DEPARTMENT OF MANAGEMENT SCIENCE

University of Strathclyde Business School

Can data mining be utilised to improve knowledge sharing within DFID?

Suzanne Beith

201790528

MSc Data Analytics

Signed Statement

The contents of this report were created by the author, material obtained from other sources has been appropriately referenced.

Signed:

Summary

DFID manages many aid projects across the world, however, knowledge sharing between staff the who manage these projects is not common practice. The MI and Analytics department began working on clustering project data. This thesis continues this research by examining other possible approaches to text mining and cluster analysis. It highlights a gap in the literature around assessing clustering results when there is no ground truth with which to compare the results. K-Means, PAM, DBSCAN and Mean Shift algorithms were applied to the text data which had been processed, using a bag-of-words approach. Partitional algorithms were more successful, with K-Means providing the most sensible looking result in a business context. To evaluate the success of the clustering effort, semi-structured interviews were conducted with end users. Due to the small sample of clusters tested and only two interviews being conducted, a strong conclusion cannot be drawn. However, the results do show some initial success and provide a basis on which the work may continue.

Table of Contents

SCENE SETTING

CLIENT REPORT

1. Scene Setting

1.1 Introduction to DFID

The Department for International Development (DFID) is a division of the UK government which manages the provision of aid. Department goals include, ending extreme poverty, strengthening global peace and security, strengthening resilience and response to crisis, promoting global prosperity and delivering value for money (1). The UK is a member of the Organisation for Economic Co-operation and Development (OECD) and the Development Assistance Committee (DAC). DAC facilitates communication between 29 donor countries and European Commissions, setting the definition and classification for reporting on financing of international development. In addition, DAC set guidelines on what financial aid may be considered as Official Development Assistance (ODA). In short, to be considered ODA, aid must flow from an official agency to a country/territory on the ODA list with the purpose of promoting development in economy and welfare (2).

The International Aid Transparency Initiative (IATI) sets the standard for sharing of aid information across all organisations working in international development. Once published, this information is available for anyone to use. DFID values the using and publishing of IATI data for transparency and improvement. DFID use these data to better understand the reach and/or collective effort of initiatives, to identify relationships between partners, to better understand priorities and to allow managers to see a more high-level view of portfolio data (3). These data are also used in DFID's public Development Tracker, allowing the general-public to view information on aid spending (4).

There are two main types of aid, the first is bilateral, where the "recipient country, sector or project is known." The other is Multilateral, when a contribution is made to an aid organisation on the DAC list, without the definition of a sector or recipient. As this type of donation is put into a central fund, there is no way to identify contributions by country. It is important to note, that funds given to a multilateral organisation not on the DAC list are also considered bilateral, and some bilateral aid may be channelled through a multilateral organisation (2).

In 1969, the Pearson Commission published their report "Partners in Development", which proposed an ODA target of 0.7% Gross National Income (GNI). The following year, the OECD agreed this was a suitable target, and in 2005 the EU members (at that time) agreed to reach this target by 2015 [5]. Figure 1 illustrates the ODA spending for countries in 2016. The UK met the 0.7% target and was in the top 6 largest contributors in terms of ODA to GNI ratio (6).

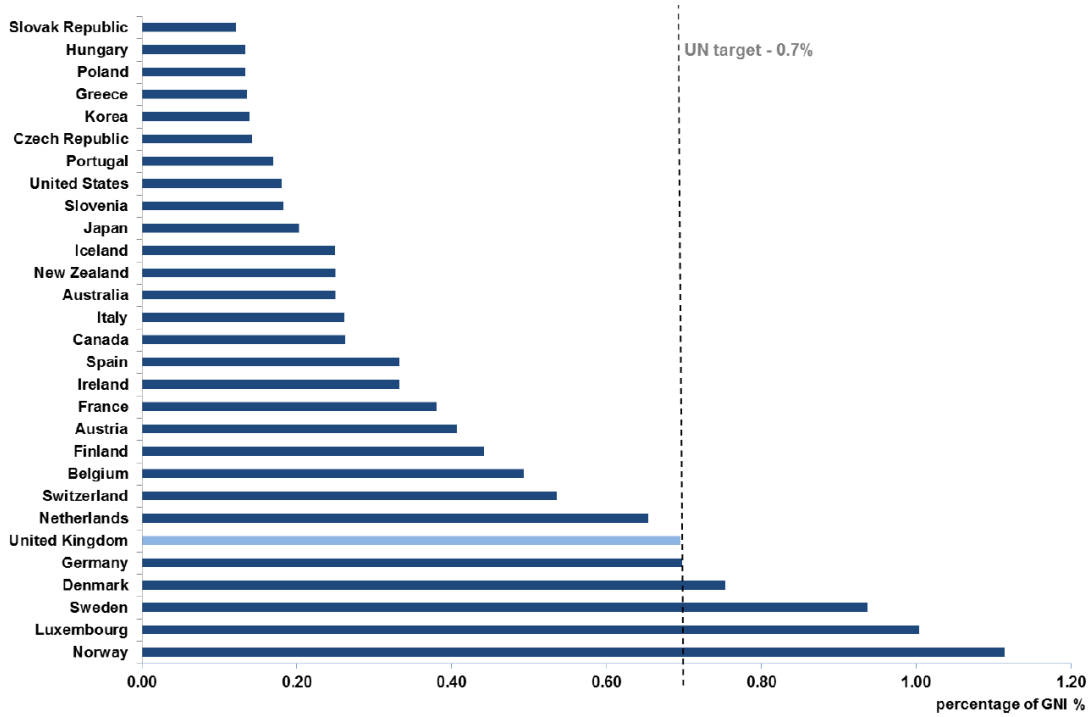


Figure 1: ODA:GNI spend by country, 2016

Bilateral ODA aid covers a broad range of sectors including humanitarian, multisector (environmental protection, research, and urban/rural development), health and education. Figure 2 shows the 2016 bilateral ODA spend on each sector, accompanied by the rank for 2015 in brackets. In both years, the largest spend was on humanitarian aid (6).

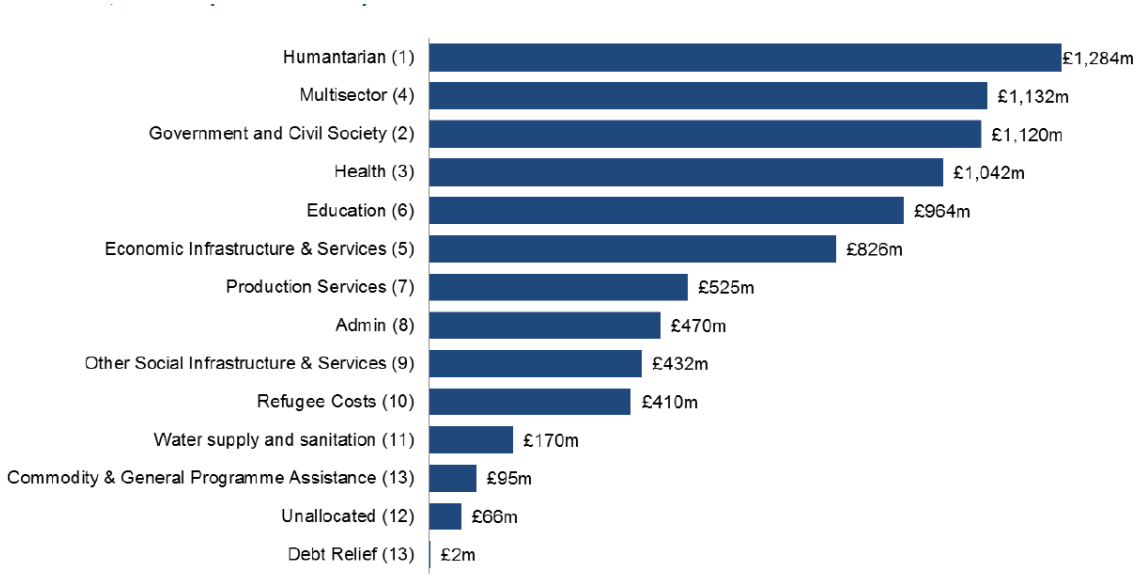


Figure 2: UK Bilateral aid spending by sector 2016

1.2 Research Objectives

The aim of this project is to use data mining tools and techniques to facilitate knowledge sharing between PMs and SROs working on similar projects within DFID. Whilst consulting with the department, key areas requiring research were highlighted, these include:

1) Approaches to text analysis
2) Clustering techniques
    a. Clustering techniques which do not require the number of clusters to be known a priori
    b. Methods to identify the number of clusters when unknown
    c. Clustering techniques which do require the number of clusters to be defined
3) Ways to measure the success of clustering

Based upon the research conducted, various approaches will be tested to assess which technique/combination of techniques provide the best result. All work must be conducted using the R programming language as this is the technology used by the department.

1.3 Project Plan

Literature Review Deadline: 11th June

- Natural Language Processing / text analysis
- Clustering algorithms
- Methods for estimating numbers of clusters
- Measure for comparing success of clusters
- Background reading on international development etc. for context
- Appropriate size of dataset for clustering

Dataset Preparation Deadline: 25th June

- Selection of appropriate data attributes for clustering
- Data cleaning
- Creation of an acronym dictionary
- Assess clustering tendency of the data

Implementation of Clustering Deadline: 6th August

- Calculating the appropriate number of clusters/topics for partitional clustering
- Implementation of clustering algorithms found through literature review
- Measure the success of each clustering algorithm

Analysis and Discussion Deadline: 27th August

- Analyse the results of different approaches completed and discuss the outcome

Contingency Time: 28th August – 17th September

- Additional time available should any issues arise.

Complete Write Up Deadline: 26th September

- Complete the final draft of dissertation for submission.

References

1. About us [Internet]. GOV.UK. 2018 [cited 22 May 2018]. Available from: https://www.gov.uk/government/organisations/department-for-international-development/about
2. DFID. ANNEX 1: UNDERSTANDING AID EXPENDITURE STATISTICS. 2018.
3. DFID IATI Guidelines (Policy) [Internet]. GOV.UK. 2018 [cited 23 May 2018]. Available from: https://www.gov.uk/government/publications/dfid-iati-guidelines/dfid-iati-guidelines-policy
4. Development Tracker [Internet]. Devtracker.dfid.gov.uk. 2018 [cited 23 May 2018]. Available from: https://devtracker.dfid.gov.uk/
5. The 0.7% ODA/GNI target - a history - OECD [Internet]. Oecd.org. 2018 [cited 24 May 2018]. Available from: http://www.oecd.org/dac/stats/the07odagnitarget-ahistory.htm
6. DFID. STATISTICS ON INTERNATIONAL DEVELOPMENT 2017 - Final 2016 UK ODA spend statistics. 2017.

Department for International Development

University of Strathclyde Glasgow

Can data mining be utilised to improve knowledge sharing within DFID?

Suzie Beith
MSc Data Analytics
September 2018

Executive Summary

DFID manages a large portfolio of aid projects across the globe. There is a great deal of data associated with each project, however, staff may be unaware of similar projects. Technical staff often belong to a cadre, giving them an opportunity to meet others working in the same sector or on similar projects. Programme Managers (PM) and Senior Responsible Officers (SRO) do not belong to any such group, making knowledge sharing difficult. The aim of this project is to build upon the preliminary work undertaken by the MI and Analytics department, researching various text analysis and clustering methods to create groups of similar projects. The aim is creating groups of similar projects, making it easy for PMs and SROs to identify who to contact should they wish to learn about another project.

This report initially researches various approaches to text analysis including TFIDF, Word2Vec, Doc2Vec and GloVe. In addition, many methods of clustering are examined, covering hierarchical, partitional and density based models. The desired number of clusters is unknown and there is no existing result with which to compare the results of clustering efforts. Therefore, various approaches to obtaining the number of clusters are researched, as are criterion to measure the effectiveness of clustering. A gap in the literature around assessing clustering results when there is no ground truth is identified. Reference is made to expert knowledge, however, no methodology for such assessment is provided.

The Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology is used to implement modelling using data obtained from the DFID analytics platform. CRISP-DM was selected due to its iterative nature and inclusion of business context in this process. Text analysis using the bag-of-words (BOW) model was implemented, with K-Means, PAM, DBSCAN and Mean Shift algorithms being applied. Many feature combinations were tested with each algorithm, always coming back to the business problem at hand. The final model which was presented to end users for testing was creating using the K-Means algorithm on project Title, Purpose, DAC 3, DAC 5 and Risk Description data.

Due to gaps in the literature, social science and human computer interaction texts were examined to create a set of non-leading interview questions. A small sample of clusters were presented to a handful of end users to assess if the results made sense in a business context. Half of the clusters were valid, however, it was highlighted that further distinction within the two of the groups could be made. Two clusters had conflicting results from interviewees and participants found the connection between projects in the remaining clusters unclear.

While this is a small sample of clusters, tested on a small sample of staff, it shows some success and provides a basis on which to continue working. Recommendations are presented which include further testing of text analysis approaches, fine tuning user interviews and evaluation of the eventual implementation to help further improve the model.

Acknowledgements

Working with DFID has been a great experience and I would like to take this opportunity to express my gratitude to Tom Wilkinson, Christie Hay and Calum Campbell for making this project possible. The team were a great source of advice and encouragement throughout the project, and for this I am truly thankful. I would like to thank all the staff at DFID for making me feel welcome. Furthermore, I would like to express my appreciation to my university supervisor, Kerem Akartunali for his support not only during this dissertation, but throughout the whole MSc. Thank you all.

Table of Contents

List of Figures

List of Tables

1. Introduction

DFID manages many projects from varying sectors across the world. There is documentation associated with each project, including Business Cases, Annual Reports (AR) and Project Completion Reports (PCR), in addition to other information such as Project Quality (PQ) score, spend and purpose. DFID staff and project managers may be unaware of similar projects, meaning they are unable to access the appropriate reports or contact staff who have managed "similar" projects. The similarity of a project may not be solely related to the sector of the project and may be influenced by factors such as geopolitics and/or economics. As such, there are no predefined groups for the projects.

1.1 Project Aims

The aim of this project is to identify whether text analysis and clustering can be applied to enable Programme Managers (PM) and Senior Responsible Officers (SRO) to share knowledge more effectively.

1.2 Previous Work

DFID have completed some initial work, in which the project is considered as a text analysis problem. As the reports available have been collected over decades of work, they are all varying format making text mining a difficult task. Therefore, text analysis was conducted on data obtained from DFID analytics, DFID's excel based tool for analysis of finance, project and supplier data. The features included were project title, project purpose, DAC 3-digit sector code (e.g. education, health) and DAC 5-digit sector code (e.g. primary education, secondary education).

Initially, text analysis was also conducted on the 127 DAC 5 sector names to evaluate what the most frequent terms were. Figure 1 shows the most frequently occurring important words, whilst figure 2 shows the most frequently occurring pairs of words (bi-grams). Both word clouds highlight that health and education projects are very prominent. Figure 1 shows that there are many projects centred around policy, management and administration too.

Figure 3: Most frequently occurring words

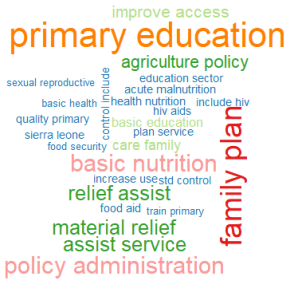


Figure 4: Most frequently occurring pairs of words

Hierarchical clustering was selected as it does not require the user to specify the number of clusters, unlike other commonly used clustering methods such a k-means. Figure 3 shows the hierarchical cluster which resulted from this process. There appear to be 5 distinct clusters of varying density. The central clade (branch) appears to contain 2 distinct clusters which are very different to each other, shown by the long leave.

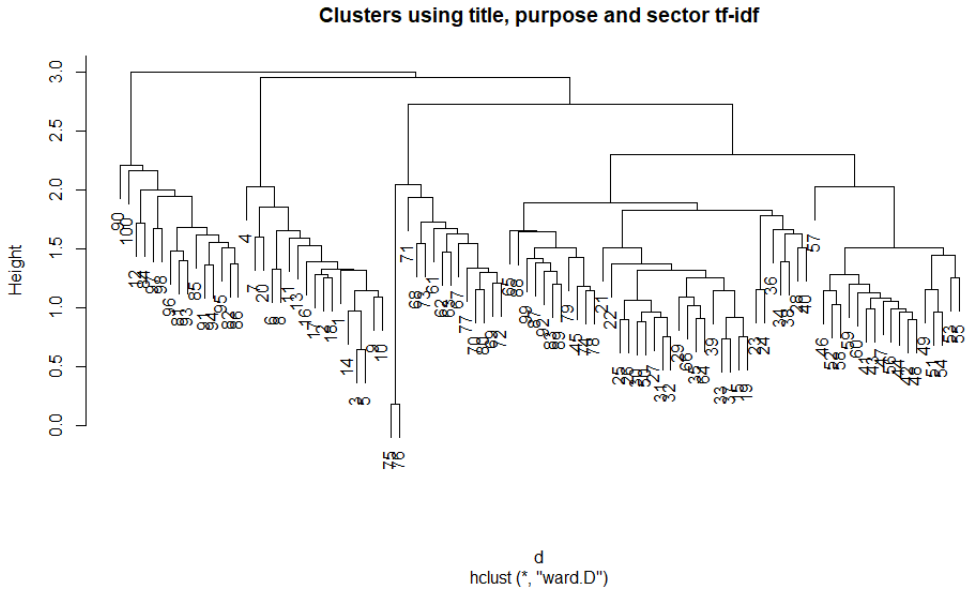


Figure 5: Hierarchical cluster created using text from project

The clusters created through this work cover topics of health and nutrition, education, humanitarian relief, reproductive health and agriculture. This analysis is based on word clouds created to show the top terms in each cluster.

1.3 Research Objectives

The aim of this project is to use data mining tools and techniques with the aim of facilitating knowledge sharing between PMs and SROs working on similar projects within DFID. Whilst consulting with the department, key areas requiring research were highlighted, these include:

1) Approaches to text analysis
2) Clustering techniques
    a. Clustering techniques which do not require the number of clusters to be known a priori
    b. Methods to identify the number of clusters when unknown
    c. Clustering techniques which do require the number of clusters to be defined
3) Ways to measure the success of clustering

Based upon the research conducted, various approaches will be tested to assess which technique/combination of techniques provide the best result. All work must be conducted using the R programming language as this is the technology used by the department.

