|  |  |
| --- | --- |
| Description: 18897 CIC A4 Portrait WordTemp_cropped.jpg | **ASSIGNMENT COVER SHEET** |
|  | |
|  | |
| |  |  | | --- | --- | | **SUBJECT NUMBER & NAME** | 36106 Data, Algorithms and Meaning | | **NAME OF STUDENT**  **(PRINT CLEARLY - SURNAME, FIRST NAME)** | Kent, William | | **STUDENT ID NUMBER** | 13285337 | | **STUDENT EMAIL** | William.J.Kent@student.uts.edu.au | | **STUDENT CONTACT NUMBER** | 0410 912 997 | | **DUE DATE** | 28/10/2018 | | **ASSESSMENT ITEM NUMBER/TITLE** | Assessment task 3A: Analysis and Interpretation of Unstructured Data | | * I confirm that the work submitted conforms with the university’s guidelines on academic integrity.   *Refer to the UTS policy on ‘Advice to Students on Good Academic Practice’*: <http://www.gsu.uts.edu.au/policies/academicpractice.html>   * I am aware of the penalties for plagiarism. This assignment is my own work and I have not handed in this assignment (either part or completely) for assessment in another subject. * If this assignment is submitted after the due date I understand that it will incur a penalty for lateness unless I have previously had an extension of time approved and have attached the written confirmation of this extension.   Please provide details of extensions granted here if applicable \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  **Signature of Student:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ **­­Date:** \_\_\_\_\_/\_\_\_\_\_ /\_\_\_\_\_  If submitted electronically tick here to indicate you agree with the above   | | | |

# ASSESSMENT TASK 3A

# Analysis an Interpretation of Unstructured Data

William Kent

Student ID: 13285337

## Brief

A folder containing 41 text documents of unknown content has been provided and insights into the contents and themes within the documents is required. The files are UTF-8 encoded.

## A Bag of Words

At a very basic level a text document is a combination of words from which meaning arises. Evaluation of the most common words in a document, or set of documents, can provide insight into the topics and themes covered.

Words that contain no important significance like *the*, *to* and *and* were removed along with any digits. The words within the documents were also reduced to their base form, for example organise, organised, and organising were reduced to *organis*. This helps to bring key words and themes to the fore.

After this processing a simple word count was completed. Figure 1 shows the 50 most common words found within the documents. As can be seen words like *project*, *risk* and *manag(e)* stand out and seem to indicate that most documents are about topics such as project or risk management.



Figure 1 – Word cloud showing 50 most common words found in the 41 documents.

Delving further into the list of prominent words returns *time*, *task* and *work*. These may be related to project or risk management topics but may also hint at another set of documents about time management.

There are also further interesting words like *word*, *cluster*, and *term*. These may be linked to a further category within the documents.

## Word Association

Although the single words hint at the contents of the documents it is not often that we use words in isolation; we link words to form meaning. To further our understanding of the documents the most common word associations were identified. The top 20-word associations can be seen in Figure 2.

Unsurprisingly the words *project manag(e)* and *risk manag(e)* come to the fore as the most commonly linked word. This confirms our understanding of the documents and means at least some documents fall within these categories. The inclusion of *complet(ion) time* also fits as it appears to compliment the words *time*, *task* and *work*.



Figure 2 - Word cloud showing 20 most common words associations found in the 41 documents.

In addition to the most common word associations are links like *mont(e) carlo* and *probabl(e) distribution* which hints at further evidence of documents about risk management. More interestingly though are words such as *doc tmmapdoc*, *tmmapdoc tospac* and *text mine* which suggest there may be a set of documents about text mining and analysis.

## Term Frequency – Inverse Document Frequency

Within a collection of documents, it is not always the case that the most common words cover the contents of all documents. A single long document about a specific topic may skew results and provide a false sense of what the documents are about. To account for this the term frequency – inverse document frequency (TF-IDF) weighting method diminishes the importance of words that occur more frequently and promotes more descriptive words in specific documents ([Jones, 1972](#_ENREF_2)).

The word cloud in Figure 3 highlights the word *risk* as having the most importance after the TF-IDF weighting has been applied. This reassures us that a significant number of the documents, within the collection, are about risk or risk management.



Figure 3 – Word Cloud showing the 50 most common words based on the term frequency – inverse document frequency weighting method.

