# CORPUS ANALYSIS

## BUSINESS UNDERSTANDING

**Objectives**:

1. Provide insights into the themes and content of 42 unnamed documents.
2. Re-organise the main "docs" folder into subfolders by theme and rename each document to help identify its content.

## DATA UNDERSTANDING

1. Glossary

| Term | Definition |
|---|---|
| Bag of words model | Way to extract features from document to then be used for model |
| Collocation | Series of words that co-occur more than would be expected by chance |
| Corpus | Ensemble of documents |
| Cosine similarity | Measure of similarity |
| Document-term matrix | Table with the below:<br>Rows = documents<br>Columns = terms<br>Cell = frequency of words (count) |
| IDF | Inverse document frequency – measures rarity of a term in a document |
| Stemming | Process of reducing word to it's root form |
| Stopwords | commonly used in the English language ("a", "an", "the") |
| Topic modelling | Statistical model to identify topics |
| Wordcloud | Graphical representation of most frequent words |

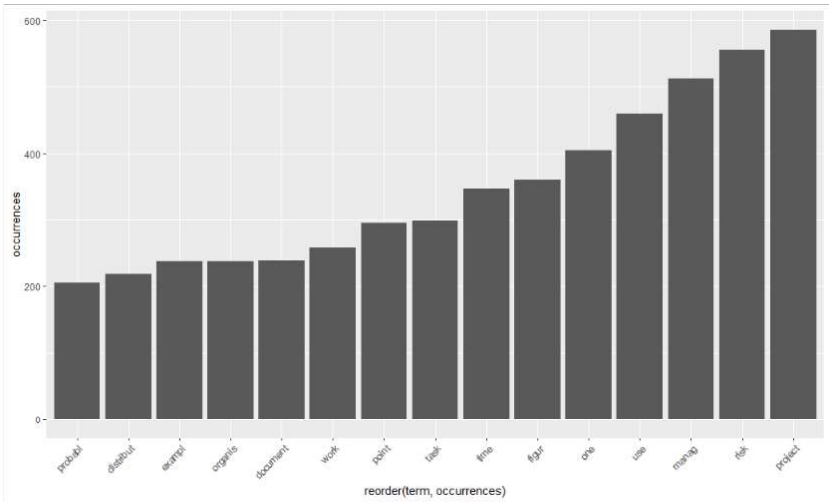
2. Assumptions

It is assumed that the documents' content is different and can be grouped into multiple topics based on the words mentioned in each document.

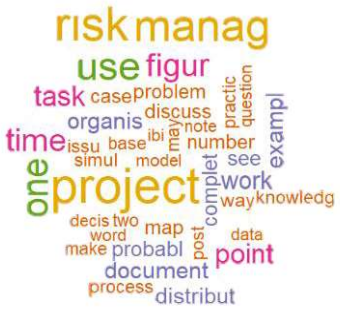
3. Data exploration

The corpus contains 42 documents of type text (".txt")

Histogram of words that occur more than 200 times in the corpus




Same representation but as a Wordcloud: risk, manag and project are the top 3 words.




## DATA PREPROCESSING

To facilitate the analysis and create useful visualisations the data needs to be processed.
Here are the transformations made on the corpus of documents:

1. Remove punctuation
2. Transform all words to lower case
3. Remove “stopwords” – to keep meaningful words only, remove “noise” and improve the accuracy of further classification.
   In addition to the common English stopwords “can”, “will”, “use” and “one” have also been removed
4. Remove numbers
5. Remove whitespace
6. Stemming to standardise the words (“conventional” to “convent”)

This allows to put the remaining words into the matrix needed for the model (Extract: Appendix 1) .

# MODELLING AND METHODS USED

To satisfy the business requirement of identifying the corpus' themes and the individual content of each document, the model used for this task is a "topic modelling" algorithm.

1. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) requires a "bag of words" in a matrix format as an input which has been created in the data pre-processing step.

LDA also requires a sampling method, in this case we are using "Gibbs", it is the most used algorithm for the LDA model.