1.4 Remainder of this report

Section 2 contains a literature review exploring different approaches to text analysis, clustering and evaluation of cluster results. Section 3 outlines CRISP-DM, the methodology used for this analysis. Section 4 documents the implementation of text analysis using the bag-of-words approach and TFIDF. Different methods used to find the number of clusters (*k*) are discussed, and the results of K-Means, PAM, Mean Shift and DBSCAN algorithms are discussed. Section 5 explains how the clustering results were assessed in a business context, interviewing end users to evaluate how useful the clustering effort is. Section 6 suggests how the project may move forward and section 7 concludes the report.

2. Literature Review

2.1 Text Analysis

Pathak states that text is "perhaps the most ubiquitous form of data", the world has an abundance of text in a variety of formats. Text mining, also known as Text Analytics, is the process of extracting latent information from text. Although text does follow linguistic rules which make it understandable to humans, it does not contain clear variables which may be used for analysis. Therefore, to analyse text, it must be processed in some way (1).

A commonly used text analysis pipeline for transforming free-form data into structured data begins with tokenisation. Anything which is not white space is considered a token. During this process it is common to convert all text to lower case, remove punctuation, numbers, special characters and stop words. Stop words are words which appear very frequently but provide no insight into the latent meaning of the document, such as prepositions and pronouns. Additional stop words may be added based upon the data being analysed. This creates a bag-of-words (BOW) for each document, in this model word order and grammar are disregarded.

Once tokenised, the data is stemmed to reduce all words to their semantic route. A Term Document Matrix (TDM) is created and the Term Frequency (TF) is calculate for each term. TF sums the amount of times a word appears in a document. To normalise the impact of long documents and reduce the effect of words which appear frequently across the corpus of documents, the next step is to calculate Term Frequency-Inverse Document Frequency (TF-IDF), where Inverse Document Frequency (IDF) increases the weight of terms which appear infrequently across the documents and decreases the weight of terms which appear often (1, 2). Not all terms appear in all documents, therefore, the matrix contains many 0 values; it is a sparse matrix.

The BOW model can be extended using n-grams, whereby, the order of words can be considered. This may be bi-grams, tri-grams or greater. Below is an example of how a simple sentence would be processed in a standard (unigram) BOW model and a bi-gram model. The text has been represented in an unprocessed state (no stop word removal, stemming, etc.) for simplicity.

Unigram BOW model

"Strengthening" "of" "the" "primary" "education" "system"

Bi-gram BOW Model

"Strengthening of" "of the" the primary" "primary education" "education system"

However, recording word order in this way hugely increases the dimensionality of the already sparse matrix. The sparse matrix produced by text analysis pipeline is said to suffer from the curse of dimensionality (3,7). A common approach to reduce the dimensionality of the TDM is Latent Semantic Analysis (LSA). LSA performs Singular Value Decomposition (SVD), which factorises the matrix, thus, creating linear composites of the TD-IDF features. As a result, LSA can capture semantic spaces in the data (4), terms which have similar meaning are likely to appear in similar context (5). As the matrix produced by TF-IDF is sparse, Langer suggests LSA be included as part of the standard text processing pipeline. Implementation of LSA is a computationally intensive process (2). As such, this may prove problematic for large datasets.

Another approach to dimensionality reduction commonly used in text analysis problems is Principal Component Analysis (PCA) (47). The aim of this approach is to extract the most pertinent information from the data while reducing the dimensionality (50). PCA works by creating new variables, known as Principal Components (PC). These components are "linear combinations of the original variables", calculated using SVD of the table or matrix. These new variables have associated values for each observation, known as factor scores. The first PC must have the largest variance among all PCs. The second component calculated must be orthogonal (not correlated) to the first, the third must be orthogonal to the second, and so on (48-50). Eigenvectors indicate the direction of PCs and eigenvalues denote the size and variance (49). One technique to identify the number of components to retain is plotting eigenvalues according to their size, such that the resulting plot has an elbow where the line changes from steep to flat. Components after the elbow (where the line goes flat) are discarded (50). An example of this can be found in figure 16. An alternative approach is to discard components who's eigenvalue is less than the average eigenvalue (50). A disadvantage of LSA and PCA is that it can be difficult to understand the dimensions constructed (47).

The BOW model has been successfully used for text analysis for many decades and produces document representations which are easy to interpret (6). However, it does suffer from sparsity as previously discussed. In recent years, other methods for text analysis have been created such as Word2Vec and Doc2Vec, both models utilise shallow neural networks to train the model. The traditional text analysis approach previously discussed does not consider the context of surrounding words, whereas new models do. There are two architectures of Word2Vec, Continuous Bag-of-Words (CBOW) and continuous skip-gram. The models are very similar, neural networks with one hidden layer. The architecture is based on the product of two-word vectors (9). Both models take a one-hot

(binary) encoded matrix as input. CBOW predicts the current word based upon the surrounding words which provide context. While, the skip-gram model attempts to classify a word context based on the word itself (8, 9). Figure 4 shows the network architecture for both models.

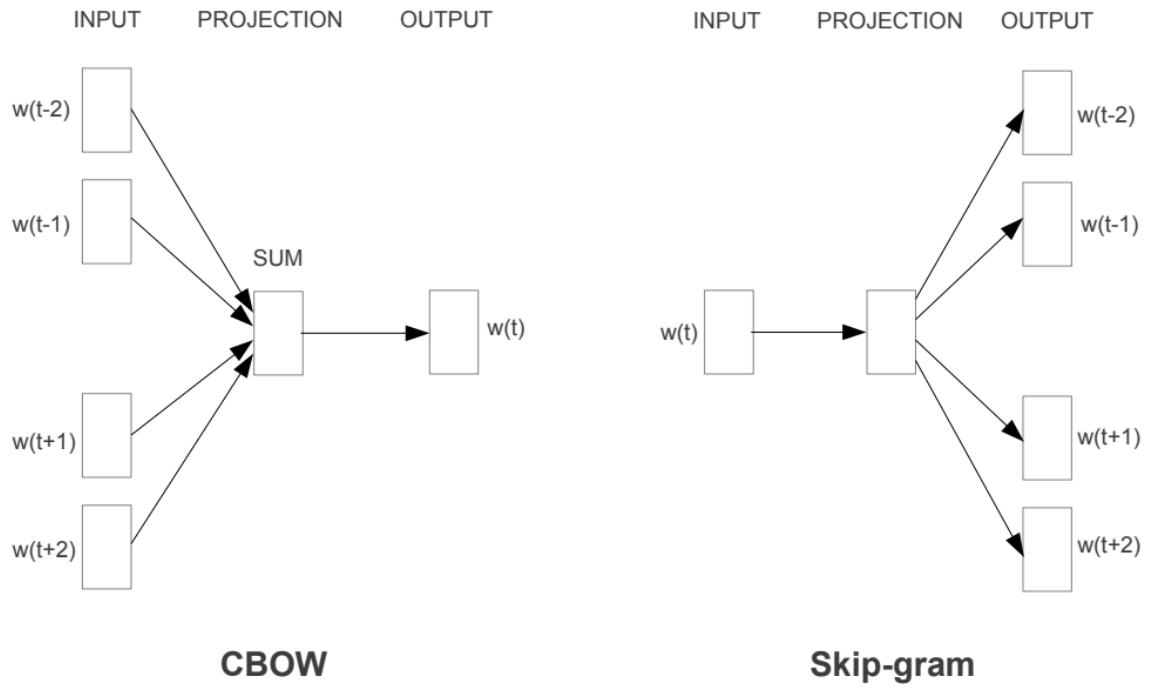


Figure 6: Network architecture of CBOW and skip-gram models (8)

These models are sometimes referred to as shallow window-based models, as they learn based upon adjacent words using a shallow neural network (11). The CBOW model is quicker to train and gives better accuracy for frequently appearing words, while skip-gram is slower to train, it has been shown to give good accuracy for both frequent and rare words when tested (10). Both Word2Vec models assume that words which appear in similar contexts have similar meanings. Therefore, when words are represented in vector space similar words will be in close proximity to each other, while unrelated word will be further away.

Doc2Vec builds on Word2Vec, using the same architecture and training approach. However, Doc2Vec embeds documents along with their words. A word vector will be created for each word, and a document vector for each document. Whilst Doc2Vec has shown to be effective in document clustering and classification tasks it suffers from a "lack of intuitive interpretation behind its generated document vectors" (6). As with any neural network based learning, it is difficult for humans to interpret what is happening in the model. It is unclear what the representation of document features are in terms of vector value. In some use cases this will not matter, but often, being able to analyse how the model is treating the data is an important part of analysis. It must also be noted, at the time of writing there was no Doc2Vec implantation available for R.

Another disadvantage of shallow window-based models is they do not operate on the corpus co-occurrence statistics. As they scan context windows (neighbouring words) across the entire corpus, they fail to take advantage of word repetition. The Global Vectors for Word Representation (GloVe) model proposed by Pennington, Socher, and Manning (9) captures the global corpus statistics. The authors suggest studying the ratio of each words co-occurrence probability in the context of other words, *Pik|Pjk*. For example, the ratio of *P(solid | ice) | P(solid | steam)* will be large. If *k* is changed to gas, which is more related to steam and less related to ice, the ratio will be small. If *k* were changed to water or fashion, the ratio expected will be close to one. The authors argue that this demonstrates when implementing word vector learning, the ratios of co-occurrence probabilities should be used rather than the probabilities themselves. To overcome the issue of infrequent co-occurrencesa weighted least square regression is used. Training of the model is computationally expensive, with examples in the paper taking between 6 and 24 hours.

"The Euclidean distance (or cosine similarity) between two word vectors provides an effective method for measuring linguistic or semantic similarity of the corresponding words. Sometimes, the nearest neighbours according to this metric reveal rare but relevant words that lie outside an average human's vocabulary. (12)"

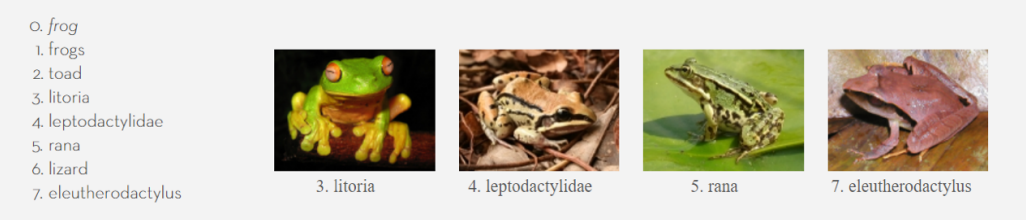


Figure 7: Words closest to frog (12)

2.2 Clustering

Clustering is concerned with discovering groupings in data, such that similar data points are placed in clusters and dissimilar data points are in separate clusters (13). However, as the purpose of clustering algorithms is to find clusters in data, an algorithm will create clusters, even if they are not inherent in the data (14). Therefore, before clustering, it is essential to assess the clustering tendency of the data. This can be done by visual methods which represent the similarity and dissimilarity between rows in the data or based on statistical evaluation. An example of a statistical approach is the Hopkins statistic (H) which may be used to assess the suitability of the data for clustering.

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$$

The Hopkins statistic null hypothesis is, the data set D has uniform distribution, meaning it has no meaningful clusters. The alternative hypothesis states that the data is not uniformly distributed, therefore, it contains inherent clusters. If H is greater than 0.5 the null hypothesis is rejected (15).

There are two main types of clustering, hard clustering where a data point is associated with only one group (16) and fuzzy clustering, where a data point may belong to more than one cluster with varying degrees of membership (17). There are many clustering algorithms, but the broad groups can be described as partitional, hierarchical and density based.

Hierarchical clustering produces nested clusters and is commonly used for text based clustering as the number of clusters is not required a priori. There are two types of hierarchical clustering, the first is Agglomerative which uses a bottom up approach (18). Initially, each data point is a cluster, these clusters are systematically amalgamated, based on some linkage criteria (19, 20). Clusters are repeatedly joined until the hierarchical tree is formed (21). Divisive clustering begins with one cluster containing all data points, this cluster is split, with each subsequent cluster being split until every data point is in an individual cluster. Divisive hierarchical clustering is computationally expensive, even on a small dataset, therefore it's not as widely used as the agglomerative method (22).

While hierarchical clustering has been successfully used for document clustering, there are some disadvantages to the algorithms. Primarily, the algorithm imposes a hierarchical structure on the data which may not be reflective of the natural data structure (19). The k-means algorithm has been shown to perform better than Agglomerative nearest-neighbours hierarchical clustering in terms of document clustering. Due to the probabilistic nature of the hierarchical model and each document only containing a subset of the total lexicon, documents which contain many of the same words are often considered as nearest neighbours, even if they are not of the same class. Once clustered, any mistakes in grouping cannot be changed. K-means may outperform Agglomerative clustering in a text domain, however, this is not true for all types of data (23).

K-means is a hard, partitional method of clustering commonly used for documents. The number of clusters centres *k* is set, then the algorithm will minimise the within class Sum of Squares (SS) from each centre (24). SS in minimised in an iterative optimisation procedure. In a standard implantation K-means, centres are randomly assigned, then the cluster prototype matrix (CPM) is calculated.

$$M = [m_1 \ldots, m_k]$$

Each object is assigned to the nearest centroid, based on the minimum least squared Euclidean distance. After this, the cluster prototype matrix is recalculated. This process is repeated until there is no change to any cluster after calculation (25).

Assigning each object to the nearest cluster:

$$\mathbf{x}_j \in C_l, \text{ if } \|\mathbf{x}_j - \mathbf{m}_l\| < \|\mathbf{x}_j - \mathbf{m}_i\|$$
$$\text{for } j = 1, \ldots, N, i \neq l, \text{ and } i = 1, \ldots, K;$$

Recalculating CPM

$$\mathbf{m}_i = \frac{1}{N_i} \sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j;$$

K-means is a useful tool for clustering large datasets with a time complexity of *O(NKdT)*.
The number (*N*), of (*d*) dimensional vectors is generally less than the number of clusters (*K*) and the number of iterations (*T*). Therefore, the time complexity is approximately linear (26), the run time will increase if the size of the data or the number of clusters are increased. Disadvantage of K-means are that it is assumed that clusters are spherical, multivariate and all have the same spread (27). In fact, K-means does not perform well on data which does not meet the criteria of assumption (28).

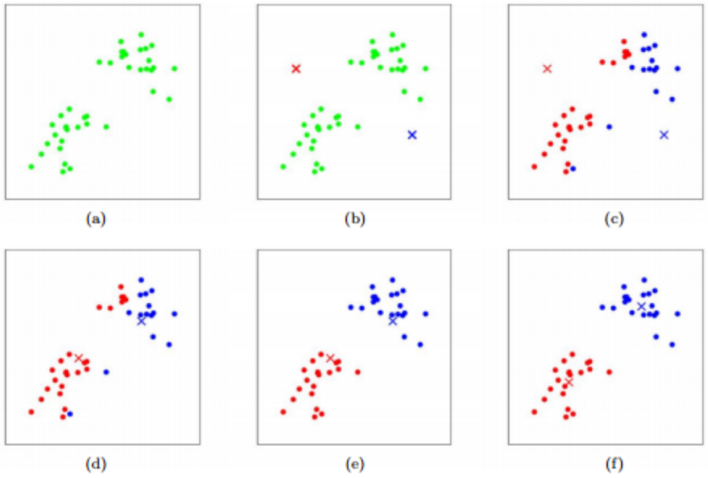


Figure 8: Visual representation of how k-means works (49)

Figure 6 provides a simple visual example of how k-means works. Plot (a) shows the dataset, plot (b) shows the initial centroids which are set at random, plots (c) to (f) demonstrate two iterations of the algorithm. The final four plots show points being assigned to their closest centroid, after this, the centroids are moved to the mean of the points assigned to them (49).

There are many similar partitional clustering algorithms such as K-Medoids, which uses median rather than mean to calculate the centroid. Compared with K-Means, K-Medoids is less sensitive to outliers due to the different method of calculating the medoid (29). One disadvantage of K-Medoids is, it requires many more iterations than K-means to converge. This is due to the iterative approach of using randomised-interchange *k* representatives which are incrementally improved (30). The algorithm initially chooses *k* centrally located points at random to be initial medoids and associates all other points to the closest medoid. After this, each medoid is swapped with each associated data point and the average dissimilarity between the 'new' centroid and all other points in the cluster are calculated. This process is repeated until no more changes are possible (31). This creates a great expense in terms of time and computation, which will be magnified when using a large dataset. To address this issue, the Clustering Large Applications (CLARA) algorithm was created. CLARA extends K-Medoids to deal with large datasets (32). To reduce the time it takes the model to converge, CLARA optimises its clusters using samples from the data, rather than the whole dataset (29). Clearly, clusters optimised in this way may not be representative of the whole dataset and the success of clustering will be very dependent on the size of samples used.

Another extension of K-Means is K-Modes, which is used for clustering categorical data. This algorithm uses a simple dissimilarity measure for categorical objects, uses mode to calculate centroids and a frequency based approach to update modes. While this may be useful for some datasets, often data is a mix of numeric and categorical data. The K-Prototypes algorithm addresses this issue by combining the K-Means and K-Modes models. The dissimilarity measure used considers both numeric and categorical data. s^rrepresents the squared Euclidean distance which measure numerical dissimilarity and s^c measures dissimilarity of categorical attributes calculated by the number of mismatches of categories between two objects. The dissimilarity is $s^r + \gamma s^c$, γ is a weight which stops favouring either type of attribute (29). The algorithm is efficient as it uses the same method as K-Means.

$$d_2(X, Y) = \sum_{j=1}^{p} (x_j - y_j)^2 + \gamma \sum_{j=p+1}^{m} \delta(x_j, y_j)$$

The K-Prototypes algorithm shown above is made up two parts, the first is the squared Euclidean distance for measuring dissimilarity of numeric attributes and the second, is the simple matching dissimilarity measure for categorical data. Huang (29) showed that the run time of the algorithm increased linearly as the number of clusters and records were increased. As with any of the *k* based algorithms, the number of clusters is required a priori and it does not perform well on non-spherical data which is not uniformly distributed.

