# THE ANATOMY OF AN UNKNOWN CORPUS

Name: Wenying Wu
Date: 30/05/2021

# INTRODUCTION

The report details the process and analysis of insights into the content and themes of the documents in a folder in my manager's computer utilising software R's text mining tools.

# BUSINESS UNDERSTANDING

The folder named 'doc' is buried in my manager's computer. He is interested in the contents and themes of the documents in the directory but does not willing to manually go through documents and manually interpret them. Therefore, he sent the folder to me to provide insights into documents using computational text analytic methods.

# DATA UNDERSTANDING & PREPARATION

## Data Understanding

The original data is 42 text documents with a total size of 560kb in the 'doc' folder. The contents and themes of these documents are unknown unless read through manually.

## Data Preparation

The first step to performing text mining in R is to convert the documents to be analysed into a corpus (a form of a collection of documents in the R environment). Then perform the below data cleaning steps to covert the corpus into an appropriate format for text analysis:

- **Remove punctuation marks.**
  Punctuation marks do not add any value to text analysis and are insignificant.
- **Transform all letters to lowercase.**
  R is case-sensitive, performing this step ensures same words with a different case in the corpus are treated by R as one word. For example, APPLE and apple are 2 unique words for R, but they are apple and apple (1 unique word exist 2 times) after case transformation.
- **Remove numbers and digits.**
  Numbers and digits do not add any value to the analysis of words and hidden themes of documents.
- **Remove stop words.**
  Stop words are the most used words in languages, like to, is, are, in, the and so on in English. Although stop words are commonly used, they do not add any meaning to the document.
- **Remove whitespaces.**
  Whitespaces do not add any value to text analysis and are insignificant.
- **Stemming.**
  Stemming refers to trimming words to their stem by removing suffixes. Similar to case transformation, R treats words with different suffixes as different words, for example, eat and eating are 2 unique words, but they are eating and eat (1 unique word exists 2 times) after stemming.

An example of the document before and after transformation is in Appendix 1.

## EXPLORATORY DATA ANALYSIS (EDA)

The abovementioned corpus is transformed into a document tern matrix (DTM) to perform EDA and further text analysis. DTM is a matrix listing all occurrences of words in the pre-cleaned corpus, grouped by document (Awati, n.d.). Below is a simple clarification example of DTM:

Doc A: apples are red

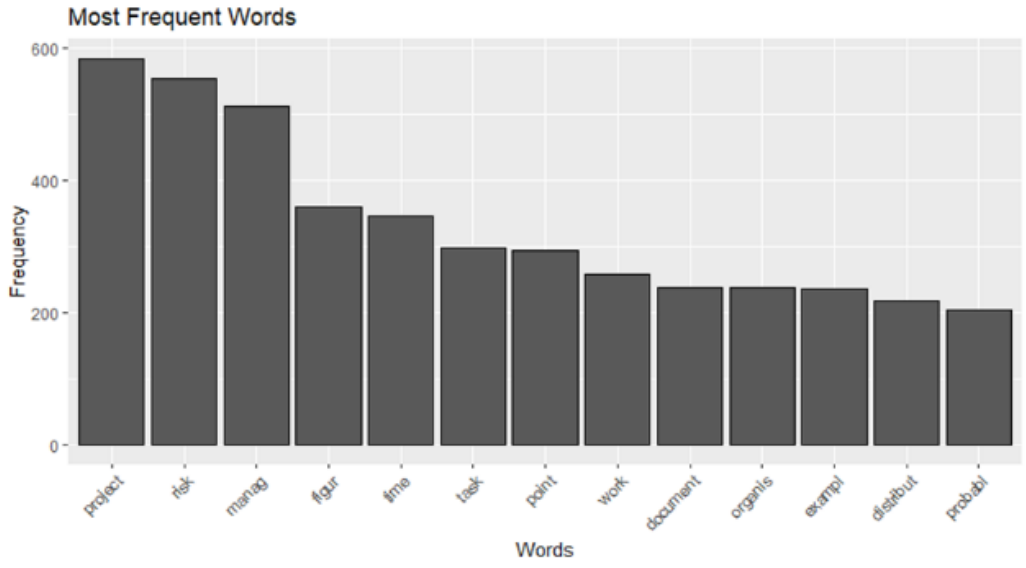
Doc B: apples are good

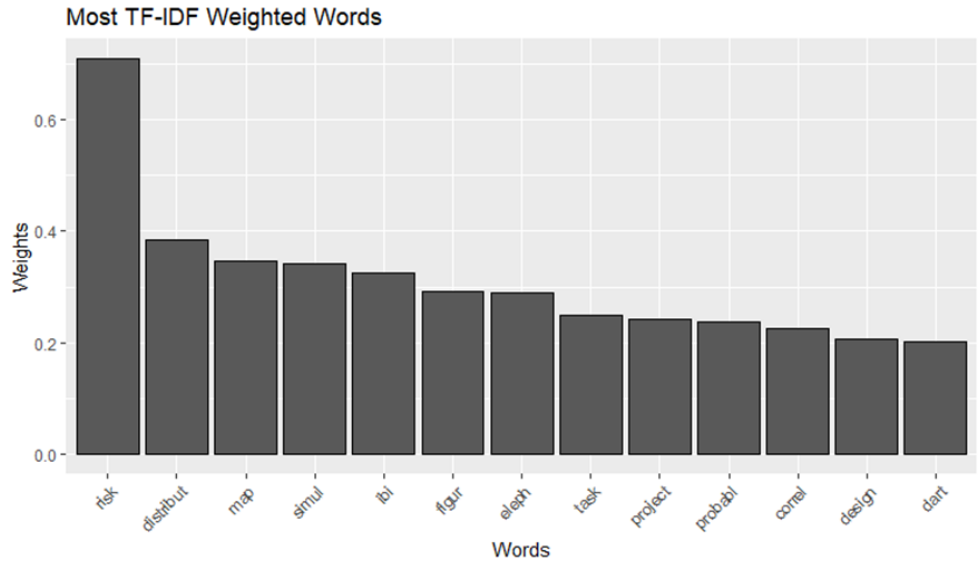
| DTM | apples | are | red | good |
|---|---|---|---|---|
| Doc A | 1 | 1 | 1 | 0 |
| Doc B | 1 | 1 | 0 | 1 |

