following steps –

Data cleaning

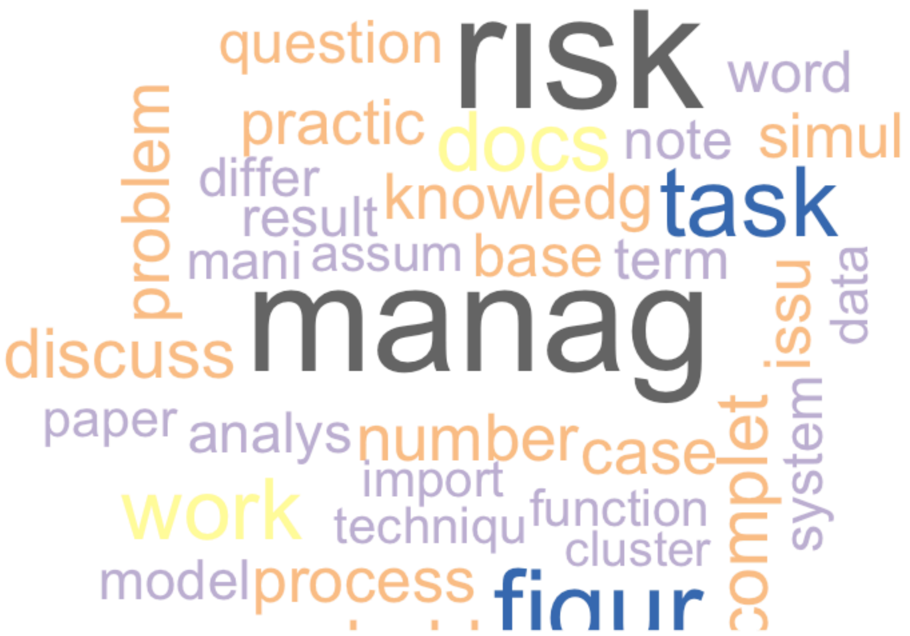
- Remove punctuations
- Remove numbers (under consideration, this being a text analysis process, numbers wouldn't contribute considerable meaning)
- Remove stop words (Ex: an, the, enough, though, etc.) and any word lesser than 4 or 20 alphabets long

Data transformation

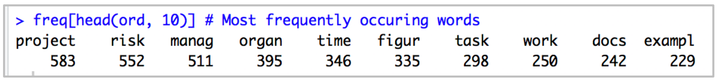
- Convert all text to lowercase (as R being case-sensitive reads 'tom' and 'Tom' as two different characters)
- Stem the words to root-word (Ex: organization, organize -> organ)

Text analysis

The analysis is begun by analysing the corpus for most frequently and less frequently used words.



Most frequently used

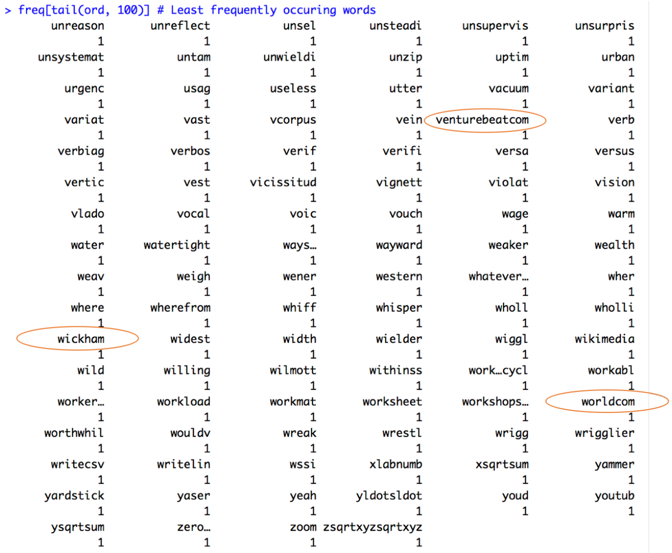


```
> freq[head(ord, 10)] # Most frequently occuring words
project    risk   manag   organ    time   figur    task    work    docs  exampl 
    583     552     511     395     346     335     298     250     242     229 
```

Looking at the most frequently used words, the general subject of discussion in all the documents seem to be around project & risk management, figures and documents quantifying the organizations path.