Interestingly the importance of the word project has been reduced; it is not as prominent. This may suggest it is over used in only a few documents and previous analysis was giving it more importance than it was due.

There are still terms that suggest at a category of documents associated with text mining with the word *tmmapdoc* being present as well as words like *algorithm* and *model* being highlighted as document subjects.

The interesting words that come to the fore after TF-IDF weighting has been applied are ones like *hubbard* and *eleph(ant)*. There may be a subset of documents about elephants, or about someone named Hubbard, or there may be an elephant in the room. This shows a limitation with this type of analysis. Whereas humans can determine the difference between a document about elephants and a document using a metaphor to describe problems, or difficult situations, a machine doesn’t have this cognitive ability. As such we can’t jump to conclusions about a subset of the documents being about elephants, or someone named Hubbard.

## Clustering

Through the words within the documents we have been able to glean the major topics and themes. What we haven’t been able to determine is, can the documents be grouped in anyway to help with our understanding. To aid this a method known as hierarchical clustering can be used to group similar documents.

Hierarchical clustering is an algorithm that groups similar documents or objects ([King, 1967](#_ENREF_3)). The end-point is when each cluster contains only a single document. Figure 4 shows the cluster dendrogram and the similarity between the documents in the folder. From this analysis there appears to be a large homogenous group of documents that are similar and three other groups that form their own distinct subsets.

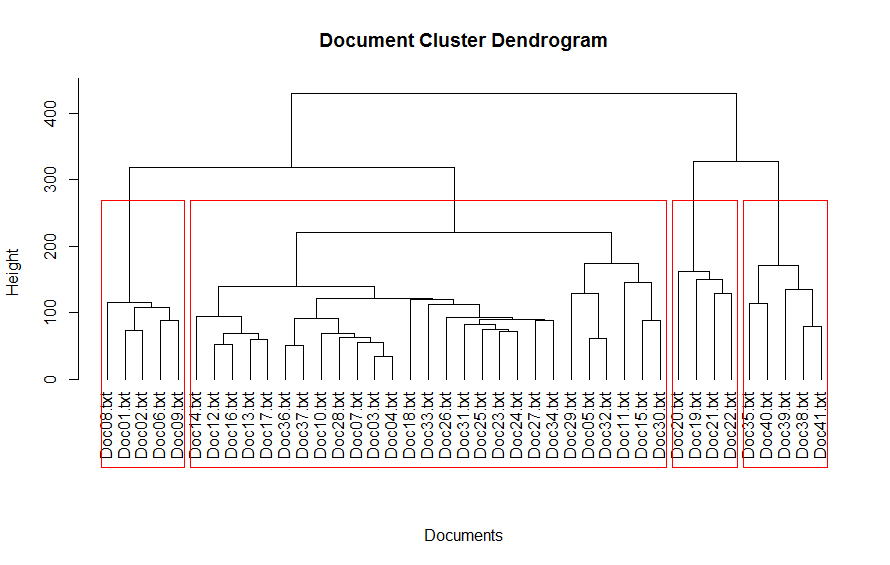


Figure 4 – Dendrogram highlighting the breakdown of the documents into distinct groups. The red boxes highlight the four distinct groups.

## Optimal Number of Clusters

In the hierarchical clustering analysis, I intuitively divided the dendrogram into 4 groups (or clusters). To take a more analytical approach the “elbow” method can be applied to help define how many groups is optimal. In this method the total intra-cluster variation is calculated for different numbers of groups. The idea is that the resulting graph will show a distinctive bend at a particular number of groups. After this point the reduction in variation is minimal.

Figure 5 shows the calculated plot using the “elbow” method, unfortunately without a distinctive bend. This makes determining the right number of groups to use more difficult.