This report utilised 2 methods for gaining insights into words in documents. The first is using word frequency and the second is using term frequency-inverse document frequency (TF-IDF). Word frequency is simply counting how many times a word appears in the corpus (42 documents), while TF-IDF is a weight that evaluates the importance of a word to a document in the corpus (42 documents). In other words, TF-IDF increases when a word in a document is high but is scaled down when the occurrence of a word in the corpus (42 documents) is low (*Tf-Idf : A Single-Page Tutorial - Information Retrieval and Text Mining*, n.d.).

Below are the major findings from EDA, miscellaneous results are in Appendix 2.

There are 4246 unique words in the corpus after data cleaning, below graphs show the most frequent words from word frequency and the most TF-IDF weighted words in the 42 documents.






Interestingly, these two graphs provide a different view of the documents. From word frequency, the three most frequent words are ‘project’, ‘risk’ and ‘manag’, while the TF-IDF suggests that the three most weighted words are ‘risk’, ‘distribut’ and ‘map’.

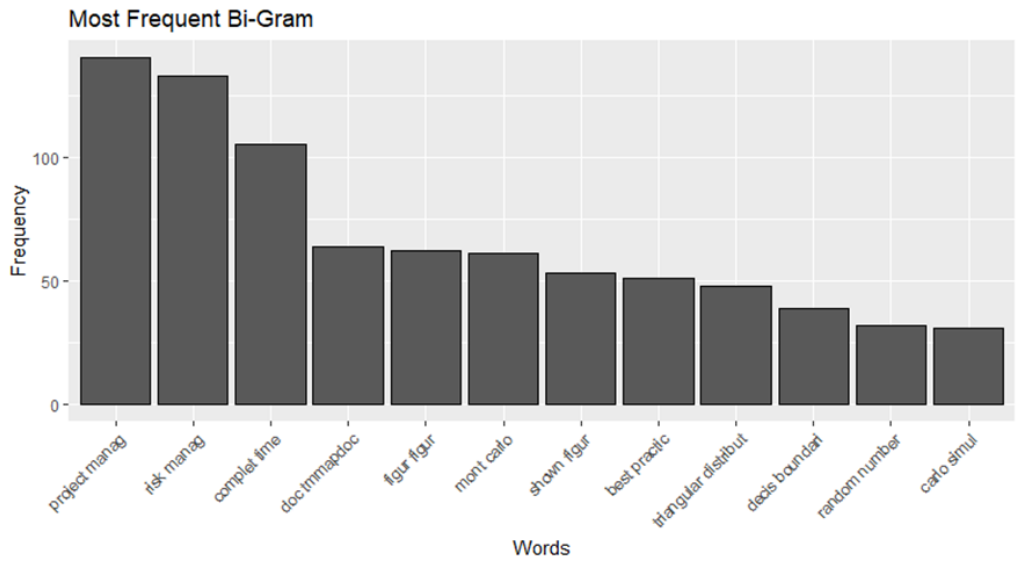
Another representation of the most frequent words and the most weighted words is the word cloud shown below. Bigger word size means more frequency or weight.


**Frequency word cloud**


**TF-IDF word cloud**

Though there are differences in these two methods, the common words like ‘project’, ‘risk’, ‘organis’ and ‘document’ suggest that the 42 documents are likely to be related to an organisation’s project risk management documentation.

The below Most frequent Bigram (2 words combination occurrence) suggest the three most frequent two-words combination are ‘project manag’, ‘risk manag’ and ‘complete time’. This also suggests the 42 documents are likely to be related to project management and risk management.