Least frequently used

At times, it could be the least frequent words used in a document that reveals something unique about it.

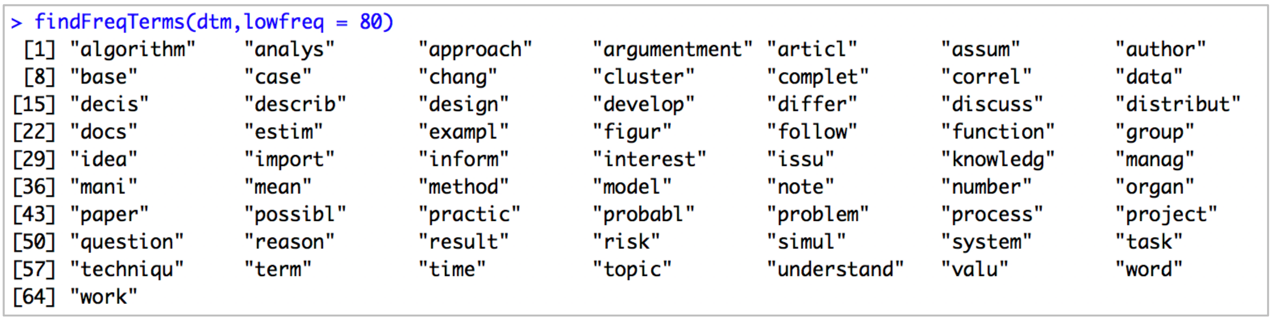


```
> freq[tail(ord, 100)] # Least frequently occuring words
        unreason       unreflect           unsel        unsteadi      unsupervis       unsurpris
               1               1               1               1               1               1
      unsystemat           untam        unwieldi           unzip           uptim           urban
               1               1               1               1               1               1
          urgenc            usag         useless           utter          vacuum         variant
               1               1               1               1               1               1
          variat            vast         vcorpus            vein  venturebeatcom            verb
               1               1               1               1               1               1
         verbiag          verbos           verif          verifi           versa          versus
               1               1               1               1               1               1
          vertic            vest      vicissitud         vignett          violat          vision
               1               1               1               1               1               1
           vlado           vocal            voic           vouch            wage            warm
               1               1               1               1               1               1
           water      watertight           ways…         wayward          weaker          wealth
               1               1               1               1               1               1
            weav           weigh           wener         western       whatever…            wher
               1               1               1               1               1               1
           where       wherefrom           whiff         whisper           wholl          wholli
               1               1               1               1               1               1
         wickham          widest           width         wielder           wiggl       wikimedia
               1               1               1               1               1               1
            wild         willing         wilmott        withinss       work…cycl         workabl
               1               1               1               1               1               1
         worker…        workload         workmat       worksheet      workshops…        worldcom
               1               1               1               1               1               1
       worthwhil          wouldv           wreak          wrestl           wrigg       wrigglier
               1               1               1               1               1               1
        writecsv        writelin            wssi        xlabnumb        xsqrtsum          yammer
               1               1               1               1               1               1
       yardstick           yaser            yeah      yldotsldot            youd          youtub
               1               1               1               1               1               1
        ysqrtsum           zero…            zoom zsqrtxyzsqrtxyz
               1               1               1               1
```

Perhaps a document discussing about 'venturebeat.com' or 'worldcom'? Another document that had something to do with 'Wickham'?