When the number of clusters is not known a priori, many approaches have been suggested to calculate the optimum number of clusters. Many of these approaches are subjective and are open to interpretation, as such, there is no definitive method of finding the optimum value for *k*. The elbow-

method (see Figure 7) is a commonly used method of finding the number of clusters. As the aim of K-means clustering is to reduce the within cluster SS, the elbow method suggests computing the clustering method multiple times, incrementing *k* by one each time. For each *k* computed, calculate the within cluster SS. Plot all SS values, the result should be a curve with a bend, this is the "elbow" which denotes the point at which increasing the number of clusters does not continue to reduce the SS value (33).

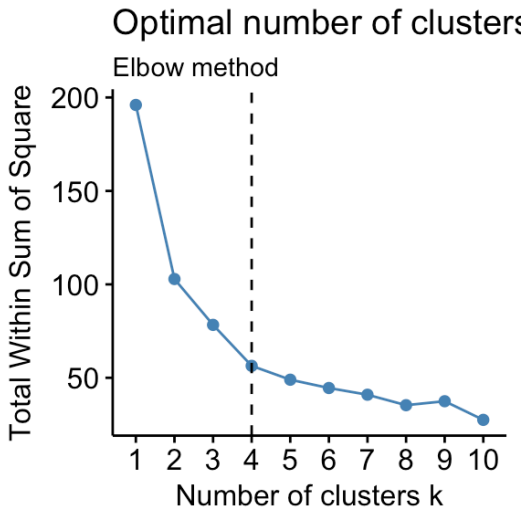


Figure 9: Example of finding k using the elbow method (33)

Silhouette Analysis works on a similar principal, compute the partitioning algorithm multiple times calculating the average silhouette score each time.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}.$$

a(i) is the average dissimilarity between *i* and all other objects in cluster C_r (the same cluster). *b(i)* is the minimum average dissimilarity of the *ith* object to all other objects in clusters (34). The *k* value which provides the highest silhouette score provides the "optimal" number of clusters (35). Figure 8 is the Silhouette analysis of the same data as with the Elbow method in Figure 7. Both approaches are heuristic and provide different estimates as to the "best" value of *k*, in addition, the business context has not been considered at any point during this process.

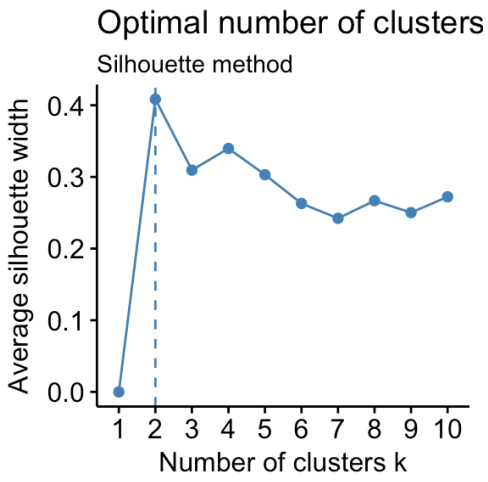


Figure 10: Example of Silhouette Analysis plot (33)

Another approach used to compute the number of clusters is the X-Means algorithm. X-Means is an extension of K-Means which automatically iterates through the process of incrementing *k* by 1 and computing the resulting clusters. X-Means adds new centroids in areas where they are required, whereas K-Means can only place the number of centroids defined by the user. The algorithm is fed a range of values in which *k* may plausibly lie. Posterior probabilities are used to score models created during the process.

$$BIC(M_j) = \hat{l}_j(D) - \frac{p_j}{2} \cdot \log R$$

$l_j(D)$is the log-likelihood of data according to the *jth* model, taken at the maximum-likelihood point, and p_j is the number of parameters in M_j (36). The BIC value is calculated for each model, the model with the lowest BIC score is considered most likely to be the best model (50). A major advantage of this model is the time saved by automation. However, once again, as an extension of K-Means this solution does not work well for non-spherical data which is not uniformly distributed.

Mean Shift is a non-parametric, density based algorithm, Unlike the K-Means algorithm and its extensions, Mean Shift does not assume any shape of clusters. The number of clusters is not required a priori, making the algorithm suitable for classification of unknown data. Like many other algorithms, it does make some assumption about the density of clusters, typically clusters are assumed to have a Gaussian distribution (37). The algorithm works in an iterative manner, initially setting a random seed and window size *W*.

$$\sum_{x \in W} xH(x)$$

The mean of *W* is calculated, then the search window is shifted based upon the new mean. This process is repeated until convergence. It is said that all points in a cluster fall within the attraction basin of a mode. Attraction basins are regions in which all trajectories lead to the same mode, demonstrated in Figure 9 (38).

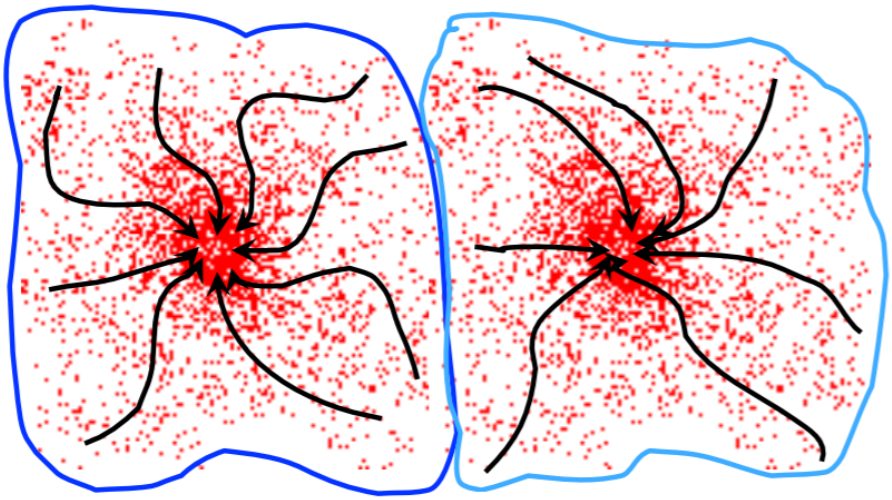


Figure 11: Visual representation of Mean Shift Cluster attraction basins (38)

While there are aspects of Mean Shift which may be beneficial in certain business cases, it does suffer from some disadvantages. Primarily, the computational complexity caused by repeatedly shifting many windows, many times. The window size has an impact on the result and selecting this is not a trivial task, as poor selection can cause modes to merge (38). Mean shift is commonly used for image visualisation however there are examples of its application in a document clustering context (37). A sparse matrix is often the result of text analysis; however, mean shift requires a dense matrix (39).

Density Based Spatial Clustering of Applications with Noise (DBSCAN) is one of the most well-known density based algorithms (40). DBSCAN became popular due to its ability to discover arbitrary shaped clusters and because the number of clusters is not required a priori (41). It works on the premise that data points which are *density-reachable* from each other, belong to the same cluster. DBSCAN estimates the density around a point using the ϵ-*neigbourhood* concept. The neighbourhood of data point p is a set of points within a specific radius ϵ around p. d is a distance measure and $\epsilon \in \mathbb{R}^+$.(40)

$$N_\epsilon(p) = \{q \mid d(p,q) < \epsilon\}$$

This, along with *minPts* are used to detect dense regions in the data and classify points into *core*, *border* or *noise* points. Points are defined as core if $N_\epsilon(p) \geq$ *minPts*, as border if they are not core but are in the neighbourhood of a core point, and as noise if they do not match either criteria (40). Figure 10 illustrates the assignment of points.

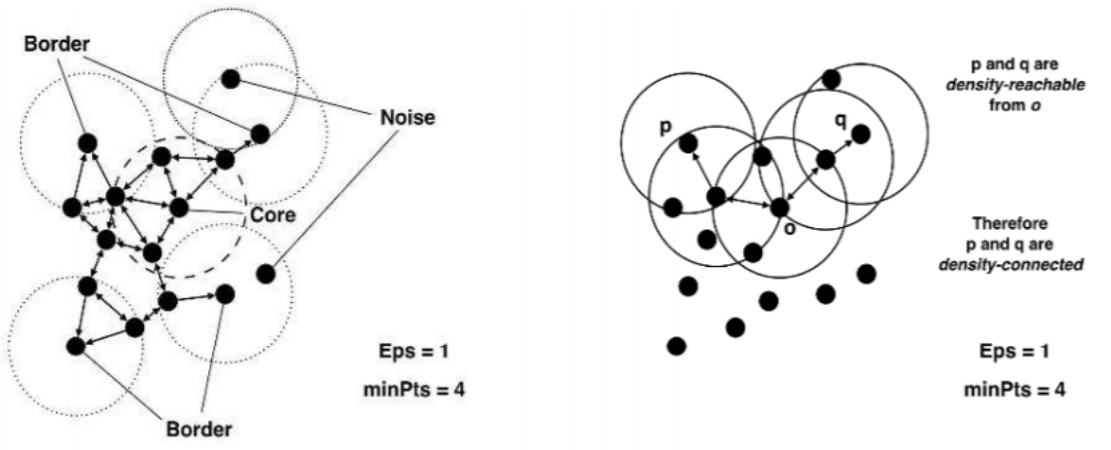


Figure 12: Assignment of points (40)

Different clustering algorithms will produce different data clusters, even the same algorithm can produce different clusters dependent on the parameters (35). Therefore, measuring each resulting set of clusters is important to identify the best algorithm parameters for the dataset. There are three different ways to assess the results of clustering (42). External criteria may be used if the cluster labels are known a priori, as the clustering results can be compared with known labels. This is often to referred to as ground truth. There are four external clustering criteria, F-Measure, Normalised Mutual Information (NMI) Measure, Entropy and Purity. It is rare to have ground truth with which to compare clustering results (51).

Internal criteria assess the success of clustering based upon information obtained from within clusters, this approach does not reference any external sources. The final approach is the relative criteria, which assess the clustering result produced by the same algorithm, run multiple times using different parameters. There are a selection of metrics which may be measured. An example is the Dunn Index which attempts to identify clusters which are tight but well separated (42), however, it can be over sensitive to noise. In theory, a higher Dunn index signifies better quality clusters (41). This measure will work for assessing the success of K-Means and its relatives as they are spherical with a defined centre point, but it will not prove useful for clusters of arbitrary shape resulting from some density based algorithms (43). Another example is the Calinski-Harabaz (CH) index which evaluates clusters based on the between and within sum of squares. A higher CH score is indicative of better clusters (52).

While these assessment criteria may be useful in a mathematical sense, the clusters must be assessed in a business context to ensure they are useful to the organisation. There is a lack of literature in the area of cluster assessment when ground truth is unknown. Kurgan and Musilek (44) highlight that much of the research conducted has been on improving Data Mining (DM) algorithms rather than evaluating knowledge extraction. Social Sciences (SC) have a great deal of literature about methods of collecting qualitative data. There are many ways to collect data such as questionnaires, interviews and

focus groups. Interviews are most appropriate for this work as the research focuses on complex issues and speaking with PMs and SROs will give valuable insights.

There are multiple types of interview; structured interviews have a rigid structure with predefined questions which are closed in nature. The purpose of this is to make data collection uniform, for easier analysis. Semi-structured interviews have a set list of topics to cover, however there is flexibility in the order in which topics are discussed and there may be many questions which are open in nature, allowing the participant(s) to talk in more depth about topics they have greater experience of. Finally, unstructured interviews are focused on the participant(s), the interviewer may introduce a topic to start the conversation, after this the interviewee(s) lead the discussion (53). Semi-structured and unstructured interviews will provide a more in depth understanding of the topics being discussed, however, the data may be more difficult to analyse than that of structured interviews, due to the non-uniform answers. Denscombe (53), highlights that there is more to conducting a successful interview than having a well-planned set of non-leading questions to discuss. Other factors which must be considered are the attentiveness of the interviewer, their ability to tolerate silence, knowing when to prompt the interviewee(s), knowing when to use probes to learn more about a specific topic, using checks to summarise what has been said and to conduct the interview in a non-judgemental manner (54). While interviews are a useful means of collecting data, they are time consuming, can be costly in terms of participant time, the data may be of an inconsistent structure, especially in the case of semi-structured and unstructured interviews and the answers provided may vary depending on the identity of the researcher (55).

In conclusion, there are multiple approaches to text analysis and clustering, however, it is not clear which will provide the best results for a given dataset. There is much literature with regards to clustering algorithms and great efforts have been made to enhance them, but there is a lack of research covering how to assess the results when there is no ground truth with which to compare the results. Therefore, qualitative data collections techniques from social sciences were researched as it will be necessary to examine the results using expert knowledge. Clusters are only valuable if they are useful in a business context, regardless of how well they perform in any mathematical metric.

3. Methodology

It is not clear which features or clustering algorithm will produce the best results for a given dataset. A commonly used Data Mining (DM) process has the following stages: Variable Selection, Clustering Algorithm Selection, Validation of Results, and Interpretation. However, this process does not account for the goals of the end user (56). Numeric validation metrics may be useful for comparing the results of different algorithm on the same data set, but they do not give any insight into how useful the resulting clusters are in a business context. As the purpose of this project is to facilitate knowledge sharing between PMs and SROs, the clusters must be evaluated in a business context to establish whether a technique has been successful or not. Kim (57), suggests comparison between clustering results should

only be assessed from a business perspective, rather than through mathematical metrics. Assessing clustering results in this way would pose challenges in terms of access to staff and time. Therefore, a combination of mathematical and user based evaluation were used; interviewing users to assess results from every algorithm and feature combination is not practical.

The methodology used for this project is the Cross-Industry Standard Process for Data Mining (CRISP-DM) model. The CRISP-DM model is an iterative model comprised of six stages, shown in Figure 11 (58). This approach has been adopted as unlike the traditional methodology previously described, CRISP-DM incorporates iteration within different stages of the process.

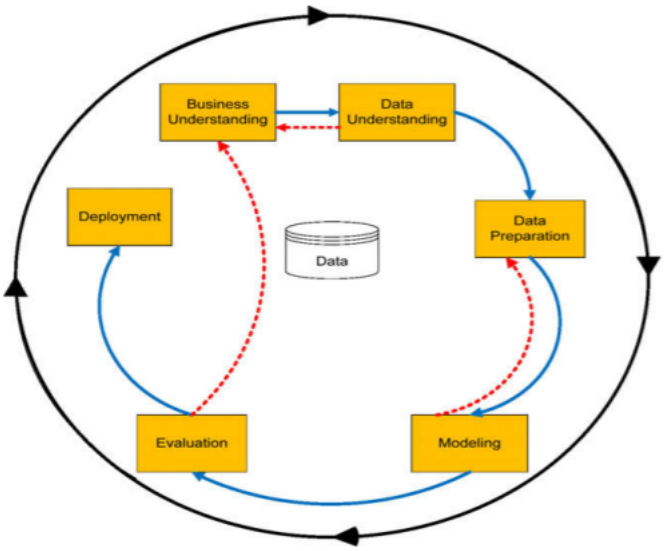


Figure 13: CRISP-DM Model (58)

The model recognises that as part of the DM process, going back and forth between different stages is often required. For example, feature selection will often take many iterations to discover which combination of attributes work best with the algorithm being tested. These iterations are denoted by solid blue lines when moving forward and dotted red line when moving back to make changes. The entire model is encapsulated in a large black circle, demonstrating that once the model is deployed, other business questions may be raised, resulting in improvements to the existing model (59).

4. Implementation using CRISP-DM Methodology

4.1 Business Understanding

The business objective of this DM task is to create groups of projects which will facilitate the sharing of knowledge between PMs and SROs. Given a group of projects, staff members should be able to identify who they can contact, with the aim of gaining some insight which will benefit the programme they are working on. DM results must be useful in a business context to be considered successful.

4.2 Data Understanding

There is a large amount of data available via the DFID analytics platform. Table 1 contains a data description of all features which were tested during the analysis. For security purposes, samples of the data have not been provided.

Feature	Description	Length
Title	Project title	Short, generally one sentence
Purpose	Description of the project	Ranges from 1 sentence to 1 paragraph
DAC3	Sector code information	1 sentence, there can be multiple sector codes per project
DAC 5	Sector code information	1 sentence, there can be multiple sector codes per project
Business Case	Business case produced at the design stage of a programme	Documents containing several paragraphs. Layout varies between documents. Many documents appear to have missing text.
Risk Description	Description of the risks associated with a component.	1 – 3 sentences, a programme may have multiple risk associated.
Income Group	Income category	Short, between one and three words
SID Continent	Continent of project	Short, Generally one word
Benefitting Country	Country Benefitting from project	Short, Generally one word
Fragile State/Region	Fragility of the region in which the project is	Short, generally two words

Table 1: Attributes tested during analysis process (a data sample is not provided due to security issues)

The first six features listed were all included in the analysis as they are created during the design stage of a project. The end goal would be for PMs/SROs to add details of their newly designed project to the system and be returned a list of suggested projects which enables them to easily identify staff members who may be able to share their experiences with them. The remaining five features were initially considered for inclusion, however, some of the attributes would not be available for projects in the design stage and other were discounted at a later stage when considering the business context of the problem at hand.

The code developed to create the dataset used during this project can be found in appendix A. During the iterations of testing it became apparent that administrative projects were included, the client highlighted that these projects are not useful in this analysis, therefore, they were removed.

4.3 Data Cleaning

Rows with missing data were removed as no suitable value can substitute missing text. All data was subject to the same text analysis pipeline in preparation for application of algorithms. Attributes were concatenated into a single string, creating a bag of words for each project. It should be noted that a

project is made up of several components, therefore, may contain multiple DAC3, DAC5 and risk descriptions. Common contractions were removed, common abbreviations were replaced, the text was tokenised (making each word or symbol a token), numbers, punctuation and URLs were removed. After this all words were converted to lower case, stop words were removed, then words were returned to their route form, known as stemming.

Due to some quirks with the Quanteda package, additional data cleaning was implemented after the document corpus was created. Quanteda only removes items which are tokens (with space at either side) meaning there were still numbers and special some characters in the text.

Quanteda will remove the number two

"2" "million" "people"

Quanteda will not remove the number two

"2million" "people"

Therefore, numbers and special characters (including punctuation) were removed using the tm_map function of the tm package. See appendix B for code which cleans data and plots top terms in corpus.