The parameters have been adjusted multiple times to find the optimum number of topics of 5.

**Top 10 keywords for each topic (words are stemmed).**

| | Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 |
|---|---|---|---|---|---|
| 1 | task | map | organis | document | risk |
| 2 | time | issu | work | word | project |
| 3 | distribut | knowledg | manag | cluster | manag |
| 4 | probabl | ibi | project | topic | problem |
| 5 | complet | question | practic | figur | use |
| 6 | figur | discuss | chang | point | may |
| 7 | simul | idea | best | use | author |
| 8 | number | point | organ | term | model |
| 9 | correl | said | process | doc | base |
| 10 | use | argument | system | data | process |

Refer to appendix number 2 - **LDA MODEL + GIBBS SAMPLING OUTPUT** to view the topic associated to each document.

Using these topics single words and by looking at some of the most frequent associations of words in a document the 5 below topics have been identified:

**Topic 1:** "Probability Distribution Completion Time" (7 documents)
**Topic 2:** "Issue mapping" (7 Documents)
**Topic 3:** "Project Management Best Practice" (10 Documents)
**Topic 4:** "Machine learning - Text mining" (6 Documents)
**Topic 5:** "Risk Management" (12 Documents)

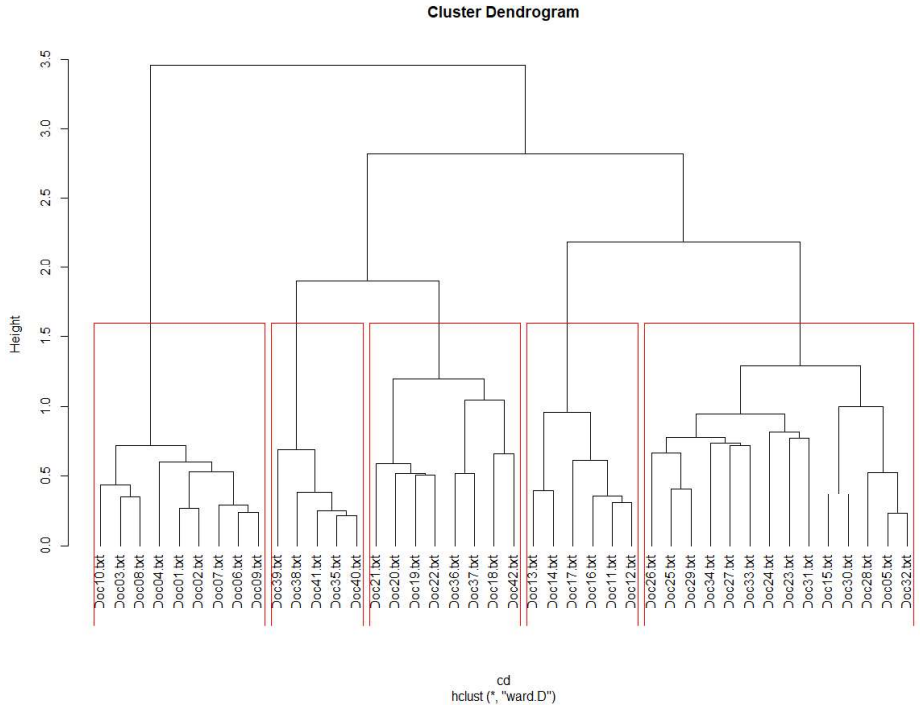
Example of the frequent associations for topic3 that helped identifying the topic of "Project Management Best Practice"