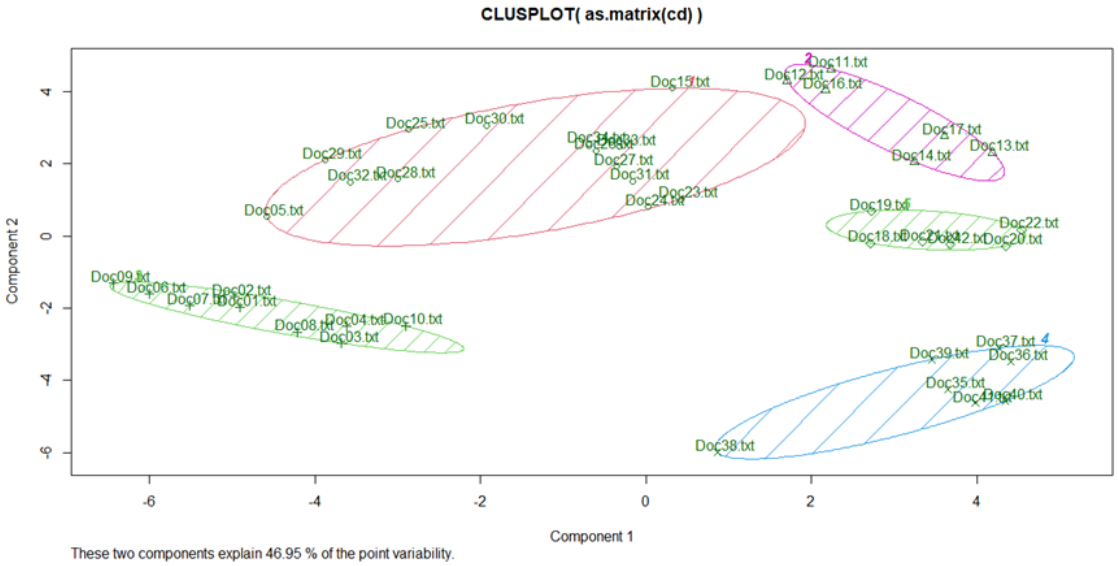



# CLUSTERING

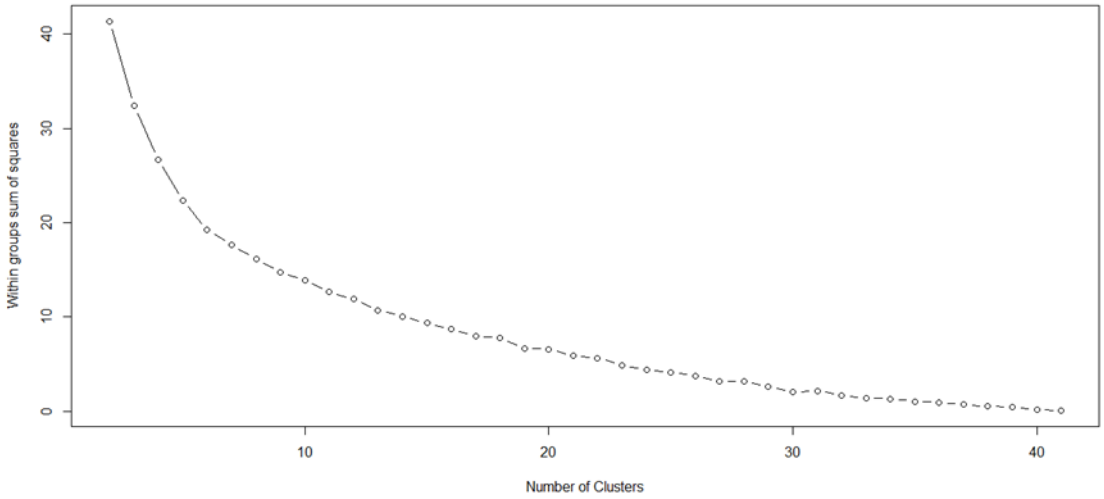
Clustering means create clusters (groups) that contain similar documents. Documents in one cluster are similar to each other but different from documents in other clusters. This section shows K-mean clustering because it is more practical to find the optimal cluster number. Furthermore, the similarity between documents is computed by cosine distance in this report. Cosine distance and Euclidean distance are two common document similarity measurements, however, given cosine distance takes document size (word count difference) into account, it is more advantageous than Euclidean distance (Prabhakaran, n.d.).

## K-mean Clustering

Below CLUSPLOT is the result from K-mean clustering for K=5. (K represents the cluster number) Each cluster is shown in a different colour in CLUSPLOT, doc numbers fall in the same circle means they are clustered in the same group. For example, Doc 11,12,13,14,16 and 17 are in the same cluster.



Below Elbow Plot for K-mean Clustering illustrates why K=5 is chosen. The optimal cluster number is the point where the within-group sum of squares (Y-axis in Elbow plot) decreasing rate slows down. Hence, K=5 is the optimal cluster number for the K-mean clustering method.

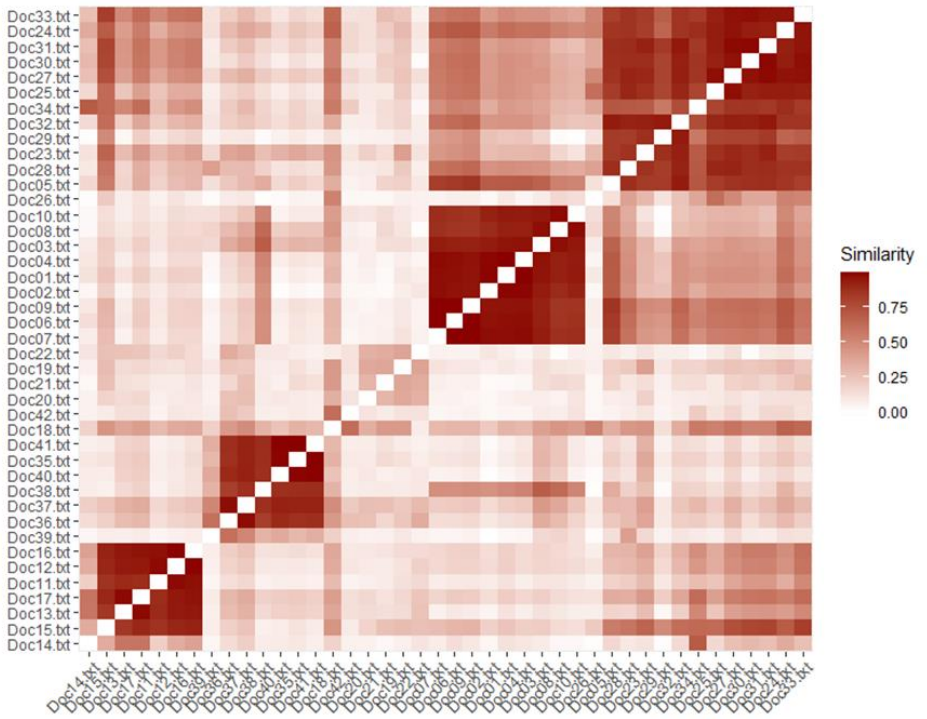



Hierarchical clustering is an alternative clustering method and Network graph is another visualisation tool in text analytic, both detailed in Appendix 3.