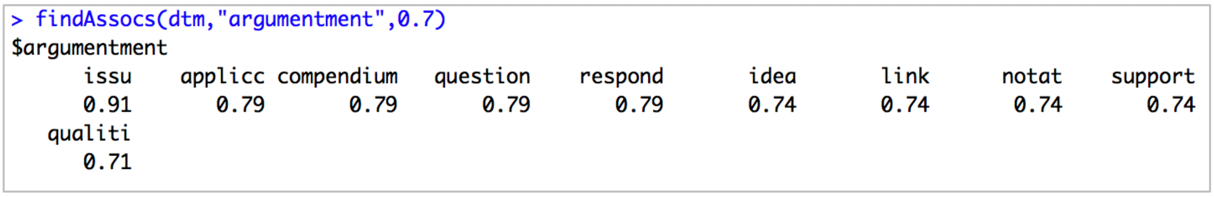
Most frequently used across documents

Taking a look at some of the words that recur a minimum of at least 80 times throughout the corpus.



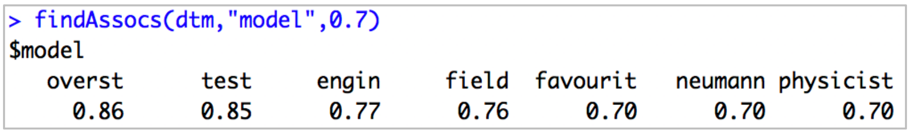
```
> findFreqTerms(dtm,lowfreq = 80)
 [1] "algorithm"    "analys"       "approach"     "argumentment" "articl"       "assum"        "author"      
 [8] "base"         "case"         "chang"        "cluster"      "complet"      "correl"       "data"        
[15] "decis"        "describ"      "design"       "develop"      "differ"       "discuss"      "distribut"   
[22] "docs"         "estim"        "exampl"       "figur"        "follow"       "function"     "group"       
[29] "idea"         "import"       "inform"       "interest"     "issu"         "knowledg"     "manag"       
[36] "mani"         "mean"         "method"       "model"        "note"         "number"       "organ"       
[43] "paper"        "possibl"      "practic"      "probabl"      "problem"      "process"      "project"     
[50] "question"     "reason"       "result"       "risk"         "simul"        "system"       "task"        
[57] "techniqu"     "term"         "time"         "topic"        "understand"   "valu"         "word"        
[64] "work"        
```

Diving deep further to understand the cooccurrence of these words with the words that are used in conjunction with these could provide better insights.



```
> findAssocs(dtm,"argumentment",0.7)
$argumentment
       issu    applicc compendium   question    respond       idea       link      notat    support
       0.91       0.79       0.79       0.79       0.79       0.74       0.74       0.74       0.74
    qualiti
       0.71
```