Creating a list of stop words was an iterative process with the list of words growing dramatically as the analysis progressed. A document term matrix was created and the most frequent terms in the corpus were plotted, this highlighted that many "business" words which do not contribute to the meaning of the document were appearing. An example of the word cloud used to determine which terms to add to the stop world is shown in figure 12. All possible conjugations of each root shown were added to the list. With each iteration, more words were added to the list. Much later in the analysis process, the decision was taken (in consultation with analytics team at DFID) to remove countries, continents and demonyms to prohibit clusters based on geography from forming. Staff working in a country office will already be in contact with staff from that office, therefore it is a redundant grouping to make from a business perspective. The final list of stop words can be found in appendix C.

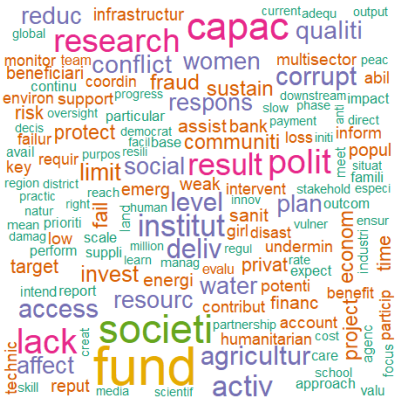


Figure 14: Top terms across corpus

In addition to removing stop words, terms which were more than 40% sparse were also removed. After this, the TF-IDF of the corpus was calculated and finally, the corpus was converted to a matrix. The final step was necessary for assessing clustering tendency, PCA and some heuristic methods of estimating, k which are covered in the next section.

To determine if there are inherent clusters in the data, the fviz_dist function from the factoextra package was used; see appendix D. It works by creating a visual representation of the Euclidean distance between rows (documents). Similarity is denoted by red tones, while dissimilarity is represented by blue tones. This was applied during every iteration of feature selection to ensure that clusters were inherent in the combinations of attributes being tested. Figure 13 demonstrates that Title, Purpose, DAC3 and DAC5 data contains clusters. This approach was favoured over the Hopkins statistic as it provides a clear, easy to interpret evaluation of the data.

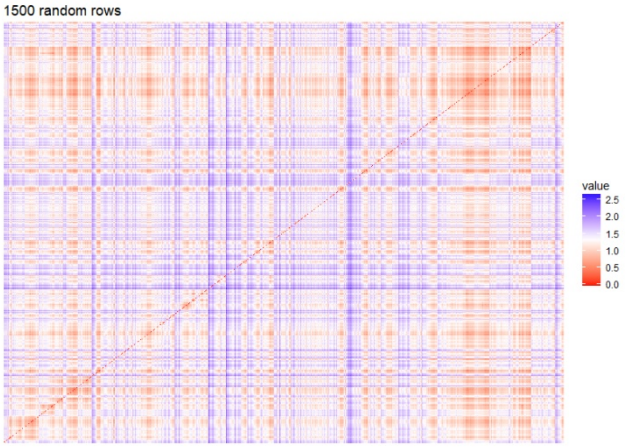


Figure 15: Clustering tendency of title, purpose, DAC3 and DAC 5 data

In general, most of the attributes/combinations of attributes proved to contain similarity between documents, apart from business cases which showed no similarity (figure 14). An additional issue encountered with the business cases was the amount of time it took to analyse the text. With a sample of 500 randomly selected documents it took several hours to process the data. While it is possible the business cases could have been useful when combined with other features, the computational cost of processing them and their apparent dissimilarity lead to the decision to omit them from the analysis. To make the business cases viable for use in this process, it may be beneficial to take a specific section, however, due to the irregular structure of the documents this could prove challenging. Alternatively, it may also require a more extensive list of stop words, arriving at this list would prove a time-consuming process as there is such a large volume of text.

Figure 16: Tendency of business case data

The final step before applying any clustering algorithm was to apply PCA to reduce the dimensionality of the matrix. The resulting plot of the cumulative frequency of variance within the data helps to determine which components may be omitted. Principal components which explain more than 98% of variance in the data may be removed. During the initial iterations of the project using features, Title, Purpose, DAC3 and DAC5, a sample of 1500 random rows were used for analysis. Figure 15 is an example of the plot used to assess how many features may be removed from the matrix. At around 1000 components, 98% of the variance is explained. Therefore, only 1000 components are used for clustering. See appendix E for PCA code.

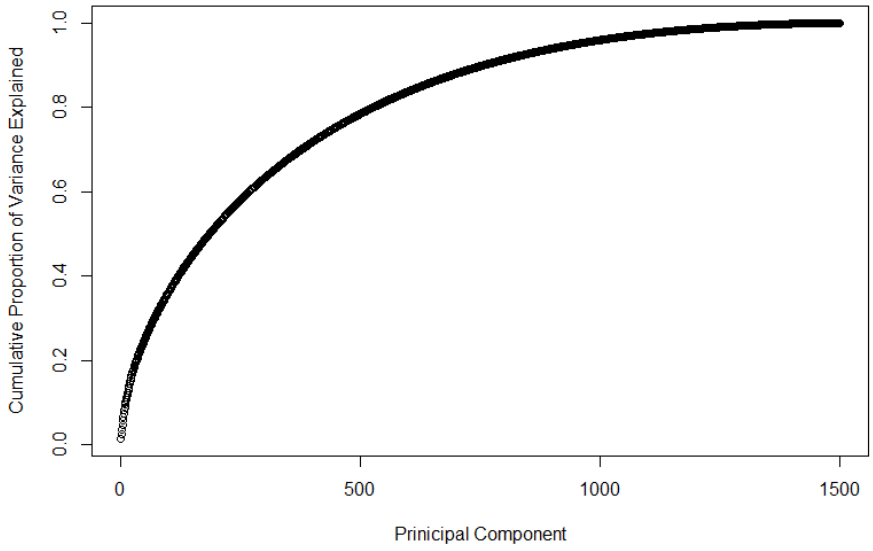


Figure 17: Principal Component Analysis - Cumulative Proportion of Variance Explained – 1500 random projects

It should be highlighted that PCA is not always necessary. After the risk description data was added, the sample size fell to 596 projects. The plot in figure 16 demonstrates that no dimensionality reduction is possible from this small sample of data.

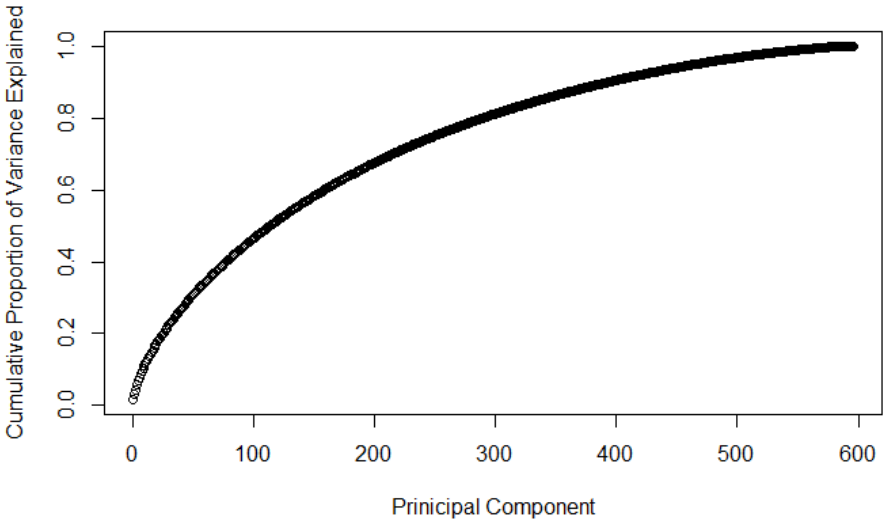


Figure 18: PCA - Cumulative Proportion of Variance Explained - Sample of 596 (Title, Purpose, DAC3, DAC5 and Risk Description)

4.4 Modelling

Algorithms tested during this analysis were K-Means, PAM, PamK, DBSCAN and Mean Shift. These algorithms were selected to test partitional and density based algorithms on this data, and also to provide a comparison of algorithms which require *k* a priori and those which do not. Whilst other algorithms were discussed in the literature review, not all were implemented. This was predominantly due to time constraints of the project and delays in installation of required R packages.

4.4.1 Partioning Algorithms

K-Means

Multiple approaches were used to estimate the best value of k. Elbow and Silhouette, heuristic methods consistently returned results of between 2 and 10 clusters, depending on the set of features being tested. However, DFID staff were interested in pushing the number of clusters to see if it were possible to create many small clusters which may be easier to interpret.

The Elbow function cannot be edited and only calculates the within sum of squares (WSS) for values between 2 and 10. Therefore to ascertain if more clusters would be useful for this data, code was created to calculate the WSS for *k* with values between 2 and 200; all resulting values were placed in one plot. K-Means has a linear time constraint, increasing the *k* value will increase the run time, therefore, the code created utilises multiple cores in the machine to speed up the process.

Figure 17 shows the WSS plot for Title, Purpose, DAC3, DAC5 and risk data, this plot shows that the WSS decreases as the number of plots increases and there appears to be a small elbow close to 200. Code for finding k (Elbow, Silhouette and WSS 2-200) can be found in appendix F.

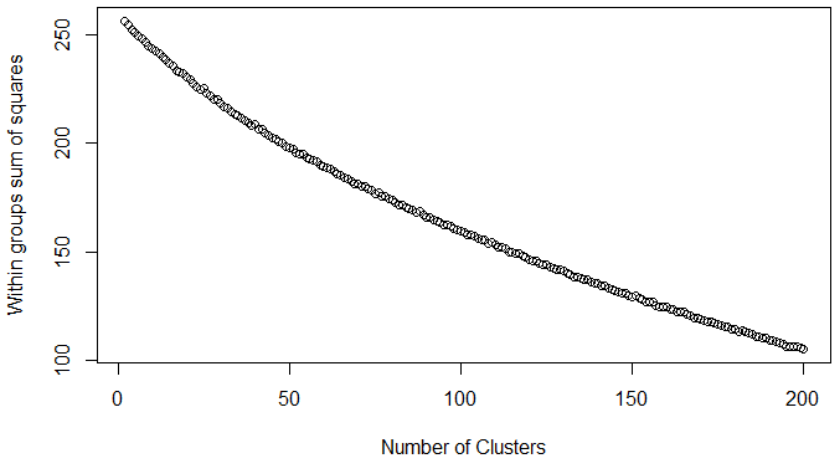


Figure 19: Within Sum of Squares. Title, Purpose, DAC3, DAC5 and Risk Description Data

Feature selection at *k* values varying from 50 to 200 (in increments of 25) were tested, to allow for analysis of word clouds from the resulting clusters. The code to run K-Means was also written to utilise multiple clusters; see appendix G. As expected, the runtime of the algorithm increased as the value of k increased, even when using multiple cores. Table 2 shows a comparison of the results of K-Means clustering on the same data sample. While there is more data available when risk is omitted, it would not produce a fair comparison of the results.

Features	**K**	**Calinski-Harabasz Score**
Title, Purpose, DAC3, DAC5	50	6.007
Title, Purpose, DAC3, DAC5	75	5.252
Title, Purpose, DAC3, DAC5	100	5.061
Title, Purpose, DAC3, DAC5, Risk Description	50	3.926
Title, Purpose, DAC3, DAC5, Risk Description	75	3.568
Title, Purpose, DAC3, DAC5, Risk Description	100	3.361
Title, Purpose, DAC3, DAC5, Risk Description	125	3.239
Title, Purpose, DAC3, DAC5, Risk Description	150	3.169

Table 2: Comparison of K-Means clustering on varying features

Clustering efforts were stopped when it became apparent that over clustering was happening (most clusters containing a single project). While clusters created without risk data score higher on the Calinski-Harabasz (CH) index and create a more even distribution of clusters, they may be less useful in a business context. Discussion with DFID staff highlighted that it may be more beneficial if PMs

were able to contact staff whose projects had some similarity in terms of risk to help inform their approach to a risk mitigation strategy.

The difference in CH score of clusters created using risk is negligible. What was interesting about these clusters is the difference in distribution. The lowest k values didn't appear to cluster very well, while the higher k values appeared to over cluster. Based upon the distribution of clusters and the sensibility of cluster word clouds, a *k* value of 100 was used as the final set of clusters. Figure 18 shows the distribution of clusters for all features listed at *k* equal to 100.

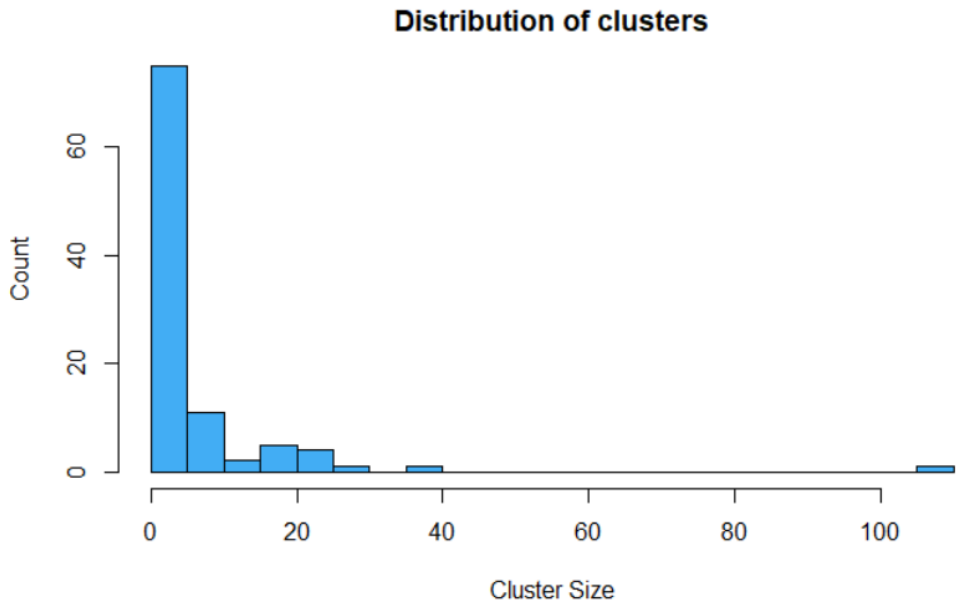


Figure 20: Distribution of clusters

The resulting clusters were generally a manageable size, except one, which contained 109 projects and appeared to have a broad range of topics included. Further clustering was tested on the largest cluster to try and make it a more manageable size. As clustering on the same features would not produce any further result, it was clustered based on Title, Purpose and Risk Data alone. These documents do appear to have similarity when the tendency is assessed, however it does appear that there is a large amount of documents which are similar (figure 19).

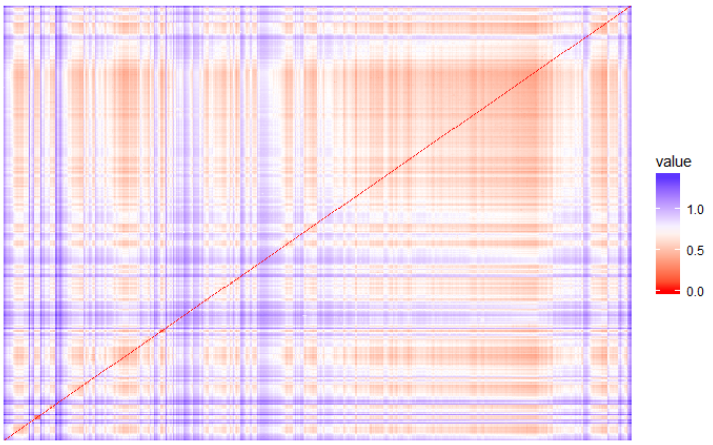


Figure 21: Tendency of Title, Purpose and Risk Data

Partitional and density based algorithms did not prove successful as they consistently over cluster, producing one large cluster and many single point clusters. This was not entirely surprising given the illustration of the tendency. The hierarchical code which had already been created by DFID was applied and it successfully created 6 further clusters (figure 20).

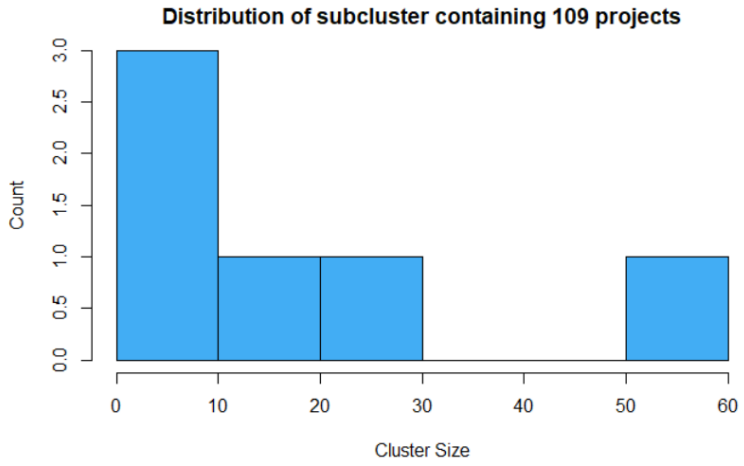


Figure 22: Distribution of sub-clusters

PAM

There are two versions of the PAM algorithm, PAM and PAMk. PAMk does not require *k* a priori, a range of plausible *k* values is given to the algorithm and it will evaluate what the best value is. Unfortunately, this proved to be unsuccessful as a result of 2 clusters was consistently returned. While it may make sense to the algorithm it does not make sense in a business context. The standard PAM algorithm was tested in the same way as K-Means, with multiple features and multiple k values. However, when looking at the distribution of clusters, it appeared to over cluster more than K-Means. See appendix H for code.

4.4.2 Density Based Algorithms

DBSCAN

While DBSCAN does not require the number of clusters, a value for epsilon and minPts must be provided. To calculate these values, the K-Nearest Neighbour (KNN) distance was plotted using the best *k* value obtained during that iteration of feature selection. The k value became the minPts value and the epsilon value was taken from the plot by finding where the elbow fell (60). Figure 21 provides an example of the KNN plot used to determine epsilon. See appendix I for code.