## LATENT SEMANTIC ANALYSIS

Latent semantic analysis refers to an analysis of the meaning in the corpus. It can be used to compare the similarity between documents in the sense of semantic meaning and find the nearest neighbour of a specific word (what words occur when a specific word appears).

The semantic similarity was plotted below, darker red colour refers to stronger semantic similarity. Looking at the previous cluster example (cluster containing Doc 11,12,13,14,16 and 17) at the bottom left corner, it suggests that Doc 11, 12,13,14,16 and 17 does have a strong semantic similarity. However, Doc 15 also has strong semantic similarity with these docs but not included in the K-mean or Hierarchical clustering while included in one cluster in Fast Greedy and Louvain methods in the Network graph. This discrepancy suggests we must dive deeper into text analytics to figure out which cluster method is more appropriate in this scenario.




Analysis of the nearest neighbour of some most frequent words analysis is in Appendix 3.

## TOPIC MODELLING

By far we have only the optimal cluster number from K-mean clustering but need to figure out what the topics are about. Topic modelling utilising Latent Dirichlet Allocation (LDA) algorithm was performed for both K=5 and K=6 to figure out what topics are these clusters presenting and to re-examine the optimal cluster number. Below are the topics modelling results for both K=5 and K=6, showing the 8 most frequent words for each topic.

| | Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 |
|---|---|---|---|---|---|
| [1,] | "organis" | "map" | "document" | "task" | "risk" |
| [2,] | "work" | "issu" | "word" | "time" | "project" |
| [3,] | "manag" | "knowledg" | "cluster" | "distribut" | "manag" |
| [4,] | "practic" | "ibi" | "topic" | "probabl" | "process" |
| [5,] | "project" | "point" | "algorithm" | "complet" | "model" |
| [6,] | "chang" | "discuss" | "doc" | "figur" | "problem" |
| [7,] | "organ" | "question" | "figur" | "simul" | "becaus" |
| [8,] | "best" | "idea" | "data" | "number" | "techniqu" |

K=5

**Interpretation of output:**

Topic 1: Project management and work organisation

Topic 2: Mapping problems, discussion on issues/ points/ ideas

Topic 3: Text and data analytics involving clustering, topic, document and algorithm

Topic 4: Task analysis including timing, distribution and completion

Topic 5: Project and risk management

| | Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 | Topic 6 |
|---|---|---|---|---|---|---|
| [1,] | "organis" | "risk" | "map" | "document" | "task" | "model" |
| [2,] | "work" | "project" | "issu" | "word" | "time" | "data" |
| [3,] | "practic" | "manag" | "ibi" | "cluster" | "distribut" | "techniqu" |
| [4,] | "manag" | "process" | "knowledg" | "figur" | "probabl" | "exampl" |
| [5,] | "chang" | "author" | "question" | "topic" | "complet" | "mani" |
| [6,] | "organ" | "social" | "idea" | "doc" | "figur" | "import" |
| [7,] | "design" | "approach" | "discuss" | "algorithm" | "simul" | "method" |
| [8,] | "best" | "problem" | "point" | "plot" | "number" | "variabl" |

K=6

Interpretation of output:

Topic 1: Work organisation and management

Topic 2: Project and risk management

Topic 3: Mapping problems, discussion on issues/ points/ ideas

Topic 4: Text analytics involving clustering, topic, document and algorithm

Topic 5: Task analysis including timing, distribution and completion

Topic 6: Data modelling techniques and methods

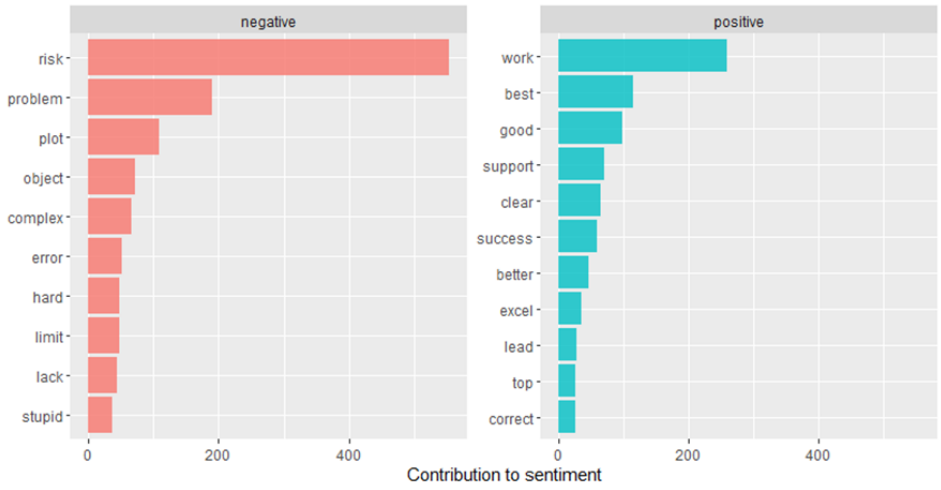
Comparing these two topic modelling results, I found that Topic 4 (K=6) is similar to Topic 6 (K=6), both are about data analytics and fall in Topic 3 (K=5). All other topics are very similar. Therefore, the optimal topic number should be 5 and the optimal cluster number should be 5.