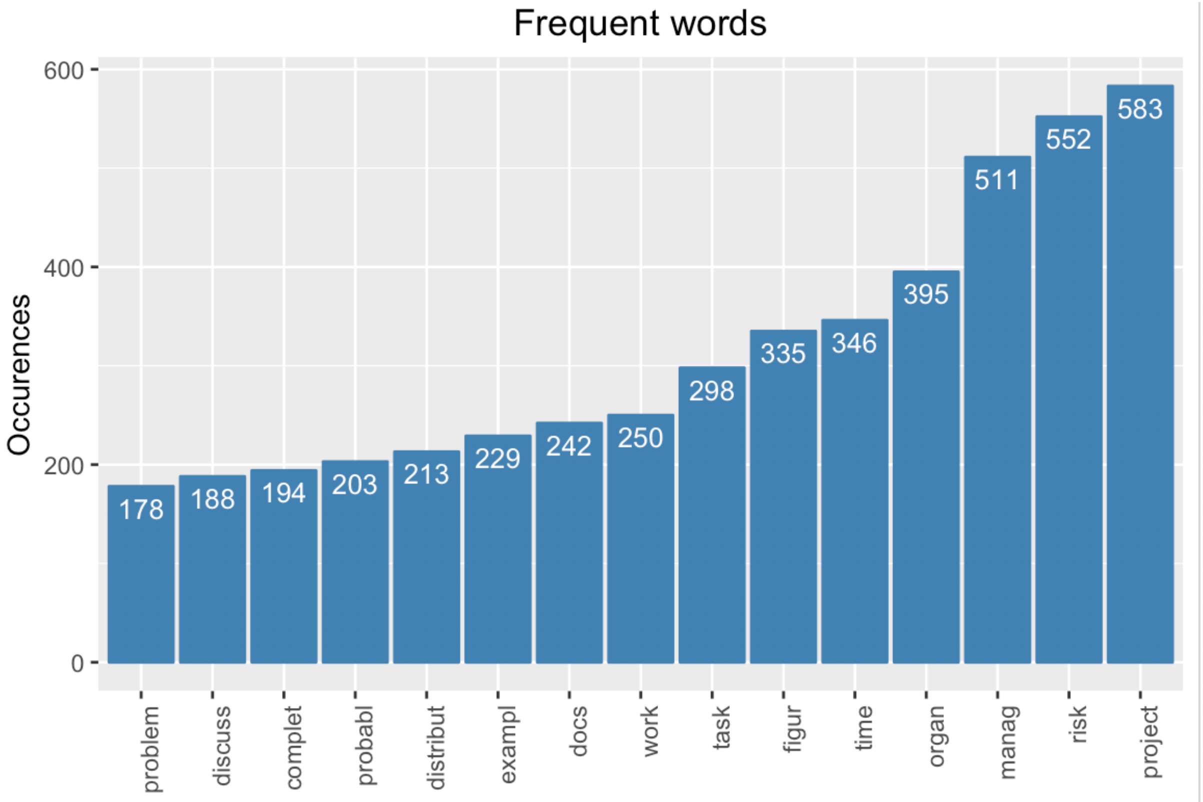
The term ‘argumentment’ seems to have a search term co-occurrence of at least 70% with words like issue, application, question, etc. which makes sense as there could have been multiple constructive discussions over these topics.



```
> findAssocs(dtm,"model",0.7)
$model
    overst      test     engin     field  favourit   neumann physicist
      0.86      0.85      0.77      0.76      0.70      0.70      0.70
```

The term ‘model’ seems to have a search term co-occurrence of at least 70% with words like test, eninge, etc. This could be a indicator that some of the documents were technical documents discussing about statistical models designed for the organisation.

A simple visualisation of the most frequently words used along with the count of times they were used would look as follows. The graph is drawn for words that occur more than 175 times throughout the corpus and each bar represent the count of times the word has been used. Most of the discussions seems to be around project and risk management.