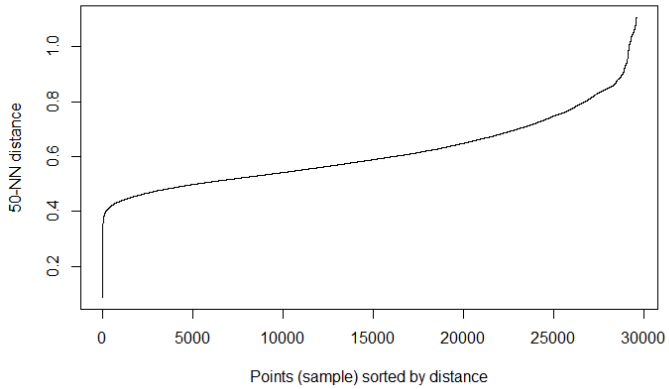


Figure 23: KNN plot used to find epsilon value

This algorithm did not prove useful in this context, it consistently either over clustered returning one large cluster and multiple small clusters, or under clustered returning only 2 large clusters. No combinations of parameters to resolve these issues could be identified.

Mean Shift

Like DBSCAN, the value of k is not required, however, a bandwidth value (*h*) is required. To calculate *h*, KNN distance is plotted using the best value for k found through previous clustering efforts; see appendix J. Too small a bandwidth results in over clustering (multiple single point clusters) while too large a bandwidth will result in large, coarse clusters (61). *h* values between 1.75 and 5 were tested, resulting in either one single large cluster or one large cluster and multiple single point clusters. It appears that density based algorithms are not appropriate for this data.

4.5 Evaluation

As it was not possible to have end users evaluate all results, a combination of different aspects have been used to evaluate clusters. The Calinski-Harabasz index evaluates cluster validity based on average, between and within cluster sum of squares (62). However this heuristic method did not always prove useful, the highest scores were returned for density based clustering algorithms, however, in a business

context these clusters were the least useful. Therefore, the number was considered along with how useful the distribution of clusters were and the top terms contained in the word cloud created for each cluster. Some word clouds clearly made sense with terms from similar contexts appearing, while other were not as clear with terms from varying sectors appearing. Clustering efforts which resulted in fewer 'broad' cluster were favoured over clustering efforts which returned multiple broad clusters.

For the sample of clusters which were evaluated by end users, an interview was created. The aim was to gain more insight into what a PM/SRO would be looking for in terms of knowledge sharing, to identify if the clusters presented made sense to them and finally to see how many projects they would like to be presented when deciding who to contact. A semi-structured approach was used as it has a clear structure, whilst having the flexibility to allow the interviewee to speak more about areas in which they have more experience (63).

Whether in the context of social science or user experience, the literature (63, 64) suggests an interview begins with some warm up questions to put the interviewee at ease. This should be a topic which the interviewee is confident and comfortable discussing. These initial questions were used to gain more knowledge about what being a PM/SRO entails and the interviewees experience of knowledge sharing in the organisation. After this, non-leading questions were asked to gauge how participants felt about the usefulness of the grouping of projects, the size of group presented and whether they felt the group would help them to contact other staff members to share knowledge. An edited[*1] version of the slides can be found in appendix K and the interview questions can be found in appendix L. The R code used to create the markdown presentation can be found in appendix M.

In conclusion, the CRISP-DM methodology was used for this analysis due to its iterative nature. Data were selected, cleaned and the tendency assessed before any algorithms were applied. Both partitional and density based algorithms were tested, with results consistently showing that partitional algorithms work more effectively on this data, in particular K-Means produced the most sensible looking results. Density based algorithms were not successful in this context which may be due to the sparse nature of the data involved. The iterative methodology was useful as the business context was considered throughout the project leading to the omission of data, which was either not available for projects in the design stage or when there was a desire to avoid creating clusters based on certain attributes.

**The most frequent risk terms have been removed from all slides as this information does not ever become public. Projects which are not yet live on dev tracker (66) have been omitted from the slides.*

5. Assessment of Clustering Results

The cluster analysis is only useful if it has value in a business context. Therefore, to assess the success of the clustering efforts, interviews were conducted with two appropriate staff within the organisation. Interview transcripts have not been included for security purposes, however the questions can be found in appendix 12. Decision fatigue is defined by the deterioration in quality of decisions made due to a prolonged session of decision making (67). To avoid decision fatigue, a sample of 10 clusters were selected. Ideally, they would have been selected at random, however, some clusters would be difficult to interpret in the limited time available, due to their size. Therefore, only clusters of 20 or fewer projects were considered for user assessment. Another consideration when selecting the sample of clusters was ensuring that both clusters and sub-clusters were represented. Therefore, 6 of the clusters created by K-Means (with less than 20 projects) were selected at random and the smallest 4 sub-clusters were also included.

To get a better understanding of the results without overwhelming interview participants, only one clustering approach was evaluated; clusters creating using Title, Purpose, DAC3, DAC5 and Risk Description data were assessed as this grouping best serves the business needs previously highlighted.

To gain insight into the roles (PM/SRO) and warm-up the interviewee, the initial questions asked about an average day in their job, if they had a sector of expertise (health, education, etc) and their experiences of knowledge sharing at DFID. These questions highlighted that identifying who to talk to is a challenge especially for PMs. Technical staff such as statisticians, are part of cadre which have regular events, making it possible for these staff to meet others who are working in similar sectors or on similar projects. PM/SROs don't have the same network which makes it more challenging for them to connect with other staff. There appeared to be a genuine interest in a tool which would aid in connecting PMs/SROs with others working on similar projects. Both participants explained that knowledge sharing is not as common as it should be and the knowledge sharing which does take place is in terms of compliance tasks and DFID procedures, not about problems faced by projects and how they were approached.

Of the 10 clusters presented, interviewees agreed that half of them appeared useful. The clusters agreed upon were related to malaria, corruption, energy, training and sanitation. However, it was highlighted that in two of the clusters a distinction in the type of project could be made. The malaria cluster contains some projects which are supporting government programmes and others which are working with the private sector to ensure good quality medications are available. While the sanitation cluster contains long term sanitation projects and humanitarian emergency response projects. Projects with a different focus will face different problems, even though they may be in same sector.

Interviewees gave conflicting results about two of the clusters. With two interviews conducted, it is difficult to make any inference about how useful these clusters are. One of the clusters which was not agreed upon was justice. Interestingly, each participant interpreted the use of this cluster in a

different way depending on their experience. The interviewee with experience in this area showed an interest in being able to see how contracts with firms managing the work on behalf of DFID were set up and what payment structures were in place to deal with the ever-changing environment in the region. The participant was aware that some of the projects were already in contact. Through the semi-structured approach to interviewing, questions were asked to explore how these projects had been connected. The connection was made through the implementers (external organisation such as an NGO) staff who had done similar work in another region. In contrast, the other participant stated that justice has a contextual difference and lessons learned in one country cannot be applied to another, even if they have a similar governance structure. The environment varies from country to country. This raises an interesting issue, different members of staff may have different interests in terms of knowledge sharing.

The other cluster which had conflicting results was the cluster related to market. The first participant stated that the group made sense, however, it was far smaller than expected. While the other, stated they weren't obviously related. When asked why they thought this, the interviewee explained that a distinction should be made between market development and market assistance as development is at an early stage of evolution, while assistance is generally to help correct markets which have failed.

Both participants were unsure of how useful three of the clusters were; all created using hierarchical clustering. The first, concerned with Aid Match made sense to all interviewees as they are similar projects, however, they are continuations of the same work undertaken in the same department. It was commented, "It's reassuring that they are together, but not necessarily useful". The other clusters which interviewees were unsure about were clusters containing a broad range of topics (see appendix K), created sub-clustering the largest K-Means cluster. Both participants could see some projects which would be useful in terms of knowledge sharing, however, there were many others in the group for which they could not understand the relation.

All participants agreed that the largest broad cluster (sub-cluster 4), also created during sub-clustering, did not make sense. They could not identify the relationship between any projects and would struggle to identify which programme to contact. This may be due to the cluster being created using only title, purpose and risk data. It was suggested that if programmes which are less similar in terms of project were being grouped, it would be useful to signify in some way what features were used. This would make it clear to staff that projects may not be similar in terms of sector but in terms of risk.

During the interviews, participants were presented groups of projects ranging in size from two to fourteen projects. Interestingly, participants consistently stated that they expected to see more projects in each cluster. In addition, it was suggested that if a tool like this were created, the user interface should allow for filtering based on project stage, with start and finish dates on display. This information will help PMs decide which project(s) will be most useful for them to contact. Another suggestion made by both interviewees was the inclusion of the implementer when the project list is presented. However, it was also noted that this could become complex as there are projects with multiple implementers.

This is a small sample of the clusters created and a very small sample of interview participants which have highlighted some conflicting view points on cluster results. More extensive testing with different PMs/SROs would be beneficial to build a more complete picture of how successful the clustering effort has been. Ideally, a selection of intended end users with different areas of expertise would be interviewed, to ensure that there was feedback from people with expert knowledge on each sector.

If a variety of PMs with different expertise could be interviewed, it would be useful to present them only with clusters from their field to gain more useful insight. Interviewees were not able to give detail of why they thought some cluster was useful (or not) beyond the titles seem similar/dissimilar. However, when they were speaking about projects from their sector they could distinguish between project types/content and use their expert knowledge to assess whether knowledge sharing would have been useful in the group. If interviewees were only presented with projects from their sector, this would help to evaluate clusters more effectively. It would also be beneficial to present clusters created using different techniques, such as the results of clustering without risk data or clusters created using other approaches to text analysis. This way it could be ascertained which approach is creating the most useful results.

In conclusion, a small sample of clusters were evaluated by a very small sample of appropriate staff. With such a small interview sample no conclusive results can be drawn, however as a preliminary test it shows that some clusters appear to be useful in a business context, while others were not as useful, or interviewees were unsure of the connection between projects in the group. It is suggested that more user interviews are conducted with many more PMs/SROs, presenting them only with clusters from their area of expertise to gain better insight into why the results are useful or not. It is also suggested that participants should be presented with clusters created using a variety of approaches to allow for comparison between different methods of text analysis and/or clustering. Ideally, each sector's cluster's should be evaluated by multiple staff from that sector with the order of clusters shuffled each time to combat the effects of decision fatigue.

6. Further Work

Due to the limited time available to conduct this research there is still work to be complete before a model could be deployed. Once a model is deployed, further evaluation will be required to fine tune how it is presented to end users. Below are some suggested steps for moving the project forward.

- Implement text analysis which takes word context into account
    - Techniques
        - Word2Vec
        - GloVe
    - Features

        - Title, Purpose, DAC3, DAC5
        - Title, Purpose, DAC3, DAC5, Risk Description
- Cluster the results using K-Means
- Conduct interviews with a variety of PMs/SROs with expertise in each sector
    - Obtain the interviewees area(s) of expertise prior to interviewing.
    - Ensure there are multiple staff with expertise in each sector interviewed (if possible). Baker and Edwards (68) state there is no correct number of interviews, however, researches must ensure there is sufficient data to answer the question. Two interviews was not enough as it lead to conflicting answers with no means of identifying consensus.
    - Present interviewee with clusters from their sector(s) of expertise only.
    - Present clusters made using varying approaches/features (as discussed) to obtain an insight into which approach produces the best results in a business context.
- Analyse the interview results
    - Create transcripts from audio recording of each interview.
    - Code (69) the responses as Useful, Not Useful or Unclear; symbols may be used for this if you so wish. Take note of any other discussion which may give further insight into the issue from the end users perspective
    - Decide which combination of attributes and text mining approach provide the best results according to expert interview results
- Put the best model into production
    - Conduct A/B testing to assess the optimal amount of projects to present staff with; bounce rate, session time and click through rate can be monitored to determine the optimal number of projects to present.

7. Conclusion

The purpose of this project was to examine approaches to text analysis, clustering and assessing clustering results in a business context. The initial research explored TFIDF, Word2Vec, Doc2Vec and GloVe. Unfortunately, there is no R implementation of Doc2Vec and due to the limited time available only TFIDF was implemented.

In addition to studying approaches to text analysis, clustering techniques were researched. There was particular interest in discovering the best way to find the number of clusters for use in partitional algorithms, and to discover algorithms which did not require *k* a priori. This research highlighted that within partitional clustering there are many options which require *k* a priori including, K-Means, K-Medoids, PAM, CLARA. The value of *k* can be estimated using the Elbow or Silhouette method, however, in implementation these ready- made functions, the *k* values are restricted to a range

from 2 to 10. Therefore, code was written which calculate the within sum of squares for k values between 2 and 200. There are partitional approaches which do not require k a priori including PamK and X-Means. In practice the results returned were not useful in a business context. Density based algorithms such as DBSCAN and Mean Shift were researched and tested. Unfortunately, they did not prove useful for this data which may be due to its sparse nature.

While mathematic procedures to assess clusters were researched, clustering is only useful if it solves the business problem at hand, to help PMs/SROs contact staff working on similar projects with the aim of facilitating knowledge sharing. There is a gap in the literature in the area of cluster assessment when ground truth is unknown, with reference made to expert knowledge but no detail on how this analysis is conducted. Therefore, social science and human computer interaction literature was examined to create non-leading interview questions. Two appropriate members of staff were presented with a small sample of clustering results from one approach only (K-Means clustering of Title, Project, DAC3, DAC5 and Risk Description data) and interviewed to assess the success the effort. Half of the clusters made sense to participants and they agreed that they could use the information to contact the appropriate staff, however, there were conflicting results for some of the clusters, while for others it was clear that both participants could not see a connection between the projects involved. The interview sample was too small to make any real assessment of the reliability of the overall result, and since only a small sample of clusters were presented a definitive answer cannot be given. However, it is does show some positive results, providing a basis on which to continue working.

Going forward, it is suggested that other forms of text analysis are implemented and the results clustered using K-Means. Rather than presenting a varied sample of clusters to interviewees, they should be presented with clusters from their area(s) of expertise only. This will ensure that all feedback received is meaningful and goes beyond a visual assessment of the project names, which arguably any person could do. In addition, it would be a good opportunity to present the results of varying text analysis approaches/features to assess which are deemed most useful by the staff who the product is intended for. With further work, this technology has the potential to improve knowledge sharing between PMs and SROs within DFID in a project context.

Appendix A - R Script Creating Data

```r
#---------------LIBRARIES---------------
library(tidytext)
library(dplyr)
library(tidyr)
library(plyr)

#---------------READ IN RAW DATA---------------

# SET WORKING DIRECTORY TO THIS FILE #

#Project ID, Title and Purpose
df_itp <- read.csv("ID_Title_Purpose.csv", stringsAsFactors = F) #16142

#ID + DAC3 Info
df_dac3 <- read.csv("ID_DAC3.csv", stringsAsFactors = F)
#Remove rows which are (998) Unallocated/unspecified
df_dac3 <- subset(df_dac3, df_dac3$DAC.3.Digit.Sector != "(998) Unallocated/unspecified")


#ID + DAC5 Info
df_dac5 <- read.csv("ID_DAC5.csv", stringsAsFactors = F)
# Remove rows which are WARN: DAC Sector Undefined
df_dac5 <- subset(df_dac5, df_dac5$DAC.5.Digit.Sector != "WARN: DAC Sector Undefined")

#ID + Income Group
df_ig <- read.csv("income_group.csv", stringsAsFactors = F)
#
# #ID + SID Continent
df_cont <- read.csv("SID_Cont.csv", stringsAsFactors = F)
#
# #ID + Benifitting Country
df_country <- read.csv("Country.csv", stringsAsFactors = F)

#ID + Fragile State/Region
df_f <- read.csv("Fragile_State_Region.csv", stringsAsFactors = F)

# Run getRiskData.R file to import the most up to date risk data
# File created by DFID staff
df_risk <- riskData
rm(riskData)
df_risk <- df_risk[,1:2]
colnames(df_risk)[1] <- "Project.ID"

# Get department office data ADDED TOWARDS END OF PROJECT
dept <- read.csv("Project_Office.csv", stringsAsFactors = F)

#ID + Stage
stage <- read.csv("stage.csv", stringsAsFactors = F)

#---------------CONCAT ALL PROJECT INFO BY PROJECT ID---------------
#Concat DAC3 Info
df_dac3 <- ddply(df_dac3, .(Project.ID), summarize, DAC3_list = paste(DAC.3.Digit.Sector, collapse=" "))

#Concat DAC5
df_dac5 <- ddply(df_dac5, .(Project.ID), summarize, DAC5_list = paste(DAC.5.Digit.Sector, collapse=" "))

#Concat Income Group
```

```r
df_ig <- ddply(df_ig, .(Project.ID), summarize, Income_Group = paste(Income.Group, collapse=" "))

#Concat SID Continent
df_cont <- ddply(df_cont, .(Project.ID), summarize, SID_Continent = paste(SID.Continent, collapse=" "))

#Concat Countries
df_country <- ddply(df_country, .(Project.ID), summarize, Benefitting_Country = paste(Benefitting.Country, collapse=" "))

#Replace N/A [16] with Not fragile state or region ##See N/A Guide##
df_f$Fragile.State.or.Region[df_f$Fragile.State.or.Region == "N/A [16]"] <- "Not fragile"
#Concat Fragility
df_f <- ddply(df_f, .(Project.ID), summarize, Fragile_State_Region = paste(Fragile.State.or.Region, collapse=" "))

#Concat Risk
df_risk <- ddply(df_risk, .(Project.ID), summarize, Risk_list = paste(RiskDescription, collapse=" "))

#Concat department data
dept <- ddply(dept, .(Project.ID), summarize, dept_list = paste(Dept...Office, collapse=" "))

#--------------CREATE ONE DATAFRAME WITH ALL DATA--------------
projectData <- merge(df_itp, df_dac3, by="Project.ID")
projectData <- merge(projectData, df_dac5, by="Project.ID")
projectData <- merge(projectData, df_ig, by="Project.ID")
projectData <- merge(projectData, df_cont, by="Project.ID")
projectData <- merge(projectData, df_country, by="Project.ID")
projectData <- merge(projectData, df_f, by="Project.ID")
projectData <- merge(projectData, df_risk, by="Project.ID")
projectData <- merge(projectData, stage, by="Project.ID")

#--------------REMOVE ADMIN PROJECTS--------------
#Admin project spend codes
admin <- read.csv("Admin_Projects_Spend.csv", stringsAsFactors = F)

#Funding type codes (admin resource, admin captial and WARN: Undefined)
admin2 <- read.csv("Admin_Funding.csv", stringsAsFactors = F)

#Remove admin spend projects from project data
projectData <- projectData[!(projectData$Project.ID %in% admin$Project.ID),]

#Remove admin funding codes from project data
projectData <- projectData[!(projectData$Project.ID %in% admin2$Project.ID),]

      # Remove admin projects from department data ADDED LATER
      dept <- dept[!(dept$Project.ID %in% admin$Project.ID),]
      dept <- dept[!(dept$Project.ID %in% admin2$Project.ID),]

      #--------------SAVE COMPLETE DATA FILE--------------
      write.csv(projectData, file="PROJECT_DATA_RISK_ADDED.csv", row.names = FALSE)

      write.csv(dept, file="department_data.csv", row.names = FALSE)
```