Documents grouped by topic (LDA) are listed below, we can see that Topic 2 is the cluster mentioned in the previous example (Doc 11-17).

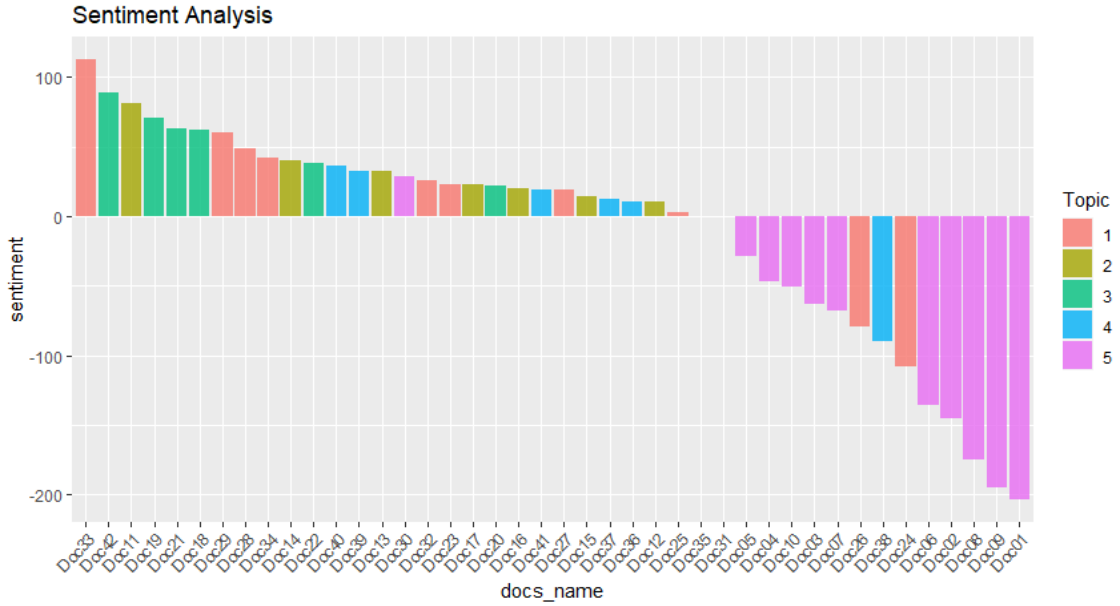
| Topic | Docs |
|---|---|
| <int> | <chr> |
| 1 | Doc23 Doc24 Doc25 Doc26 Doc27 Doc28 Doc29 Doc31 Doc32 Doc33 Doc34 |
| 2 | Doc11 Doc12 Doc13 Doc14 Doc15 Doc16 Doc17 |
| 3 | Doc18 Doc19 Doc20 Doc21 Doc22 Doc42 |
| 4 | Doc35 Doc36 Doc37 Doc38 Doc39 Doc40 Doc41 |
| 5 | Doc01 Doc02 Doc03 Doc04 Doc05 Doc06 Doc07 Doc08 Doc09 Doc10 Doc30 |

## SENTIMENT ANALYSIS

Sentiment analysis provides an overview of whether a text document carries a positive or negative sentiment. Below are the top words contributing to the sentiment analysis in this scenario. Risk contributes the most to negative sentiment, and work contributes the most to positive.




Below is the sentiment analysis graph for the 42 documents. Coloured by Topic so we can have an overall idea of each topic's sentiment. We can see overall sentiment for Topic 5 is negative, this is because this topic is about project and risk management. The document in Topic 5 with lower sentiment might be interpreted as a positive outcome in this scenario. Other topics are generally positive except for Doc 24, 26 in Topic 1 and Doc 38 in Topic 4.



# CONCLUSION

In conclusion, the 42 documents in ‘docs’ include 5 topics with different sentiments summarised in the below table.

| Topic | Name | Overall Sentiment |
|---|---|---|
| Topic 1 | Project management and work organisation | Generally Positive |
| | Doc 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34 | Doc 24&26 Negative |
| Topic 2 | Mapping problems, discussion on issues and ideas | Generally Positive |
| | Doc 11, 12, 13, 14, 15, 16, 17 | |
| Topic 3 | Text and data analytics | Generally Positive |
| | Doc 18, 19, 20, 21, 22, 42 | |
| Topic 4 | Task analysis | Generally Positive |
| | Doc 35, 36, 37, 38, 39, 40, 41 | Doc 38 Negative |
| Topic 5 | Project and risk management | Generally Negative |
| | Doc 01, 02, 03, 04, 05, 06, 07, 08, 09, 10, 30 | |

# REFERENCE


Awati, K. (n.d.). *A gentle introduction to text mining using R | Eight to Late*. Retrieved May 30, 2021, from https://eight2late.wordpress.com/2015/05/27/a-gentle-introduction-to-text-mining-using-r/

Medelyan, A. (n.d.). *5 Text Analytics Approaches: A Comprehensive Review*. Retrieved May 30, 2021, from https://getthematic.com/insights/5-text-analytics-approaches/

Prabhakaran, S. (n.d.). *Cosine Similarity - Understanding the math and how it works? (with python)*. Retrieved May 30, 2021, from https://www.machinelearningplus.com/nlp/cosine-similarity/

*Tf-idf : A Single-Page Tutorial - Information Retrieval and Text Mining*. (n.d.). Retrieved May 30, 2021, from http://www.tfidf.com/