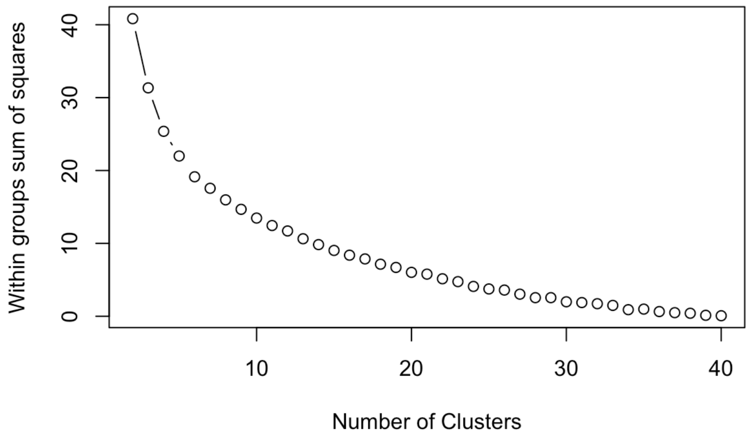


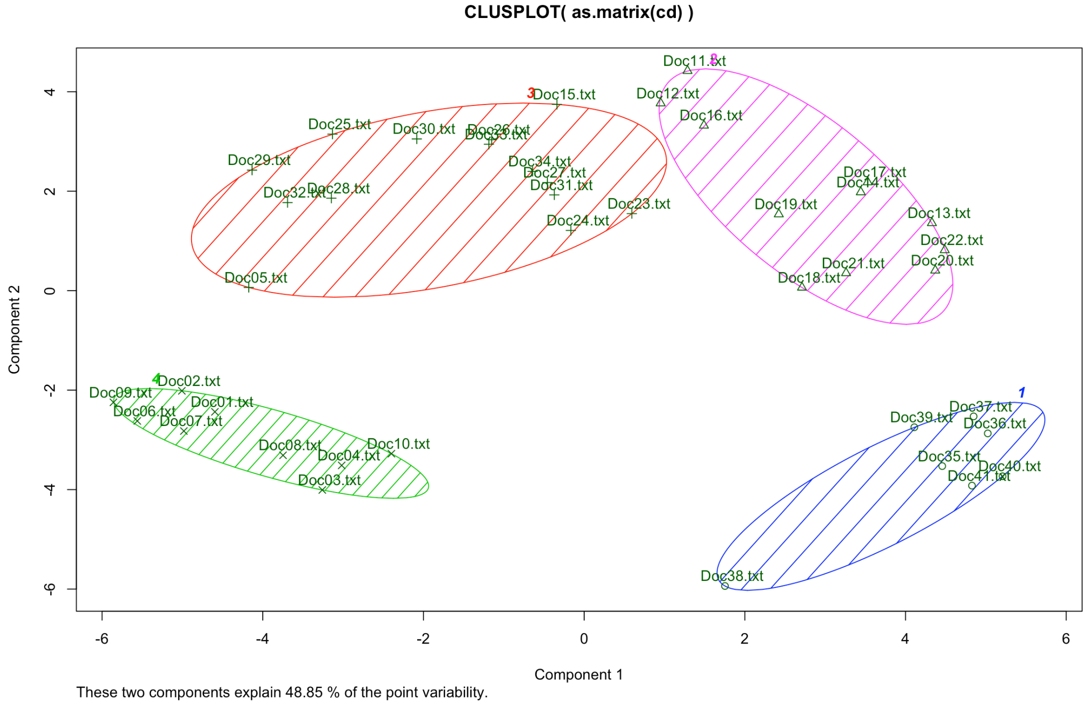
Clustering

In the following portions of the document, the corpus will be analysed with some of the more powerful tools of hierarchical clustering and k-means clustering. The choice of unit for distance measurement will be cosine distance. Cosine distance is seen as more suitable for text analysis considering the reason, size of documents cannot dictate the topic of discussion. Also, considering the varying sizes of the multiple documents in the corpus, cosine distance would be a better fit for this analysis than Euclidean distance. Furthermore, in text analysis where factors like style, sentiment, etc. are considered for grouping like documents, Euclidean distance will have the tendency to group documents based on magnitude and not direction, which could in turn lead to documents heading towards different directions on the sentiment axis could end up being grouped together based on their length. Hence, cosine distance, in this case will do a better job of clustering like documents as per our requirement for this analysis.