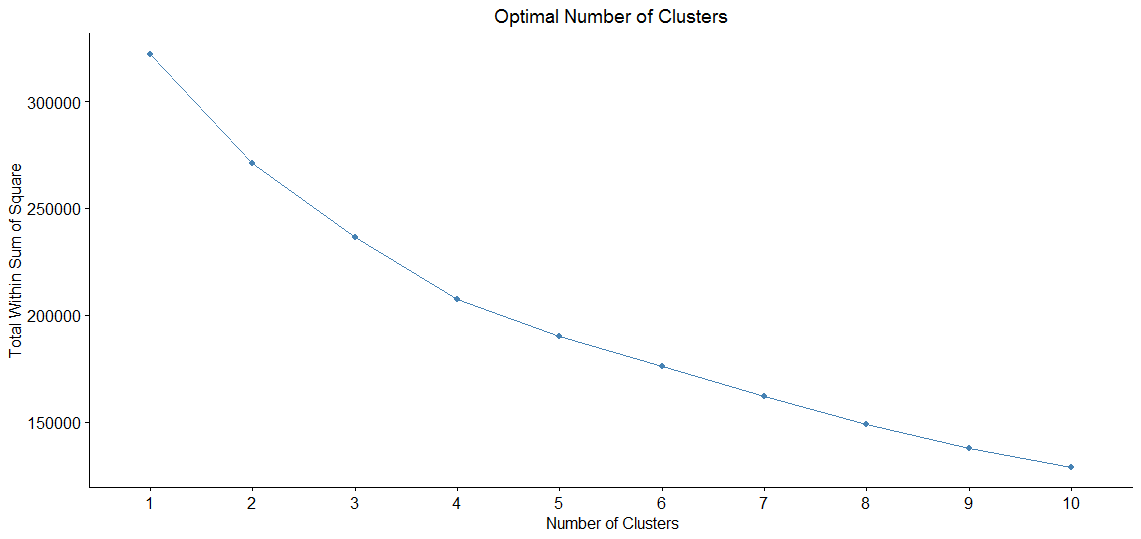


Figure 5 – Plot showing the total intra-cluster variation for different grouping levels.

## Topic Modelling

The use of hierarchical clustering provides insights into how documents relate to each other but doesn’t tell us the topics covered in each group. The use of the Latent Dirichlet Allocation (LDA) model can aid our understanding of the topics covered within each group by discovering statistical relationships between documents ([Blei et al., 2003](#_ENREF_1))

The LDA model can make statistical judgements about documents within the collection but it isn’t able to determine how many categories to group the documents into. Based on the previous investigation with hierarchical clustering and grouping we’ll continue to use 4.

The result is the categorisation of the documents into groups and the 8 most important words from each group can be seen in Figure 5. The groups seem to follow our intuitively identified classifications. There is a group that includes topics like risk and project management, another on completion times and probability themes. The previously identified terms related to text analysis are in another group. Group 3 seems to be less well defined with *best practice(e)* being the term most common to the documents.

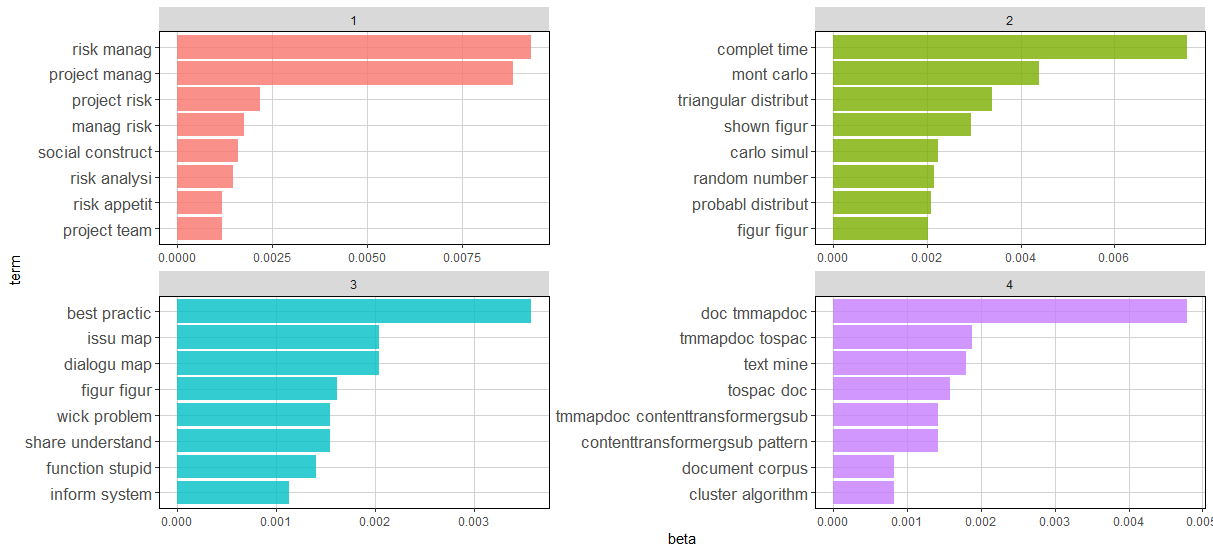


Figure 5 – Bar chart showing the 8 most important words in each category and their relative importance.