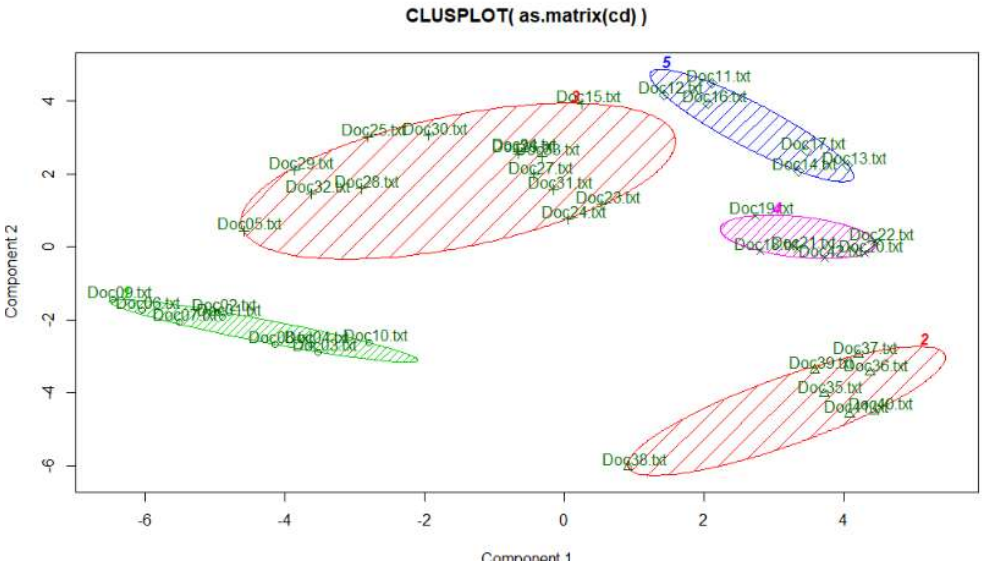
| Collocation | Count |
|---|---|
| Project Management | 53 |
| Best Practice | 44 |

2. Clustering by measuring distance between documents using Cosine similarity

The cosine similarity algorithm is used to identify how “close” documents are to each other and groups them into clusters.

The result of the cosine similarity can be plotted as a dendrogram and as a cluster plot which shows the different clusters highlighted in red below and the documents each cluster contains.




Here again, it was identified that 5 was an optimum count of clusters as the branches and the clusters are of similar height and is balanced.

The cosine method and LDA model almost provide a similar result except for the classification of 5 out of 42 documents: doc_5, doc_15, doc_30, doc_34, doc_36, doc_37.

The output using cosine similarity allows us to confirm that these documents are considered similar but does not give keywords or topic.

The selected topic attribution to a document is, as per decided by the LDA model. (APPENDIX 2)

## FINDINGS – CORPUS AND INDIVIDUAL DOCUMENT

<u>Limitations of topic modelling and further improvements:</u>
Through deep learning, the model could be fed a series of blog articles, emails, meeting notes, paper summary etc to be trained on and it could then identify the writing style and type of the document.

It is easier to group documents when there is an expectation of how many topics we need to identify (example, positive or negative reviews).

The model used here isn't capable of analysing a document type and tell whether it is a blog article


or an email but can identify a set of different topics within the documents and how identical they are as well as pulling the main keywords of an individual document.

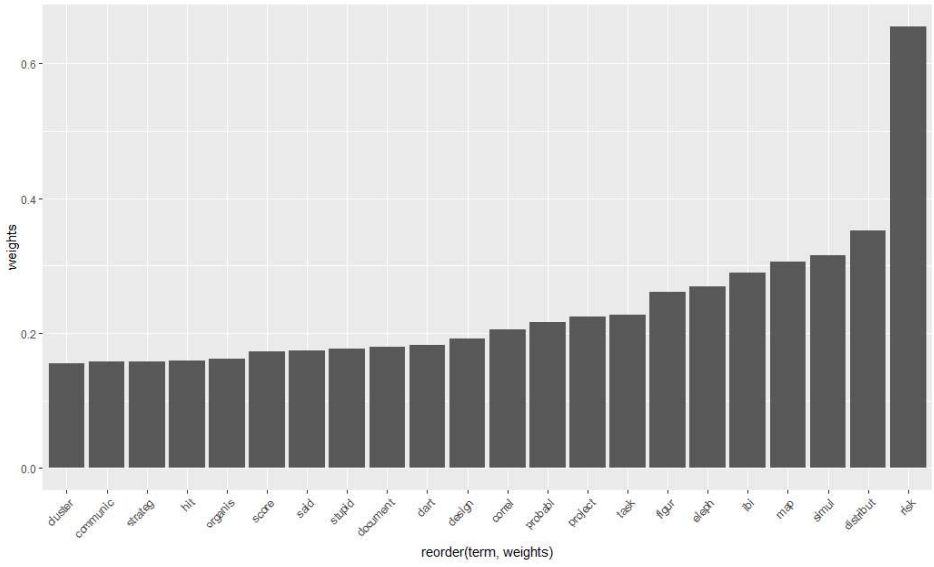
The findings of the analysis on the whole corpus come from the LDA model applied to all the documents. This allowed to identify 5 topics.