Appendix B - Cleaning Text and Creating a Document Corpus

```
#---------------ANALYISIS PIPELINE FUNCTIONS---------------
#Preprocessing function
preprocess <- function(x){
  x <- replace_contraction(x) #replace common contractions
  x <- replace_abbreviation(x) #replace common abbreviations
  x <- tokens(x, what="word",
              remove_numbers = TRUE, remove_punct = TRUE, #rmv numbs only removes strings made entirly of numbers
              # remove_symbols = TRUE, remove_hyphens = TRUE, #from quanteda package
              remove_url = TRUE) #tokenise
  x <- tokens_tolower(x) #lowercase
  x <- tokens_select(x, stp_wds$stopwords, selection = "remove") #stopword removal
  x <- tokens_wordstem(x, language = "english") #stemming
}

#concat all text fields
project_text <- paste(projectData$Project.Title, projectData$Purpose, projectData$Risk_list, projectData$DAC5_list, projectData$DAC3_list)

# clean data
clean_txt <- preprocess(project_text)
# add to dataframe
projectData$text <- clean_txt

# Find the 10 most frequent terms: term_count
term_count <- freq_terms(clean_txt, 30)

# Plot term_count
plot(term_count)

#create Corpus
text_corpus <- VCorpus(VectorSource(projectData[,7]))

#remove numbers, special characters and punctuation missed by quanteda
text_corpus <- tm_map(text_corpus, removeNumbers) # removal of numbers
toSpace <- content_transformer(function(x, pattern) gsub(pattern, " ", x))
text_corpus <- tm_map(text_corpus, toSpace, "/|@|//|$|:|:)|*|&|!|?|_|-|#|")  ## replace special characters by space
text_corpus <- tm_map(text_corpus,removePunctuation)

# create DTM - default weighting is term frequency (tf)
dtm_tf <- DocumentTermMatrix(text_corpus)
row.names(dtm_tf) <- projectData$Project.ID

# Plot most frequent terms in barchart
freq <- sort(colSums(as.matrix(dtm_tf)), decreasing=TRUE)
wf <- data.frame(word=names(freq), freq=freq)
# Plot Histogram
subset(wf, freq>250)    %>%
  ggplot(aes(word, freq)) +
  geom_bar(stat="identity", fill="darkred", colour="darkgreen") +
  theme(axis.text.x=element_text(angle=45, hjust=1))

# Plot most frequent terms in wordcloud
set.seed(100)
wordcloud(names(freq), freq, min.freq=100, colors=brewer.pal(6, "Dark2"))

#remove terms with are more than 40% sparse
removeSparseTerms(dtm_tf, 0.4)

# create DTM with weighting term frequency inverse document frequency (tf-idf)
dtm_tfidf <- DocumentTermMatrix(text_corpus, control = list(weighting = weightTfIdf))
row.names(dtm_tfidf) <- projectData$Project.ID

# Convert to matrix
#required for PCA, assessing tendency and heuristic methods of finding K
m <- as.matrix(dtm_tfidf)
```

Appendix C – Stop Words

a	ah	are	balts	bi
aa	ai	area	bamar	big
ab	aid	areas	bangladesh	bissau
abilities	aided	argentine	Bangladeshis	bj
ability	aiding	argentine	barbadian	bk
able	aim	argentinean	barbadians	bl
about	aimed	argentineans	barbados	bn
above	aims	argentines	barbuda	bo
access	al	argentinian	barbudan	bolivia
accessibility	albania	argentinians	barbudans	bolivian
accessible	albanian	armenia	barbudian	bolivians
account	albanians	armenian	barbudians	bosnia
accountable	algeria	armenians	barundi	bosnian
accounted	algerian	around	basic	bosnians
accounting	algerians	as	basics	bosotho
accounts	all	as	bb	both
achieve	almost	ascension	bd	botswana
achieved	alone	asia	be	botswanan
achieves	along	asian	became	botswanans
achieving	already	asiapoverty	because	br
acitvities	also	ask	become	brazil
across	although	asked	becomes	brazilian
across	always	asking	been	brazilians
active	am	asks	before	british
active	america	at	began	brumese
actively	american	away	behind	brunei
actively	americans	ay	being	bruneian
activeness	among	az	beings	bs
activity	an	azerbaijan	belarus	bt
activity	and	azerbaijani	belarusian	bu
ae	angola	azerbaijanis	belarusians	budget
af	angolan	b	belize	budgeted
affect	angolans	ba	belizian	budgeting
affected	anguilla	back	belizians	budgets
affects	anguillan	backed	benefitting	bulgaria
afghan	anguillans	backing	bengali	burkina
afghanistan	another	backs	benin	burkinabe
afghans	antigua	bahamas	beninese	burma
africa	any	bahamian	beninois	burundi
africa	anybody	bahamians	best	burundian
african	anyone	bahrain	better	burundians
african	anything	bahraini	between	business
africans	anywhere	bahrainis	beyond	businesses
after	ao	balkan	bg	but
ag	ap	balkans	bh	bw
again	ar	baltic	bhutan	by
against	arab	baltics	bhutanese	bz

c	co	cunha	downed	engaged
caicos	colombia	cv	downing	engagement
cambodia	colombian	cy	downs	engages
came	colombians	cypriot	due	engaging
cameroon	come	cypriots	dues	enough
cameroonian	comoran	cyprus	during	ensure
cameroonians	comorans	cz	dz	ensured
can	comorian	czech	e	ensures
cannot	comorians	czechs	ea	ensuring
capacities	comoros	d	each	er
capcity	congo	da	early	eritrea
cape	congolese	data	east	eritrean
caribbean	context	deliver	eb	eritreans
caribbeans	contexts	delivered	ec	estonia
case	contextual	delivering	ecuador	estonian
cases	contextualise	delivers	ecuadorian	estonians
cayman	contextualize	delivery	ecuadorians	et
caymanian	contextually	department	ed	ethiopia
caymanians	contract	departments	ee	ethiopian
cb	contracted	develop	eecad	ethiopians
cd	contracts	developed	ef	europe
ce	coordinate	developing	effect	european
central	coordinated	development	effected	europeans
certain	coordinates	develops	effecting	even
certainly	coordinating	dfid	effective	evenly
cf	cost	dfida	effectively	ever
cg	costa	dfids	effects	every
chad	costs	did	eg	everybody
chadian	cote	differ	egypt	everyone
chadians	could	different	egyptian	everything
chagnes	countires	differently	egyptians	everywhere
change	countries	dj	ei	f
changed	country	djibouti	either	fa
changing	country	djiboutian	el	face
chile	country	djiboutians	emirates	faces
chilean	cp	dm	emirati	fact
chileans	cq	do	emiri	facts
china	cr	do	emirian	far
chinese	croatia	does	enable	faso
ci	croatian	dominica	enabled	federation
cj	croatians	dominican	enables	federations
cl	croats	dominicans	end	felt
clear	cu	done	ended	few
clearly	cuba	donor	ending	fiji
cm	cuban	donors	ends	fijian
cn	cubans	down	engage	fijians

finance	gibraltar	guineans	id	ir
financial	gibraltarians	guineans	if	iran
find	give	guyana	il	iranian
finds	given	guyanese	impact	iranians
first	gives	gw	impact	iraq
fj	gm	gy	impacted	iraqi
focus	gn	h	impacted	iraqis
focused	go	had	impacting	is
for	goal	haiti	impacting	island
four	goals	haitian	impacts	islanders
francophone	going	haitians	impacts	islands
from	good	has	implement	israel
full	goods	have	implementation	israeli
fully	got	having	implemented	israelis
fund	govern	he	implements	issue
funded	governability	helena	important	issued
funding	governable	helenian	improve	issues
funds	governed	helenians	improvement	issuing
further	governing	hellenic	improves	it
furthered	government	hellenics	improving	its
furthering	governs	her	in	itself
furthers	gr	here	in	ivoire
g	grant	herself	include	ivorian
gambia	granted	high	including	ivorians
gambian	granting	high	income	j
gambians	grants	high	increase	jamaica
gave	great	higher	increased	jamaican
gaza	greater	highest	increasing	jamaicans
gb	greatest	him	india	japan
gd	greece	himself	indian	jm
ge	greek	his	indians	jo
general	greeks	hn	indonesia	jordan
general	grenada	honduran	indonesian	jordanian
generally	grenadian	hondurans	indonesians	jordanians
generally	grenadians	honduras	intend	jp
generlise	group	how	intended	just
georgia	grouped	however	intends	k
georgian	grouping	hr	intention	karzakh
georgians	groups	ht	interest	karzakhs
get	gt	hu	interested	kazakhstan
gets	guatemala	hungarian	interesting	kazakhstani
gh	guatemalan	hungarians	interests	kazakhstanis
ghana	guatemalans	hungary	international	ke
ghanaian	guinea	hv	intervention	keep
ghanaians	guinea	i	into	keeps
gi	guinean	ib	iq	kenya

kenyan	later	low	mauritanians	most
kenyans	latest	lowed	mauritian	mostly
key	latin	lower	mauritians	motwsana
keys	latvia	lowered	mauritius	mozambican
kg	latvian	lowering	may	mozambicans
ki	latvians	lowest	mb	mozambique
kind	lb	lowing	md	mr
kingdom	lc	lows	me	mr
kirghiz	le	lr	me	mrs
kirgiz	lead	ls	mean	ms
kiribati	leader	lt	meaning	mt
kittian	leaders	lucia	means	mu
kittians	leadership	lucian	meeting	much
kitts	leading	lucians	meets	multilateral
km	leads	lv	meets	must
kn	least	ly	member	mv
knew	lebanese	m	members	mw
know	lebanon	ma	men	mx
known	leone	macedonia	mexican	my
knows	leonean	macedonian	mexicans	my
ko	leoneans	macedonians	mexico	myself
korea	lesotho	madagascar	mg	mz
korean	less	made	middle	n
koreans	let	magyar	might	na
kosovo	lets	magyars	million	namibia
kp	level	make	millions	namibian
ky	levelled	making	mk	namibians
kyrgyz	levelling	malagasy	ml	nauru
kyrgyzstan	levels	malawi	mn	nauruan
kyrgyzstani	liberia	malawian	moldova	nauruans
kyrgyzstanis	liberian	malawians	moldovan	nb
kz	liberians	malaysia	moldovans	ne
l	libyan	malaysian	mongolia	necessary
la	libyans	malaysians	mongolian	need
lack	like	maldives	mongolians	needed
lacked	likely	maldivian	montenegrin	needing
lacking	lithuania	maldivians	montenegrins	needs
lanka	lithuanian	mali	montenegro	nepal
lankan	lithuanians	malta	montserrat	nepalese
lankans	lk	maltese	montserratian	nepali
lao	long	man	montserratians	never
laos	longer	manage	more	nevis
laotian	longest	managed	moroccan	nevisian
large	loss	management	moroccans	nevisians
largely	lossed	many	morocco	new
last	loosing	mauritania	mosotho	newer

newest	once	parted	presented	rather
next	one	parting	presenting	really
ng	only	partner	presents	region
ni	open	partnered	prevent	regional
nicaragua	opened	partners	prevented	regions
nicaraguan	opening	partnership	preventing	relief
nicaraguans	opens	parts	prevents	report
niger	or	pe	problem	reported
nigeria	or	people	problems	reporting
nigerian	order	peoples	process	reports
nigerians	ordered	per	processed	republic
nigerien	ordering	perhaps	processes	require
nigeriens	orders	persian	procure	required
no	organisation	persians	procured	requirement
nobody	organisations	pertnering	procurement	requires
non	ot	peru	program	requiring
noone	other	peruvian	programed	resource
north	others	peruvians	programing	resource
not	our	pg	programm	resourced
nothing	out	ph	programme	resourced
now	outcome	philippines	programmed	resources
nowhere	outcomes	pitcairn	programmes	resources
np	over	pk	programming	resourcing
nr	overall	pl	programs	resourcing
ns	overseas	place	progress	result
number	p	places	progressed	results
numbers	pa	plan	progressing	rica
nz	pacific	planned	progression	rican
o	pakistan	planning	project	ricans
object	pakistani	plans	project	right
objective	pakistanis	pn	projected	right
objectively	palestine	point	projects	risk
objectiveness	palestinian	pointed	provide	risk
objectives	palestinians	pointing	provided	risked
oceania	panama	points	provides	risking
of	panamaian	poland	providing	risks
off	panamaians	poles	purpose	ro
offer	papua	policied	purposes	romania
offered	papuan	policies	put	romanian
offering	papuans	policy	puts	romanians
offers	paraguay	policying	py	room
often	paraguayan	polish	q	rooms
old	paraguayans	portfolio	qualities	ru
older	parliament	portfolios	quality	russia
oldest	parliements	possible	quite	russian
on	part	present	r	russians

rw	services	somebody	system	things
rwanda	servicing	someone	systems	think
rwandan	servicing	something	sz	thinks
rwandans	several	somewhere	t	this
s	seychelles	south	ta	those
sahel	seychellois	southern	tajiks	though
said	seychelloises	soviet	tajikstan	thought
saint	sf	sp	tajikstani	thoughts
salvador	sg	sq	tajikstanis	three
salvadoran	sh	sri	take	through
salvadorans	shall	ss	taken	thus
same	she	staff	tanzania	timor
samoa	should	staffed	tanzanian	timorese
samoa	show	staffing	tanzanians	tj
samoan	showed	stakeholder	target	tm
samoan	showing	stakeholders	targeted	tn
samoans	shows	state	targeting	to
samoans	si	states	targets	to
saw	side	states	targetted	tobago
say	sides	still	targetting	tobagonian
says	sierra	still	tc	tobagonians
sb	since	stopwords	td	today
sc	singapore	strategies	team	together
sd	singaporean	strategy	teamed	togo
second	singaporeans	su	teams	togolese
seconds	sk	sub	term	tonga
sector	sl	such	termed	tongan
sector	slovak	sudan	terming	tongans
sectors	slovakia	sudanese	terms	too
sectors	slovaks	support	territories	took
see	slovene	supported	territory	toward
seem	slovenes	supporting	tg	tp
seemed	slovenia	supportive	th	tr
seeming	slovenian	supports	thai	trinidad
seems	slovenians	suppport	thailand	trinidadian
sees	small	sure	than	trinidadians
senegal	smaller	sv	that	tristan
serbia	smallest	sw	the	tt
serbian	sn	swazi	their	tunisia
serbians	so	swaziland	them	tunisian
serbs	so	swazis	then	tunisians
service	solomon	sx	there	turkey
service	somali	sy	therefore	turkish
serviced	somalia	syria	these	turkmen
serviced	somalis	syrian	they	turkmenistan
services	some	syrians	thing	turkmens

turks	until	vg	which	yemeni
turks	up	vietnam	while	yemenis
turn	upon	vietnamese	who	yet
turned	uruguay	vincent	whole	you
turning	uruguayan	vincentian	whose	young
turns	uruguayans	vincentians	why	younger
tuvalu	us	vn	wi	youngest
tuvaluan	us	vu	will	your
tuvaluans	use	w	windward	yours
tv	use	want	with	yu
two	used	wanted	within	yugoslavia
tz	used	wanting	without	yugoslavian
u	uses	wants	work	yugoslavians
ua	using	warn	work	z
ug	uy	was	worked	za
uganda	uz	way	working	zambia
ugandan	uzbek	ways	works	zambian
ugandans	uzbekistan	wb	would	zambians
ukraine	uzbekistani	we	ws	zealand
ukrainian	uzbekistanis	well	x	zealanders
ukrainians	uzbeks	wells	xa	zelanian
undefined	v	went	y	zelanians
under	vanuatu	were	ye	zimbabwe
union	vanuatuan	west	year	zimbabwean
unions	vanuatuans	what	year	zimbabweans
united	vc	when	years	zm
unless	verde	where	years	zw
unspecified	very	whether	yemen	zz

Appendix D – Assessing Clustering Tendency

```
#Visual assessment of clustering tendency, using dissimilarity matrix
fviz_dist(dist(comp), show_labels = FALSE)+
  labs(title = "Title, Purpose, Risk Description: Demonyms removed")
```

Appendix E – Principal Component Analysis

```
#---------------PRINCIPAL COMPONENT ANALYSIS (PCA)---------------#

mprcomp <- prcomp(m, scale = FALSE)

mprcomp$rotation[1:5,1:5]
dim(mprcomp$x)

#calculate standard deviation
std_dev <- mprcomp$sdev

#calculate variance
pr_var <- std_dev^2

#calculate proportion of explained variance
prop_var_ex <- pr_var/sum(pr_var)

plot(prop_var_ex, xlab="Principal Component", ylab="Proportion of Variance Explained", type = "b")

#Cumulative plot
plot(cumsum(prop_var_ex), xlab = "Prinicipal Component", ylab = "Cumulative Proportion of Variance Explained",
     type = "b")

comp <- data.frame(mprcomp$x[,1:1000])
```

Appendix F – Finding *k*

```
#---------------ESTIMATE NUMBER OF CLUSTERS FOR KMEANS---------------
#Elbow Method
fviz_nbclust(comp, kmeans, method = "wss") +
  geom_vline(xintercept = 6, linetype = 2)+
  labs(subtitle = "Elbow method")

# Silhouette method
fviz_nbclust(comp, kmeans, method = "silhouette")+
  labs(subtitle = "Silhouette method")

# Within Sum of Squares
wss <- (nrow(comp)-1)*sum(apply(comp,2,var))

# Number of clusters
try_clusters <- seq(from = 2, to = 200, by = 1)

# Calculate the number of cores
no_cores <- detectCores() - 1

# Initiate cluster
cl <- makeCluster(no_cores)
registerDoParallel(cl)

# Calculating WSS for each cluster size in parallel
wss_results <- foreach::foreach (i = try_clusters) %dopar% {
  sum(kmeans(comp, centers=i, iter.max = 1000)$withinss)
}

# Stop Clusters
stopCluster(cl)

# Plot WSS results
plot(try_clusters, unlist(wss_results), type="b", main = "Title, Purpose, Risk: Demonymns added", xlab="Number of Clusters",
     ylab="Within groups sum of squares")
```