The breakdown of individual documents into the 4 categories can be seen in Table 1. Unsurprisingly the first group, which contains *risk manag(e)* and *project manag(e)*, has the greatest number of documents. This would help propel the previously seen important words of *risk*, *manag(e)* and *project* to the fore as they are contained within more documents.

The less well-defined group of Group 3 has the second highest number of documents. This is interesting as it may mean that the third group is a catch all for a disparate set of documents.

It is important to note that the categorisation completed using the LDA model is different to hierarchical clustering as they use different methods to determine similarity.

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

Table 1 – Topic Groups and the documents within each group

## Document Relationships

With document similarity known we can show the relationship between documents. Figure 6 is a network diagram that highlights the links between documents. A line between nodes indicates a relationship between documents. The nodes are coloured to indicate which group they were categorised as being in using the LDA model.

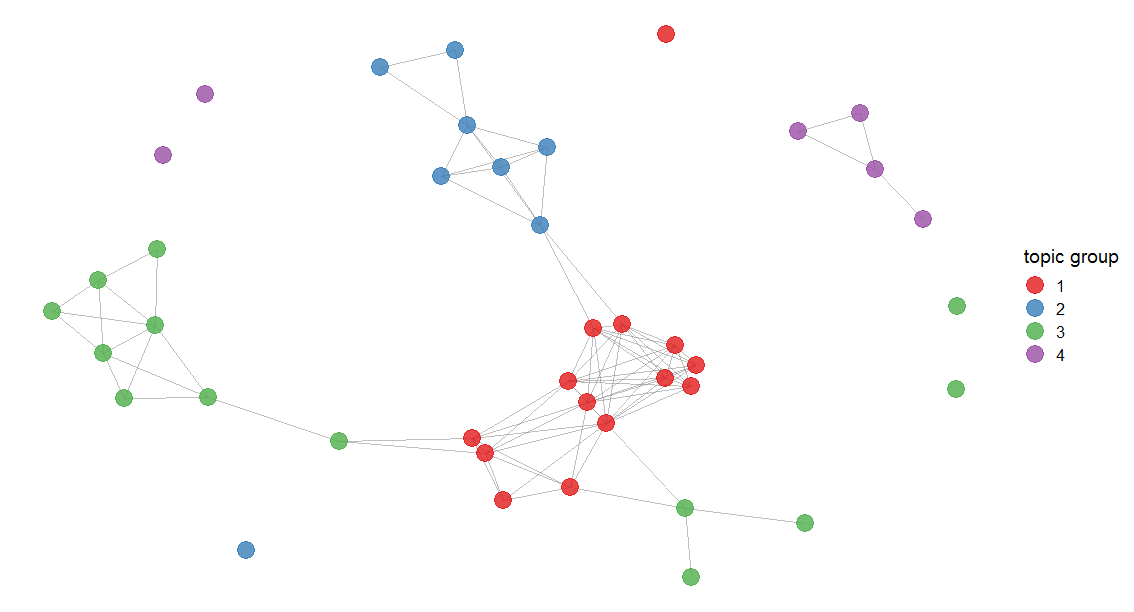


Figure 6 – Network graph showing how the documents within the folder relate to each other.

As expected the group of documents pertaining to risk and project management, highlighted in red, are central topics. They have multiple links between themselves as well as linking to the group of topics around completion time and probability as well as the group to do with best practices. The text mining documents do not link to other documents and may be less about management and more technical in nature.

## Conclusion

The primary theme of the documents in the collection appear to be around risk and project management. The folder also seems to have documents with a category around probability, monte carlo simulations and completion time, another on best practices whilst a further category of documents on text mining appear to be present. There does not appear to be any documents about elephants.

## References

## BLEI, D. M., NG, A. Y. & JORDAN, M. I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research,* 3**,** 993-1022.

## JONES, K. S. 1972. A Statistical Interpretation of Term Spesificity and its Application in Retrieval. *Journal of Documentation,* 28**,** 11-21.

## KING, B. 1967. Step-Wise Clustering Procedures. *Journal of the American Statistical Association,* 62**,** 86-101.