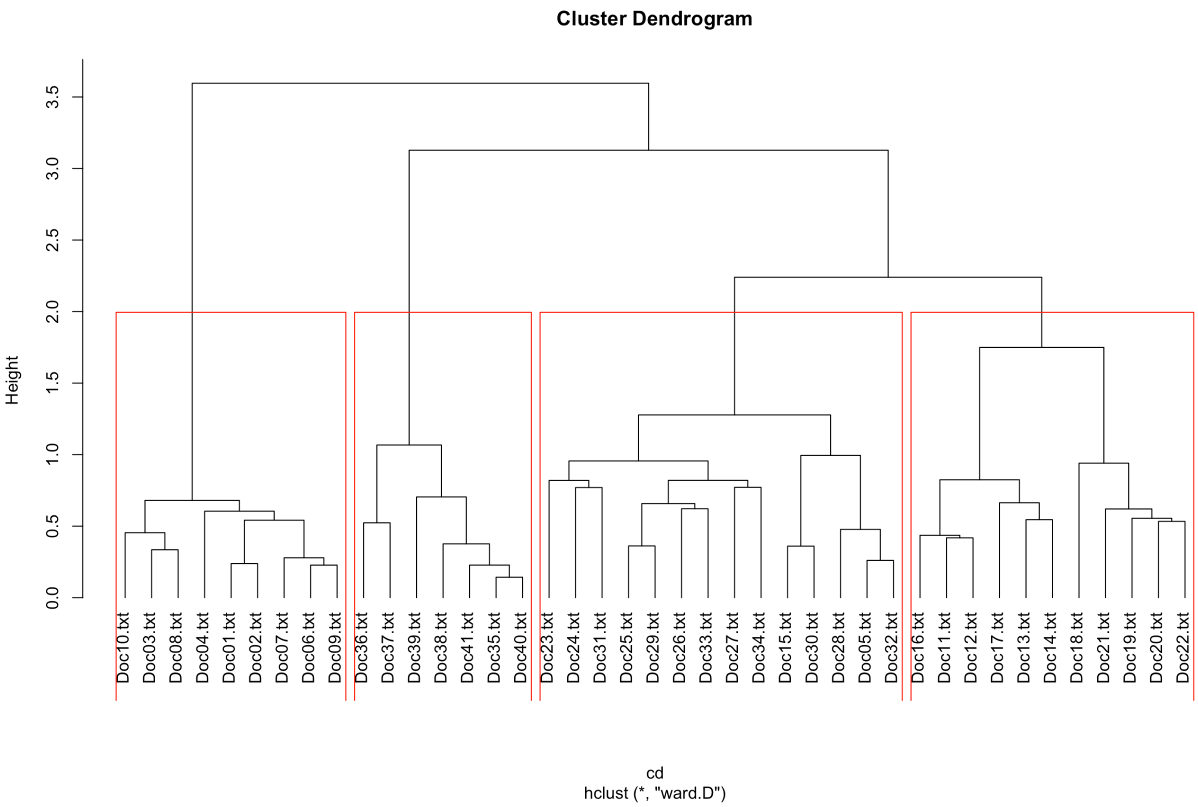
k-means clustering

Using 'elbow plot' the number of clusters for k is determined to be 4 as past this the curve seems to fairly flattens out.





Hierarchical clustering



With the clusters from k-means and hierarchical models now in hand, it can be compared and seen that the clusters formed under both measures or more or less appear to be similar, meaning there is a strong similarity between the documents under each cluster. However, the relationship between the documents in a cluster or the relationship between the clusters themselves, unfortunately, cannot be explained by either of the clustering methods. The output from the clustering technique would only serve as a starting point for further analysis, in the bigger picture providing insight that the documents under each cluster have certain similarities.

Topic modelling

To overcome the shortcoming of the previously used techniques, topic modelling of the corpus is performed in this section. Unlike the clustering techniques, the topic modelling technique helps to split the corpus into multiple topics based on the keywords. This in turn will help us understand at a wide level what the various topics covered are and how the various documents that were originally shared could be allocated to a particular topic based on the probability it holds contributed by its key words.

As in k–means clustering, topic modelling would also require the number of topics to be declared. Since, 4 was chosen for the clustering techniques based on the output of elbow plot, the same is used for topic modelling as well.