# APPENDIX

## Appendix 1

### Before Cleaning

[1] "Project managers know from experience that projects can go wrong because of events that we
ren't foreseen. Some of these may be unforeseeable that is, they could not have been anticipate
d given what was known prior to their occurrence. On the other hand  it is surprisingly common
 that known risks are ignored.  The metaphor of the elephant in the room is appropriate here be
cause these risks are quite obvious to outsiders, but apparently not to those involved in the p
roject. This is a strange state of affairs because:  Those involved in the project are best pla
ced to \"see the elephant\" They are directly affected  when the elephant goes on rampage   i.
e. the risk eventuates. This post discusses reasons why these metaphorical pachyderms are ignor
ed by those who need most to recognize their existence .  Let's get right into it then seven re
asons why risks are ignored on projects:  1. Let sleeping elephants lie: This is a situation in
 which stakeholders are aware of the risk, but don't do anything about it in the hope that it w
ill not eventuate. Consequently, they have no idea how to handle it if it does. Unfortunately,
 as Murphy assures us, sleeping elephants will wake at the most inconvenient moment.  2. It's n
ot my elephant: This is a situation where no one is willing to take responsibility for managing
 the risk. This game of \"pass the elephant\" is resolved by handing charge of the elephant to
 a reluctant mahout.  3. Deny the elephant's existence: This often manifests itself as a case o
f collective (and wilful) blindness to obvious risks.  No one acknowledges the risk, perhaps ou
t of fear of that they will be handed responsibility for it (see point 2 above).  4. The elepha
nt has powerful friends: This is a pathological situation where some stakeholders (often those
 with clout) actually increase the likelihood of a risk through bad decisions. A common example
 of this is the imposition of arbitrary deadlines, based on fantasy rather than fact.  5. The e
lephant might get up and walk away: This is wishful thinking, where the team assumes that the r
isk will magically disappear. This is the \"hope and pray\" method of risk management, quite co
mmon in some circles.  6. The elephant's not an elephant: This is a situation where a risk is m
istaken for an opportunity. Yes, this does happen. An example is when a new technology is used
 on a project: some team members may see it as an opportunity, but in reality it may pose a ris
k.  7. The elephant's dead: This is exemplified by the response, \"that is no longer a proble
m,\" when asked about the status of a risk.  The danger in these situations is that the elephan
t may only be fast asleep, not dead.  Risks that are ignored are the metaphorical pachyderms in
 the room.  Ignoring them is easy  because it involves no effort whatsoever. However, it is a s
trategy that is fraught with danger because once these risks eventuate, they can like those app
arently invisible elephants run amok and wreak havoc on projects."

### After Cleaning

project manag know experi project go wrong becaus event werent foreseen unforese anticip given
 known prior occurr hand surpris common known risk ignor metaphor eleph room appropri becaus ri
sk obvious outsid appar involv project strang state affair becaus involv project best place ele
ph direct affect eleph goe rampag ie risk eventu discuss reason whi metaphor pachyderm ignor re
cogn exist let seven reason whi risk ignor project let sleep eleph lie situat stakehold risk do
nt anyth hope eventu consequ idea handl doe unfortun murphi assur us sleep eleph wake inconveni
 moment eleph situat respons manag risk game pass eleph resolv hand charg eleph reluct mahout d
eni eleph exist manifest case collect wil blind obvious risk acknowledg risk perhap fear hand r
espons point abov eleph power friend patholog situat stakehold clout actual increas likelihood
 risk bad decis common exampl imposit arbitrari deadlin base fantasi fact eleph walk wish team
 assum risk magic disappear hope pray method risk manag common circl eleph eleph situat risk mi
staken opportun yes doe happen exampl technolog project team member opportun realiti pose risk
 eleph dead exemplifi respons longer problem status risk danger situat eleph onli fast asleep d
ead risk ignor metaphor pachyderm room ignor easi becaus involv effort whatsoev strategi fraugh
t danger becaus onc risk eventu appar invis eleph run amok wreak havoc project

# Appendix 2

## Part 1 DTM

### a. Word frequency

```
> ## Summary of the Document Term Matrix
> dtm
<<DocumentTermMatrix (documents: 42, terms: 4246)>>
Non-/sparse entries: 18006/160326
Sparsity           : 90%
Maximal term length: 65
Weighting          : term frequency (tf)
```

### b. TF-IDF

```
> ## Summary of the Document Term Matrix
> dtm_tfidf
<<DocumentTermMatrix (documents: 42, terms: 4246)>>
Non-/sparse entries: 18006/160326
Sparsity           : 90%
Maximal term length: 65
Weighting          : term frequency - inverse document frequency (normalized) (tf-idf)
>
> ## Collapse matrix by summing over columns - this gets total counts (over all docs)
 for each term
> wt_tot_tfidf <- colSums(as.matrix(dtm_tfidf))
>
> ## Number of terms
> length(wt_tot_tfidf)
[1] 4246
```

## Part 2 Frequency count

### a. Word frequency

```
> freq[head(ord)]
project    risk   manag   figur    time    task
    585     555     512     360     347     299
> ## List most frequent terms. Lower bound specified as second argument
> findFreqTerms(dtm,lowfreq=100)
 [1] "abov"       "ani"        "approach"   "articl"     "base"
 [6] "becaus"     "best"       "case"       "chang"      "cluster"
[11] "complet"    "correl"     "data"       "decis"      "develop"
[16] "differ"     "discuss"    "distribut"  "document"   "exampl"
[21] "figur"      "function"   "ibi"        "idea"       "import"
[26] "issu"       "knowledg"   "manag"      "mani"       "map"
[31] "model"      "note"       "number"     "organis"    "paper"
[36] "plot"       "point"      "practic"    "probabl"    "problem"
[41] "process"    "project"    "question"   "result"     "risk"
[46] "simul"      "system"     "task"       "techniqu"   "term"
[51] "time"       "topic"      "understand" "valu"       "word"
[56] "work"
```