Appendix G – K-Means

```
#---------------KMEANS---------------
# Calculate the number of cores
no_cores <- detectCores() - 1

# Initiate cluster
cl <- makeCluster(no_cores)

registerDoParallel(cl)

#set seed so results are reproducable
set.seed(42)

results <- foreach(i=c(25,25,25)) %do% {
  kmeans(comp, 100, iter.max = 1000, nstart = i)
}
temp_vector <- sapply(results,function(result) {result$tot.withinss})
result <- results[[which.min(temp_vector)]]

# Stop Clusters
stopCluster(cl)

#unlist clusters - changes to integer for ch assessment
result_int <- unlist(result$cluster)

#Calinski-Harabasz index
ch <- calinhara(comp, result$cluster, 100) #cn = k used

#Check the number of projects in each cluster
size <- table(result$cluster)
```

Appendix H – PAM

```
#---------------PAM (Partitioning around mediods)---------------
#PamK
pamk_clst <- pamk(comp, krange = 2:10, criterion = "asw", usepam = TRUE,
                  scaling = FALSE, alpha = 1, diss = FALSE, critout = FALSE)

#PAM
pam_clst  <- pam(comp, k = 100, diss = FALSE, metric = "euclidean", stand = FALSE,
                 cluster.only = FALSE, do.swap = FALSE, keep.diss = TRUE,
                 keep.data = TRUE, pamonce = 0)


ch <- calinhara(comp,pam_clst$clustering ,cn = 100) #cn = use k
```

Appendix I – DBSCAN

```
#---------------DBSCAN---------------
#find epsilon using KNN
kNNdist(comp, k=100, search="kd")
kNNdistplot(comp, k=100)

cluster_dbs <- dbscan::dbscan(comp,eps = 0.9, minPts = 100)

nclust <- table(cluster_dbs$cluster)

ch <- calinhara(comp,cluster_dbs$cluster,cn = 2) #cn = use resulting number of clusters
```

Appendix J – Mean Shift

```
#---------------MEAN SHIFT---------------
# Use KNN to calculate h
kNNdist(comp, k=100, search="kd")
kNNdistplot(comp, k=100)

#Run meanshift algorithm using 3 cores
options(mc.cores=3)
clustering_ms <- msClustering(comp, h=0.9, kernel = "gaussianKernel",
                              tol.stop = 1e-06, tol.epsilon = 0.001,
                              multi.core=TRUE)

nclust <- table(clustering_ms$labels)

ch <- calinhara(comp,clustering_ms$labels,cn = 16) #cn = use resulting number of clusters
```

Appendix K – Edited R Markdown Slides for User Testing

Clustering DFID Aid Data

Suzie Beith

8 August 2018

Introduction

DFID manages many projects from varying sectors across the world.

How does DFID share knowledge and learning across such a complex organisation?

Can we improve the way knowledge is shared to facilitate learning from each other?

The aim of this project is to create a groups of projects ("clusters") that could help project managers to more effectively identify staff working on similar projects or facing similar challenges.

During this chat I will:

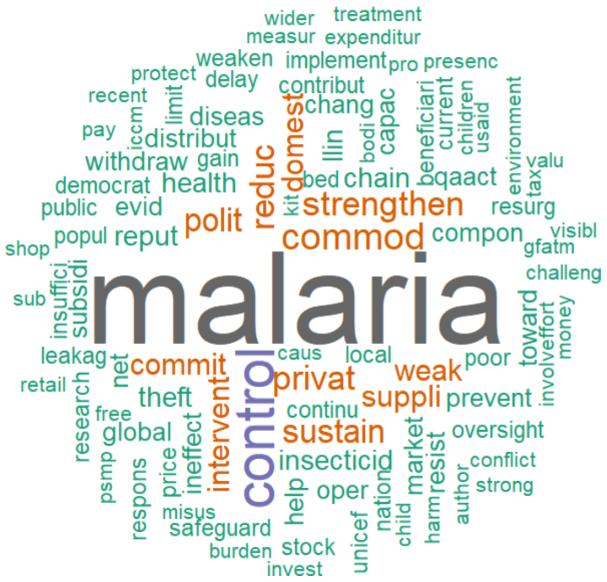
- Ask some questions about your role
- Show you a sample of the groups created
- Ask for your feedback on these groups

Sample Data

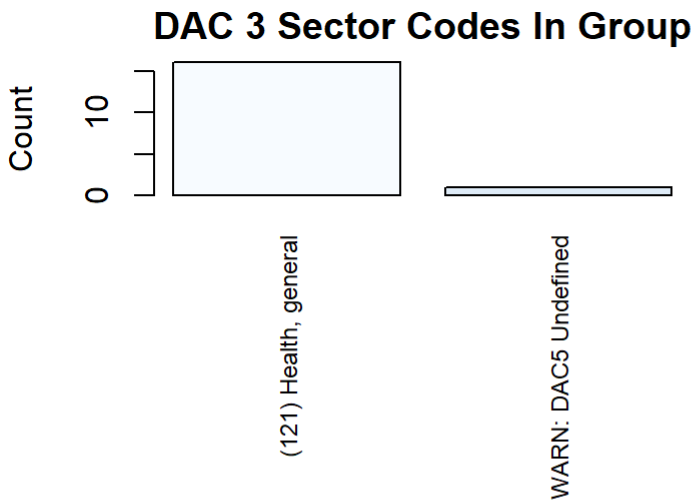
To create each group the following project attributes were used:

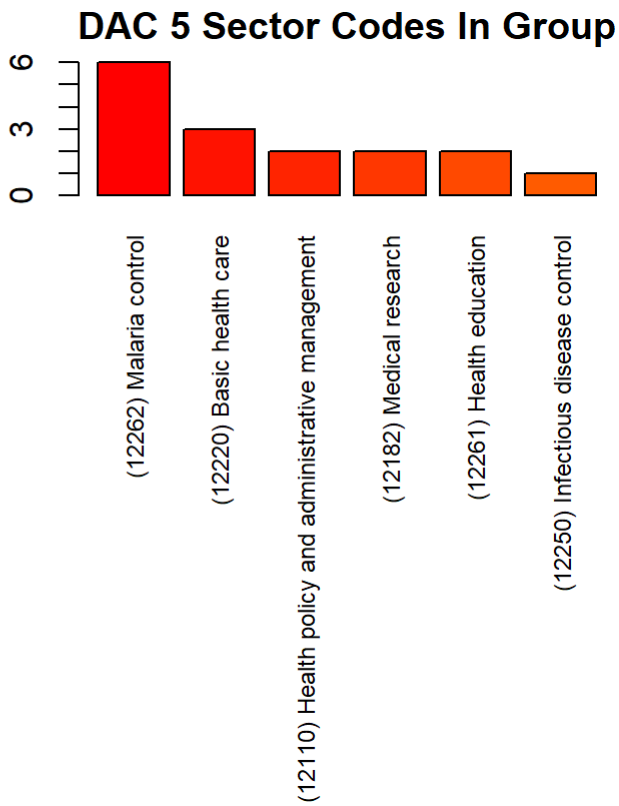
- Title
- Purpose
- DAC 3 sector code text
- DAC 5 sector code text
- Risk Description

Malaria Control



DAC 3 and 5 Sectors in Group





Projects in group

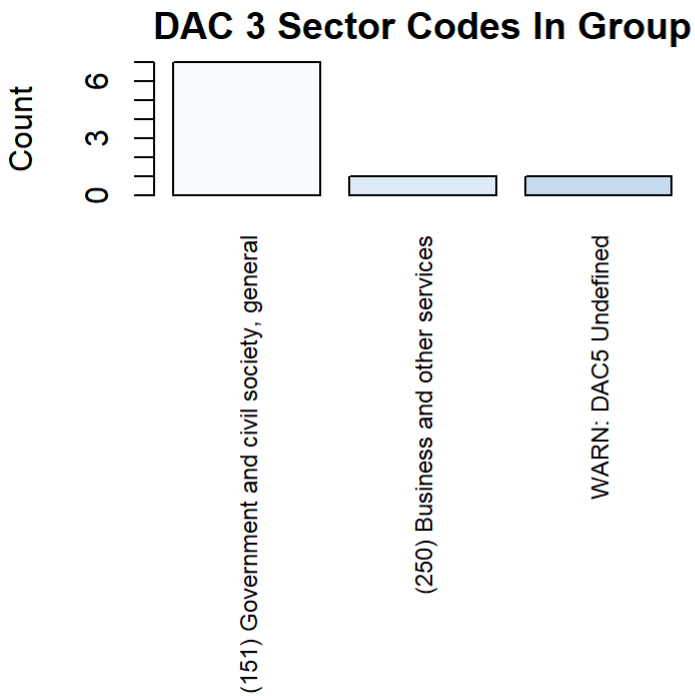
3 projects have been omitted for security purposes

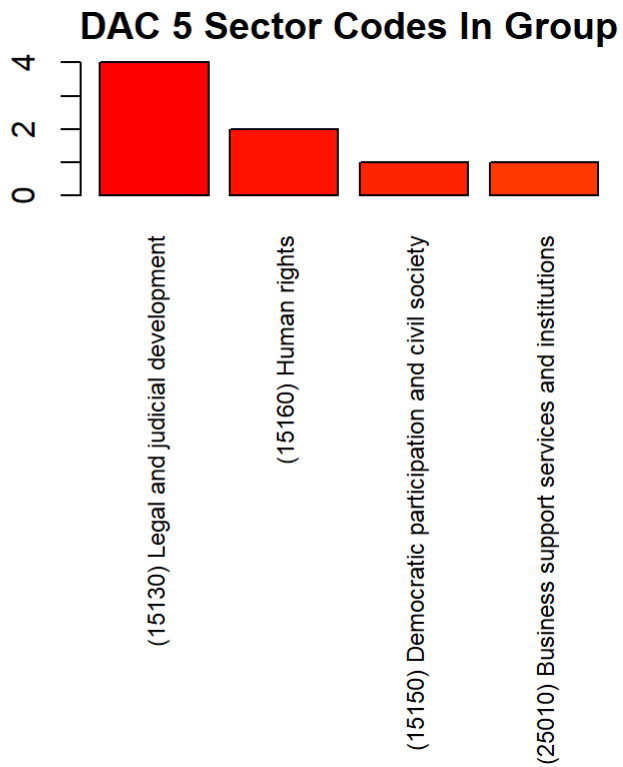
ID	Title	Department
202979	SUPPORT TO NATIONAL MALARIA PROGRAMME PHASE 2	DFID Nigeria (1018)
203155	Strengthening the use of data for malaria decision making in Africa.	Africa Regional Department (1001)
203458	Support to Malaria Control in the Democratic Republic of Congo	DFID DRC (1006)

Justice

justic
novemb
detaine
practic
independ
structur
misdirect
capac
localis
politicis
oper
contribut
abus
inform
time
perform
societi
captur
fraud
flexibl
sign
strengthen
benefit
violenc
poor
rule
actor
reduc
execut
cyclon
judiciari
money
led
citizen
fair
fail
reform
court
hmg
occur
ngos
corrupt
valu
militari
equal
miss
polic
alloc
commerci
civil
local
women
effect
human
etc
pre
children
popul
implic
elect
law
polit
prison
institut
constrain
misus
district
drug
jvg
manag
lack
undermin
joint
vulner
paraleg
recoveri
coordin
downstream
dai
formal
author
rol
communiti
giz
sexual
unabl
fluctuat
model
current
especi
fragment
offend
referr
unwil
user
salari
avoid
momentum
terror
vision

DAC 3 and 5 Sectors in Group





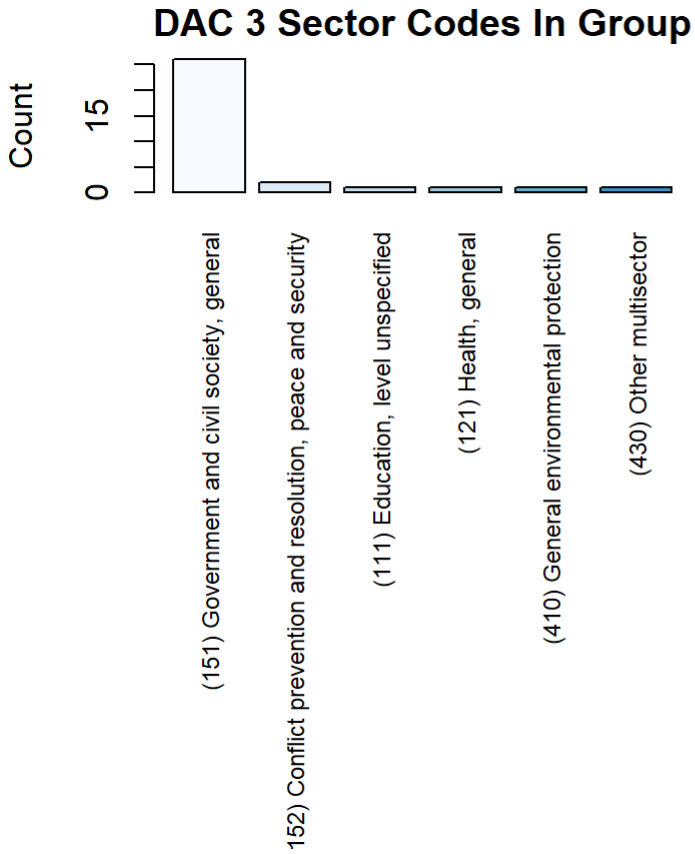
Projects in group

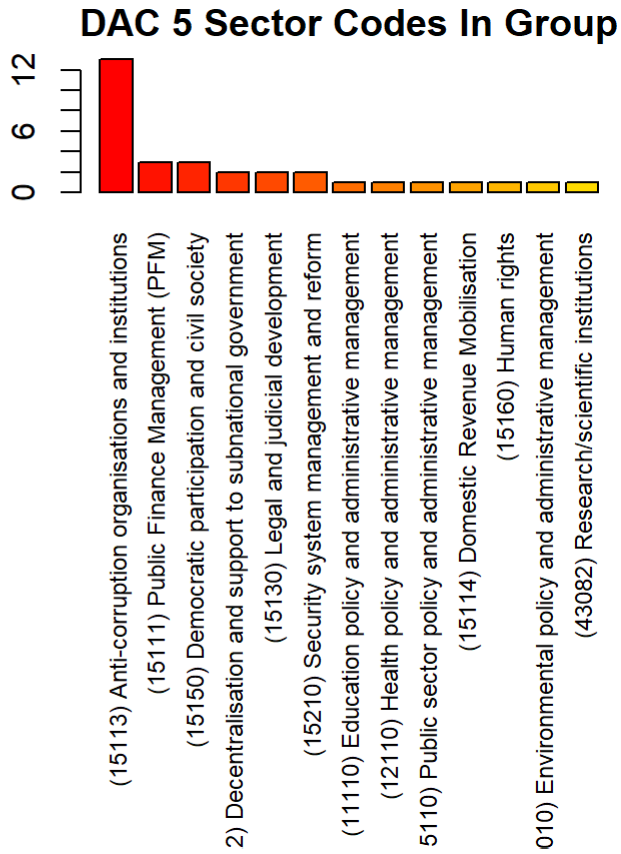
2 projects have been omitted for security purposes

ID	Title	Department
203227	Access to Justice through Paralegal and Restorative Justice Services in Bangladesh	DFID Bangladesh (1054)
204619	Strengthening Rule of Law in Pakistan	DFID Pakistan (1059)

Corruption and Government

DAC 3 and 5 Sectors in Group



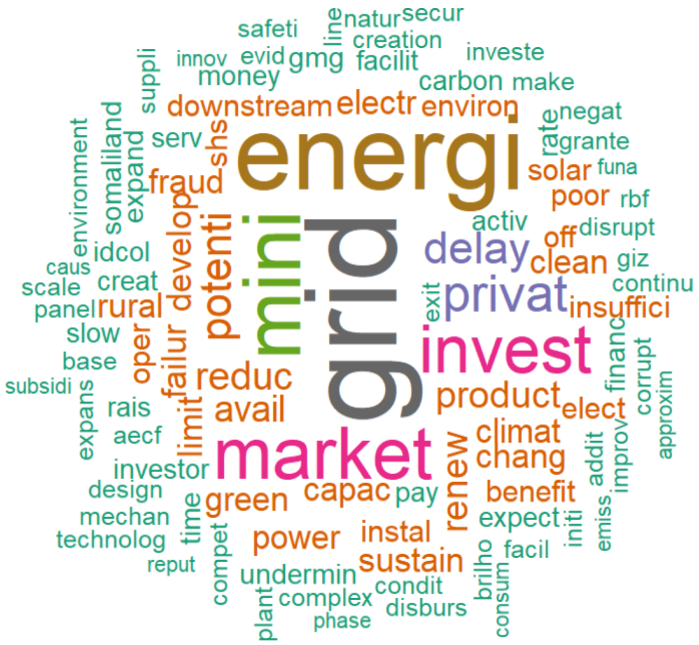


Projects in group

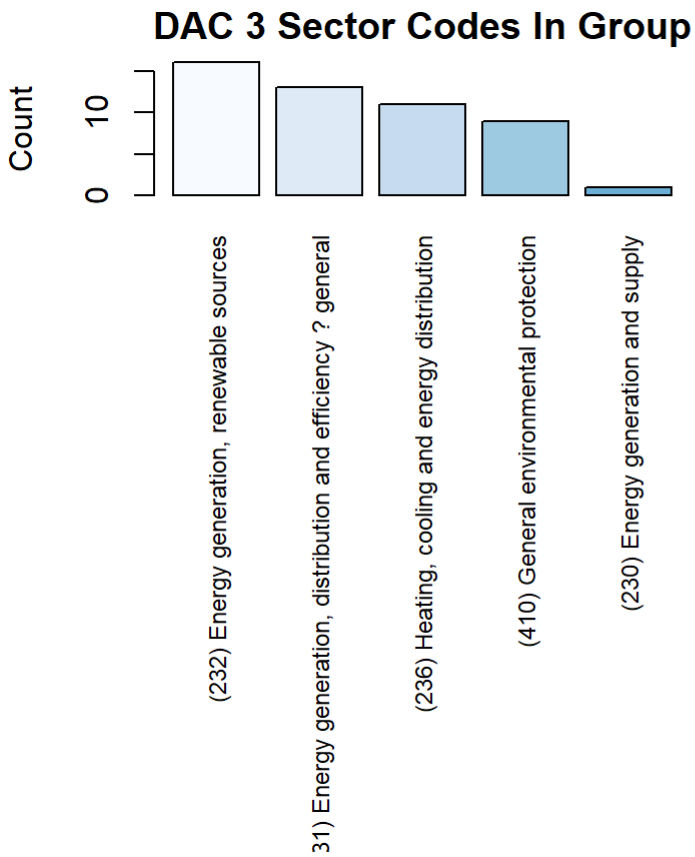
2 projects have been omitted for security purposes