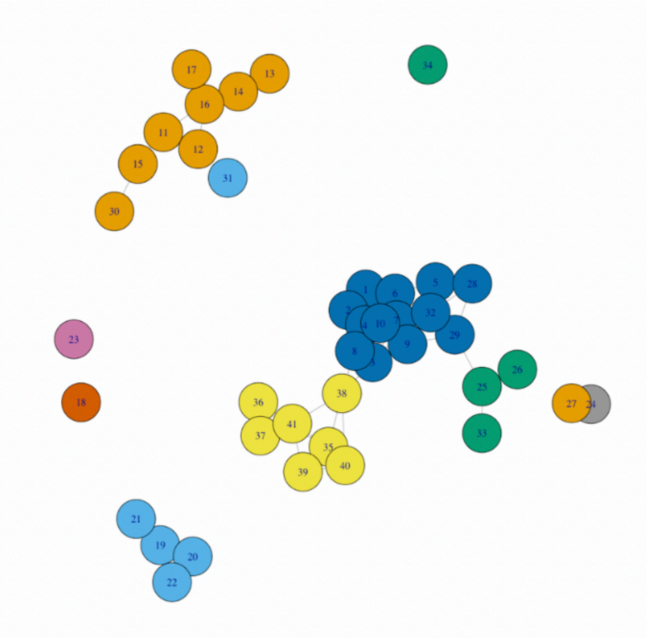
Table 1 (Topic segregation)

Topic 1	Topic 2	Topic 3	Topic 4
issu	risk	task	project
knowledg	manag	time	organ
question	analys	figur	manag
discuss	model	distribut	work
exampl	practic	docs	process
idea	problem	probabl	chang
figur	techniqu	complet	author
develop	approach	number	design

From the above table, it can be inferred that the topic 1 appears to be related to the initial stages of a project development, where knowledge of the topic seems to have been explored by brainstorming ideas, asking the right questions, figuring out the idea with examples and more. Second topic would appear to be more with risk management, analysis & modelling and discussion on techniques and approaches to solve the various business problems. The documents on the third topic could be quantifying portions of any project as we see the keywords to be around planning, figures, timelines, etc. The fourth topic could have documents that speak about the planned or on-going projects and their influence on organisation.

Network graph

Adding network graphs will give us a better visual perception of the connection between the documents that are already defied to be related using the above method of topic modelling. Fast-greedy community detection technique is used to proceed with the analysis.



The plot appears to be a close match to the topic modelling output, with the difference that the number of groups here is greater than the former.

Table 2 (Topic to document)

	Topic 1	Topic 2	Topic 3	Topic 4
Key ideas	issu	risk	task	project
	knowledg	manag	time	organ
	question	analys	figur	manag
	discuss	model	distribut	work
	exampl	practic	docs	process
	idea	problem	probabl	chang
	figur	techniqu	complet	author
	develop	approach	number	design
Associated documents	Doc11.txt	Doc01.txt	Doc19.txt	Doc05.txt
	Doc12.txt	Doc02.txt	Doc20.txt	Doc07.txt
	Doc13.txt	Doc03.txt	Doc21.txt	Doc23.txt
	Doc14.txt	Doc04.txt	Doc22.txt	Doc25.txt
	Doc15.txt	Doc06.txt	Doc35.txt	Doc26.txt
	Doc16.txt	Doc08.txt	Doc36.txt	Doc27.txt
	Doc17.txt	Doc09.txt	Doc37.txt	Doc28.txt
	Doc30.txt	Doc10.txt	Doc38.txt	Doc29.txt
		Doc18.txt	Doc39.txt	Doc31.txt
		Doc24.txt	Doc40.txt	Doc32.txt
		Doc33.txt	Doc41.txt	
		Doc34.txt		

Thus, we see that the network graphing does support the earlier identified relationship between the documents via the topic modelling technique. Drawing the understandings of the output of the two techniques, we can infer that the documents can be broadly classified into 4 types and broken down furthermore as observed in the above table.

Conclusion

Drawing conclusions by interconnecting the findings from the various techniques of clustering, topic modelling and network graphing, we see that the 41 documents part of the corpus can be widely split into four categories and further sub-divided, if required. A quick look at the way the topics are split, we can infer that the documents form a pattern from project initiation, ideas development, risk assessment, techniques and model selection to design finalisation and implementation. The final (Table 2) will provide a primary map for navigating through the documents by desired topic of interest.