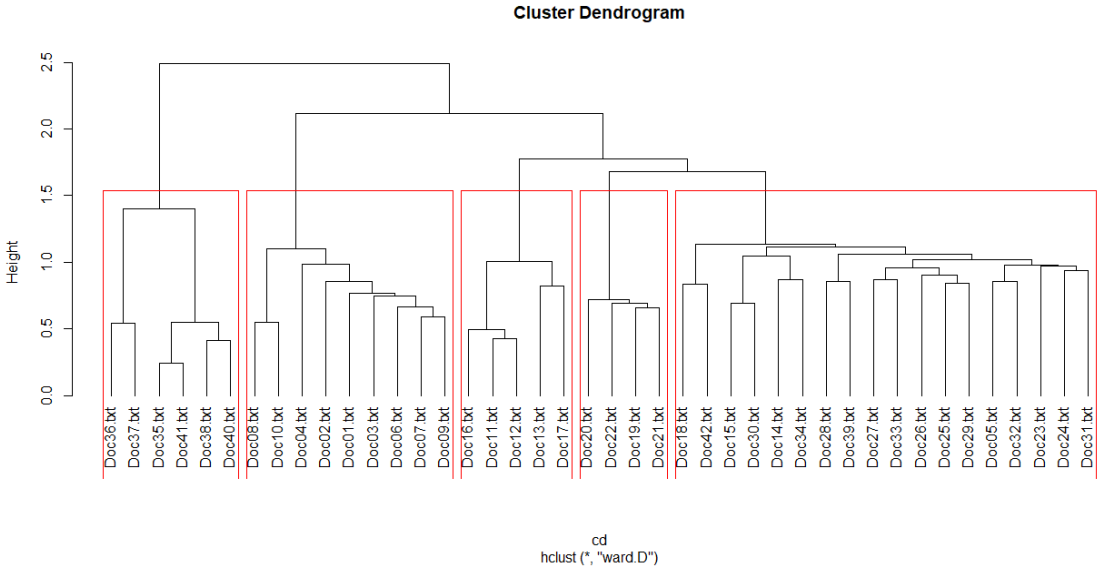
**b. TF-IDF weight**

```
> ## Inspect most frequently occurring terms
> wt_tot_tfidf[head(ord_tfidf)]
     risk distribut       map     simul       ibi     figur
0.7083808 0.3842948 0.3476514 0.3427791 0.3243720 0.2923322
> ## Inspect least frequently occurring terms
> wt_tot_tfidf[tail(ord_tfidf)]
                                          wickham
                                       0.002026425
                            wordcloudnamesfreqrfreqr
                                       0.002026425
wordcloudnamesfreqrfreqrminfreqcolorsbrewerpaldark
                                       0.002026425
                                          writelin
                                       0.002026425
                                            youtub
                                       0.002026425
                                               zip
                                       0.002026425
```

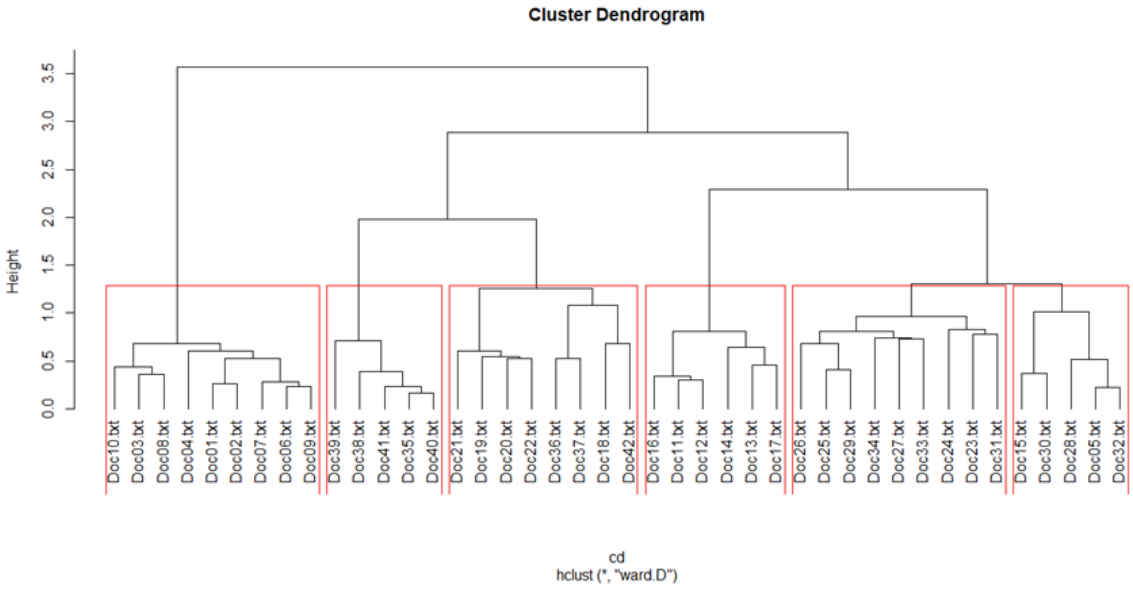
# Appendix 3

## Hierarchical clustering

Given the optimal cluster number is 5 in the K-mean clustering method. Cluster number equal 5 and 6 are trailed in Hierarchical clustering. The below graphs shows the result, each red rectangle represents one cluster, and documents inside one rectangle are in one cluster.