To have a better understanding of the content within each topic, term-frequency can be used as a method to highlight important words by answering the question "How frequent is this word in the corpus?"

It is also possible to add a weight to these words to identify how important this word is compared to the whole corpus by using term's inverse document frequency (idf).
Commonly used words will have a lower weight than "rarer" words which will have a higher weight.

Below is a new word cloud that combined term-frequency (TF) and inverse document frequency (idf) and the top words have slightly changed, risk and distribut are now the top 2 stemmed words, followed by the ones in purple.
**This is the word cloud for the whole corpus.**

**Histogram of words after applying weight**




## FINAL RESULTS

As a result of the analysis, all the documents have been classed into subfolders based on their theme and renamed with a title that summarises the content of the document using a few keywords.

To identify the content, similarly to how the topics have been identified, the 2 associations of words that were most popular in the document have been combined to form a title with 4 words. 4 words have been used as 2 was not sufficient to avoid duplicates. A few documents were named “Risk management”, by adding another set of 2 associated words it adds more information, example with document 1: Risk Management - Holt

**Example of the first 5 documents:**

| Original Document name | Topic (Subfolder) | Content (new file name) |
|---|---|---|
| Doc01.txt | Risk Management | risk manag holt suggest.txt |
| Doc02.txt | Risk Management | risk manag risk appetit.txt |
| Doc03.txt | Risk Management | risk analysi histor data.txt |
| Doc04.txt | Risk Management | risk ignor ignor metaphor.txt |
| Doc05.txt | Risk Management | project manag warn sign.txt |

See appendix number 3 for the final results of all documents.

Thus the documents have been renamed and re-arranged accordingly in an attached zip.file named “Documents”

While it is faster to do this process manually for 42 documents (opening doc to check content, type and wordcount), the model built can now be used for future similar tasks and for a much wider range of documents in a timely manner.

The process of renaming files seem to have worked better with more relevant document names for topic “Project Management Best Practice”. “Risk Management” and “Probability Distribution Completion Time” than the remaining two which include less text and some code with examples hence titles such as “figur figur” and “tmmapdoc”.

This could be reviewed by manipulating stopwords how carefully as this also impacts how they are then classified by topic and their similarity evaluation.

**APPENDICES**

**APPENDIX 1 – Example of document term matrix**

| | ridicul | rife | right | rigor | rigour | rise | risk | riski | rittel |
|---|---|---|---|---|---|---|---|---|---|
| Doc01.txt | 1 | 0 | 0 | 0 | 0 | 0 | 72 | 0 | 1 |
| Doc02.txt | 0 | 0 | 3 | 0 | 0 | 0 | 73 | 0 | 0 |
| Doc03.txt | 0 | 0 | 0 | 0 | 0 | 0 | 25 | 0 | 0 |
| Doc04.txt | 0 | 0 | 1 | 0 | 0 | 0 | 16 | 0 | 0 |
| Doc05.txt | 0 | 0 | 0 | 0 | 0 | 0 | 12 | 0 | 0 |
| Doc06.txt | 0 | 0 | 0 | 0 | 0 | 0 | 67 | 1 | 0 |
| Doc07.txt | 0 | 0 | 0 | 0 | 0 | 0 | 40 | 0 | 0 |
| Doc08.txt | 0 | 0 | 2 | 4 | 0 | 0 | 79 | 0 | 0 |
| Doc09.txt | 0 | 1 | 1 | 0 | 0 | 0 | 78 | 0 | 0 |
| Doc10.txt | 0 | 0 | 1 | 0 | 0 | 0 | 30 | 0 | 0 |