**Cluster NO. = 5**

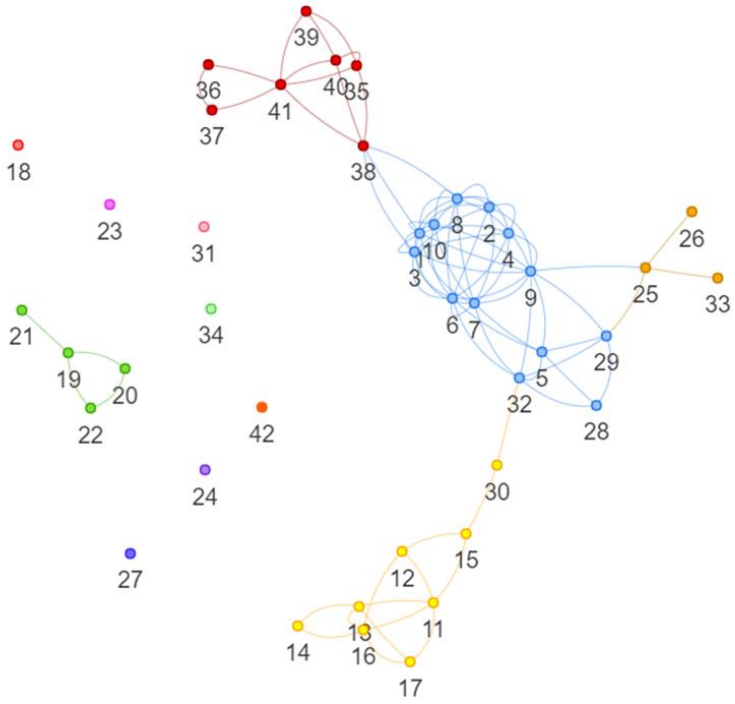



**Cluster NO. = 6**

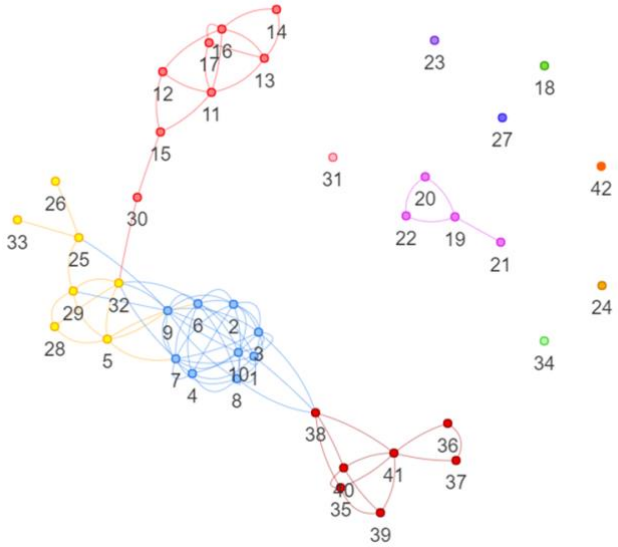
By looking at the example cluster mentioned in K-mean section, both cluster number = 5 and 6 Hierarchical clustering output the same cluster (Doc 11,12,13,14,16 and 17 are in the same cluster.

**Network Graph**

Network graph is another visualisation tool for text analytics, this section shows network graphs clustered documents by Fast Greedy and Louvain methods. The below graphs shows a network between 42 documents and these two methods yield very similar cluster. However, they indicate different clusters to K-mean clustering and Hierarchical clustering, which is normal in text analytic as text analytic is a subjective task. Especially in clustering, different method yields different results (Medelyan, n.d.).




**Network - Fast Greedy cluster**




**Network - Louvain methods cluster**

## Appendix 4

**Nearest Neighbor Analysis**

```
> neighbors("project",n=6,tvectors=LSAtk)
  project      mega      pose    troubl     carri      cycl
1.0000000 0.9792347 0.9755198 0.9733967 0.9723224 0.9694775
> neighbors("risk",n=6,tvectors=LSAtk)
     risk     dwell    danger     mitig      amok     clout
1.0000000 0.9928132 0.9909437 0.9900238 0.9896045 0.9896045
> neighbors("manag",n=6,tvectors=LSAtk)
    manag  supervis    incent  institut    consum    visibl
1.0000000 0.9904038 0.9839738 0.9829739 0.9807297 0.9795749
> neighbors("figur",n=6,tvectors=LSAtk)
    figur     shown      show     three    radius      dart
1.0000000 0.9102322 0.8994751 0.8919616 0.8832720 0.8829940
> neighbors("time",n=6,tvectors=LSAtk)
     time   complet    finish     chanc     cumul      task
1.0000000 0.9880242 0.9863386 0.9847497 0.9772343 0.9771926
> neighbors("task",n=6,tvectors=LSAtk)
     task     cumul     simul   complet    finish     delay
1.0000000 0.9965929 0.9948729 0.9930608 0.9925140 0.9910439
```

1. project: When project exists, the most likely existing words are mega, pose, trouble, carri and cycl. It could be an indicator these documents are likely to involving mega projects, considering troubles they could face, what are projects carrying and considering the project cycle.
2. risk: Risk mitigation could be a theme.
3. manag: Management of institution, the consumer could be a theme.
4. figure: Indicates some documents are likely to relate to the analysis of something.
5. time: Some documents are likely to relate to complete time, cumulative time problem.
6. task: Some documents are likely to relate to task discussion, maybe about the delay, finish or completeness.