**APPENDIX 2 – LDA MODEL + GIBBS SAMPLING OUTPUT**

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

**APPENDIX NUMBER 3 – SUBFOLDER AND DOCS RENAMED**

| Original Document name | Topic (Subfolder) | Content (new file name) |
|---|---|---|
| Doc01.txt | Risk Management | risk manag holt suggest.txt |
| Doc02.txt | Risk Management | risk manag risk appetit.txt |
| Doc03.txt | Risk Management | risk analysi histor data.txt |
| Doc04.txt | Risk Management | risk ignor ignor metaphor.txt |
| Doc05.txt | Risk Management | project manag warn sign.txt |
| Doc06.txt | Risk Management | social construct risk manag.txt |
| Doc07.txt | Risk Management | strateg risk long term.txt |
| Doc08.txt | Risk Management | risk manag mont carlo.txt |
| Doc09.txt | Risk Management | risk manag project manag.txt |
| Doc10.txt | Risk Management | score techniqu score method.txt |
| Doc11.txt | Issue Mapping | rittel kunz real time.txt |
| Doc12.txt | Issue Mapping | decis make use ibi.txt |
| Doc13.txt | Issue Mapping | figur figur shown figur.txt |
| Doc14.txt | Issue Mapping | figur figur said mari.txt |
| Doc15.txt | Issue Mapping | inform knowledg formal knowledg.txt |
| Doc16.txt | Issue Mapping | dialogu map share understand.txt |
| Doc17.txt | Issue Mapping | issu map infrastructur technolog.txt |
| Doc18.txt | Machine Learning - Text Mining | machin learn fit train.txt |
| Doc19.txt | Machine Learning - Text Mining | text mine doc tmmapdoc.txt |
| Doc20.txt | Machine Learning - Text Mining | doc tmmapdoc hierarch cluster.txt |
| Doc21.txt | Machine Learning - Text Mining | doc tmmapdoc topic topic.txt |
| Doc22.txt | Machine Learning - Text Mining | doc tmmapdoc adjac matrix.txt |
| Doc23.txt | Project Managment Best Practice | mechanis mine assembl line.txt |
| Doc24.txt | Project Managment Best Practice | scapegoat approach activ error.txt |
| Doc25.txt | Project Managment Best Practice | alvesson willmott practic manag.txt |
| Doc26.txt | Project Managment Best Practice | function stupid busi usual.txt |
| Doc27.txt | Project Managment Best Practice | side effect plan chang.txt |
| Doc28.txt | Project Managment Best Practice | organis cultur project success.txt |
| Doc29.txt | Project Managment Best Practice | project manag post bureaucrat.txt |
| Doc30.txt | Risk Management | knowledg creation project manag.txt |
| Doc31.txt | Project Managment Best Practice | system design communic path.txt |
| Doc32.txt | Project Managment Best Practice | project manag grabher note.txt |
| Doc33.txt | Project Managment Best Practice | best practic wick problem.txt |
| Doc34.txt | Risk Management | valid claim take week.txt |
| Doc35.txt | Probability Distribution Completion Time | complet time mont carlo.txt |
| Doc36.txt | Probability Distribution Completion Time | standard deviat normal distribut.txt |
| Doc37.txt | Probability Distribution Completion Time | random number mont carlo.txt |
| Doc38.txt | Probability Distribution Completion Time | complet time probabl risk.txt |
| Doc39.txt | Probability Distribution Completion Time | durat b complet time.txt |
| Doc40.txt | Probability Distribution Completion Time | complet time triangular distribut.txt |
| Doc41.txt | Probability Distribution Completion Time | triangular distribut simul run.txt |
| Doc42.txt | Machine Learning - Text Mining | decis boundari support vector.txt |