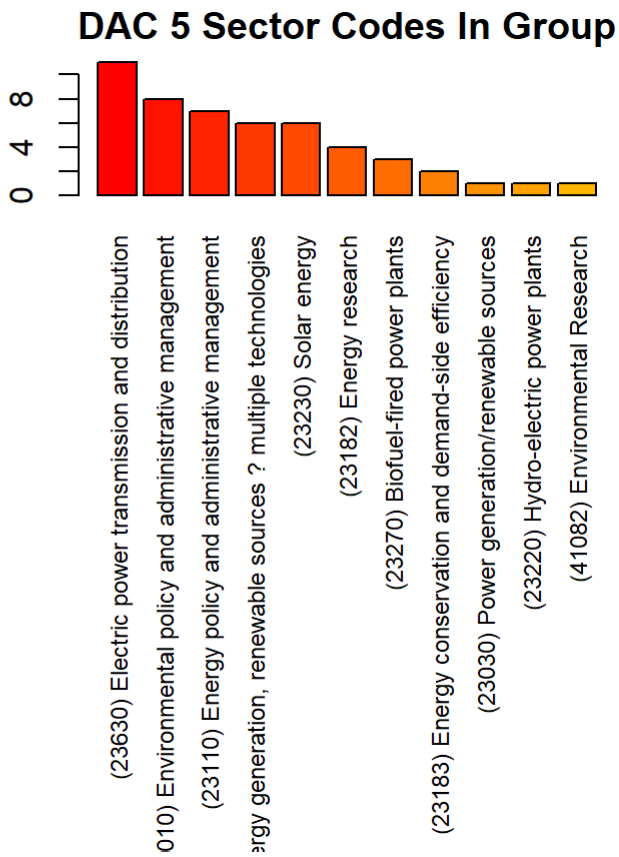
ID	Title	Department
201021	UK Action Against Corruption Programme (UKACT)	DFID MENAD Regional (1548) Governance, Open Societies & Anti-Corruption Dept (1402)
203311	Support to Anti-Corruption Initiatives	DFID Mozambique (1013)
204227	Regional Facility for Strengthening Transnational Responses to Countering Illicit Financial Flows, Corruption and Organised Crime in Africa (CIFFS Facility for Africa)	Africa Regional Department (1001) DFID Southern Africa (1014)
204232	Caribbean Anti-Corruption Programme	DFID Caribbean (1104)
204375	Strengthening Uganda's Anti-Corruption and Accountability Regime (SUGAR)	DFID Uganda (1011)
204659	Strengthening Action Against Corruption in Ghana Programme	DFID Ghana and Liberia (1017)
204819	Anti-corruption in Nigeria Programme	DFID Nigeria (1018)
205173	Tackling Serious and Organised Corruption in Malawi.(TSOC)	DFID Malawi (1012)
205181	International Action Against Corruption	Governance, Open Societies & Anti-Corruption Dept (1402)

Energy


safeti
line
natur
secur
creation
innov
evid
gmg
facilit
investe
suppli
money
carbon
make
downstream
electr
environ
negat
rate
grante
serv
shs
energi
solar
funa
fraud
poor
rbf
develop
potenti
mini
grid
activ
off
disrupt
delay
clean
giz
idcol
caus
scale
creat
privat
continu
exit
panel
rural
insuffici
slow
invest
base
oper
failur
reduc
product
elect
financ
corrupt
subsidi
avail
rais
limit
climat
approxim
expans
aecf
chang
improv
addit
investor
renew
market
design
capac
pay
benefit
green
time
mechan
expect
emiss
initi
technolog
power
instal
facil
compet
reput
sustain
brilho
undermin
condit
consum
plant
complex
disburs
phase

DAC 3 and 5 Sectors in Group



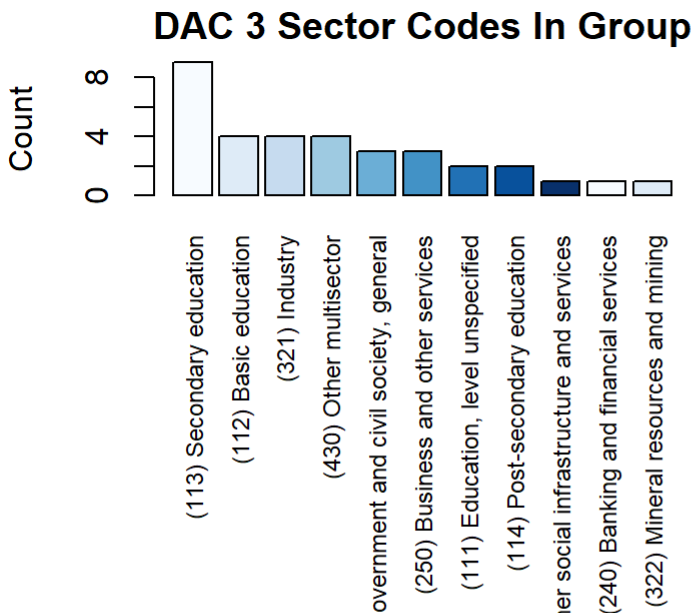


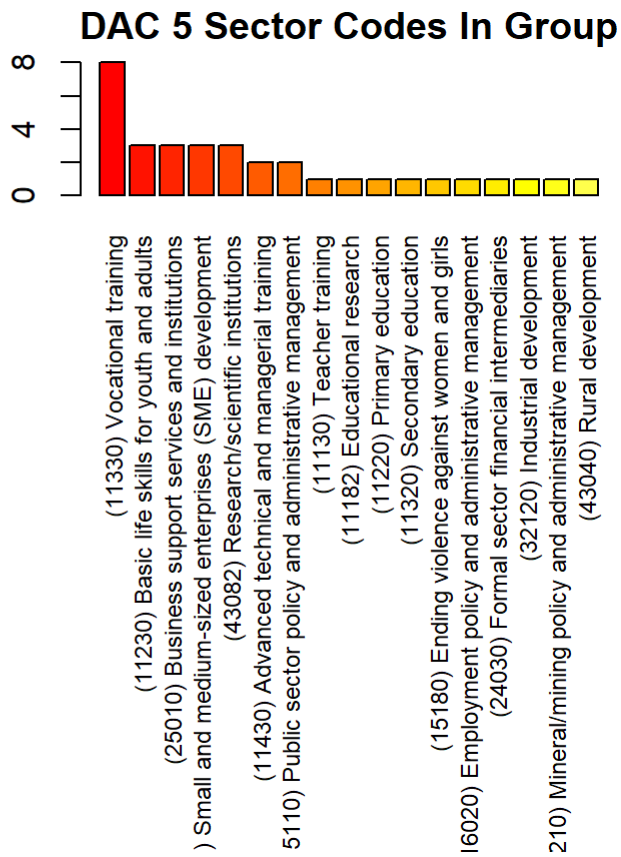
Projects in group

ID	Title	Department
201575	Renewable Energy and Adaptation Climate Technologies (Africa Enterprise Challenge Fund)	Africa Regional Department (1001) DFID Tanzania (1010)
202957	Results Based Financing for Low Carbon Energy Access	Climate and Environment Department (1519)
202976	Providing Clean Energy to the Rural Poor of Bangladesh	DFID Bangladesh (1054)
203624	On and off Grid Small Scale Renewable Energy in Uganda	DFID Uganda (1011)
203871	Energy Security and Resource Efficiency in Somaliland	DFID Somalia (1542)
203998	Green Mini-Grids Kenya	DFID Kenya (1008)
204784	Green Mini-Grids Africa Regional Facility for Market Preparation, Evidence and Policy Development	Africa Regional Department (1001) Research Department (1361)
204837	BRILHO - Energy Africa Mozambique	DFID Mozambique (1013)
205188	Increasing access to electricity in Sierra Leone	DFID Sierra Leone (1019)

Education and Economy

DAC 3 and 5 Sectors in Group





Projects in group

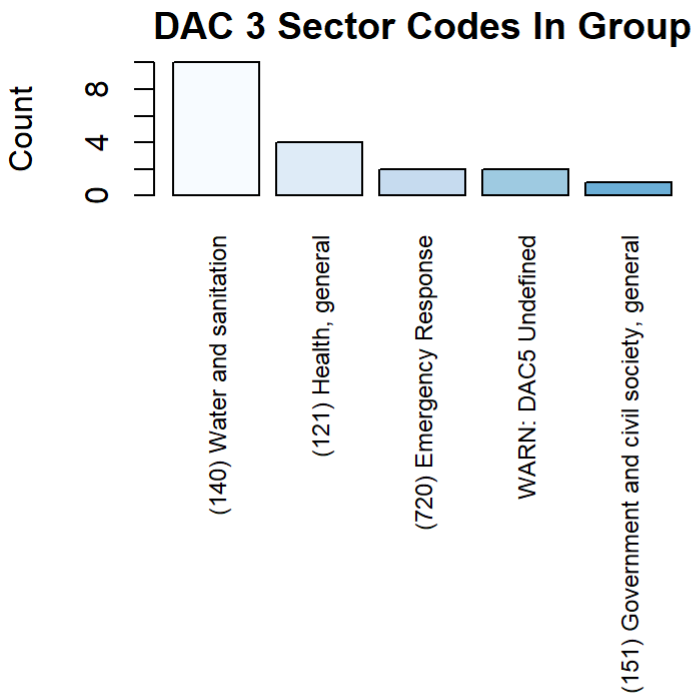
2 projects have been omitted for security purposes

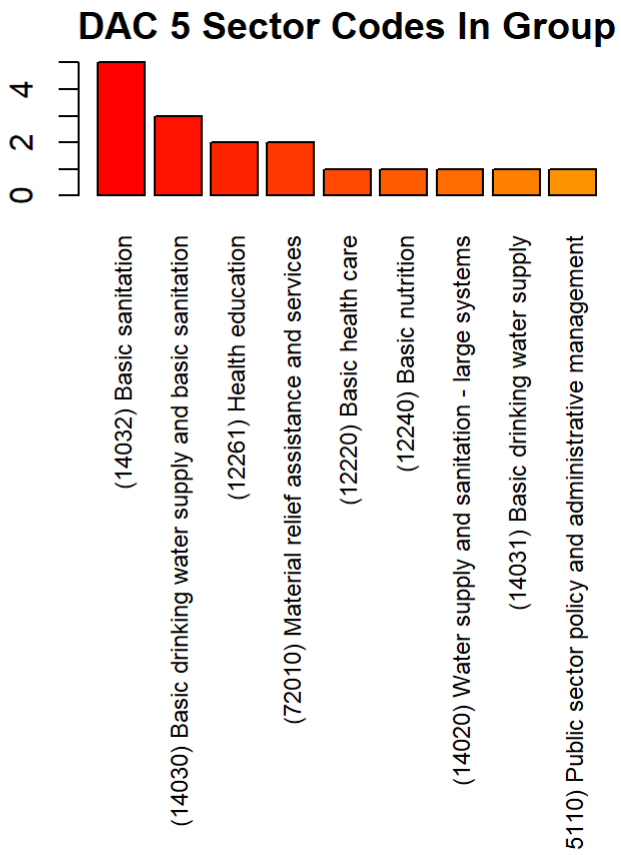
ID	Title	Department
201851	Skills and Employment Programme in Bangladesh (Sudokkho)	DFID Bangladesh (1054)
203870	Youth Education and Skills programme for economic Growth (YES4Growth)	DFID Bangladesh (1054)
204242	Skills for Employment	DFID Mozambique (1013)
204399	Skills Development Programme	DFID Pakistan (1059)
204639	Employment and Skills for Eastern Africa: E4D/SOGA	Africa Regional Department (1001) Governance, Open Societies & Anti-Corruption Dept (1402)
204857	Skills for Employment Programme	DFID Nepal (1058)
205037	Youth Skills for Economic Growth in the Eastern Caribbean	DFID Caribbean (1104)
205144	Empowering Youth in Tanzania	DFID Tanzania (1010)
300453	Jordan Labour Market Programme (JLAMP)	DFID Jordan (1567)

Sanitation


sanit
hygien
water
communiti
pbr
school
improv
rural
poor
health
approach
monitor
champion
infrastructur
sustain
household
construct
benefit
hamper
handwash
inform
addit
materi
beneficiari
agreement
environment
specif
vulner
insuffici
child
own
pay
base
latrin
nutrit
clean
local
nation
maintain
descript
rwsp
improv
evalu
approv
suppli
scheme
prioritis
payment
social
knowledg
train
editori
adequ
reduc
secur
mobil
reput
market
fraudul
refuge
district
behaviour
toward
intervent
miss
inadequ
activ
children
toilet
ngos
verif
fail
flow
grz
host
potenti
standard
humanitarian
reduct
sytem
parallel
creat
delay
infant
integr
wash
rais
author
site
releas
model
promot
status
coverag
avail
stand
contribut
diarrhoea
skill
vulner
exampl

DAC 3 and 5 Sectors in Group





Projects in group

2 projects have been omitted for security purposes

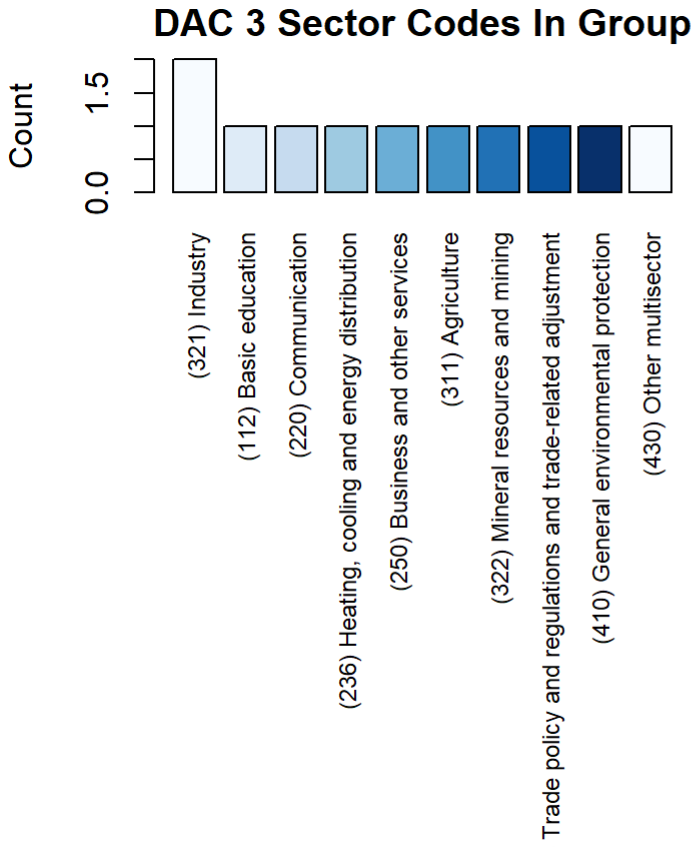
ID	Title	Department
201854	SHINE - Impact of improved Sanitation/ Hygiene and Infant Nutrition on environmental enteropathy, growth, and anemia among young children in Zimbabwe.	DFID Zimbabwe (1016) Human Development Department (1404) Research Department (1361)
202345	Sanitation and Hygiene Programme in Zambia	DFID Zambia (1015)
203187	Rural Water and Sanitation Programme Phase V	DFID Nepal (1058)
204033	Support to Rural Water Supply, Sanitation & Hygiene in Tanzania	DFID Tanzania (1010)

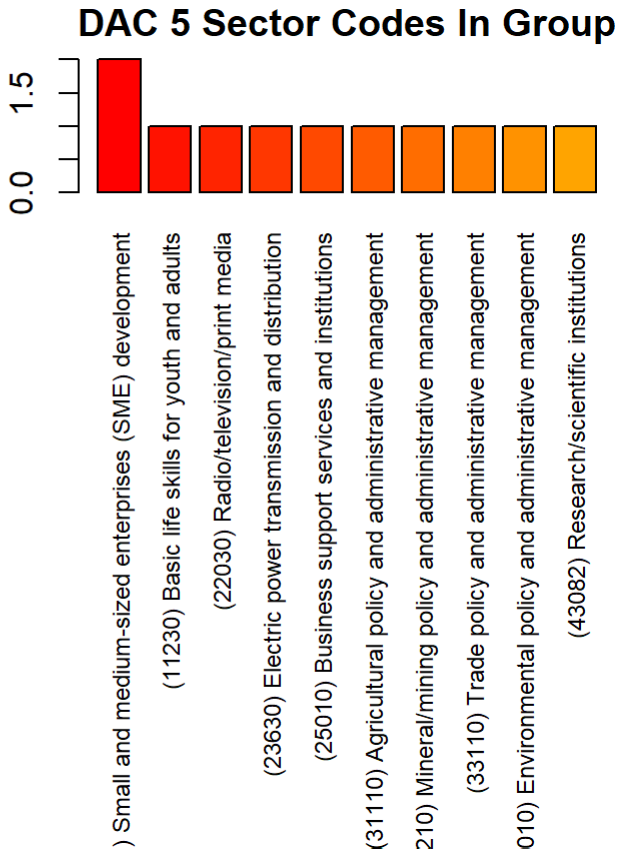
Broad Topics

Large cluster containing 109 projects. Further clustering was performed. A sample of subclusters are included, however, some are too large to be viewed in a presentation.

Sub-cluster 1

DAC 3 and 5 Sectors in Group



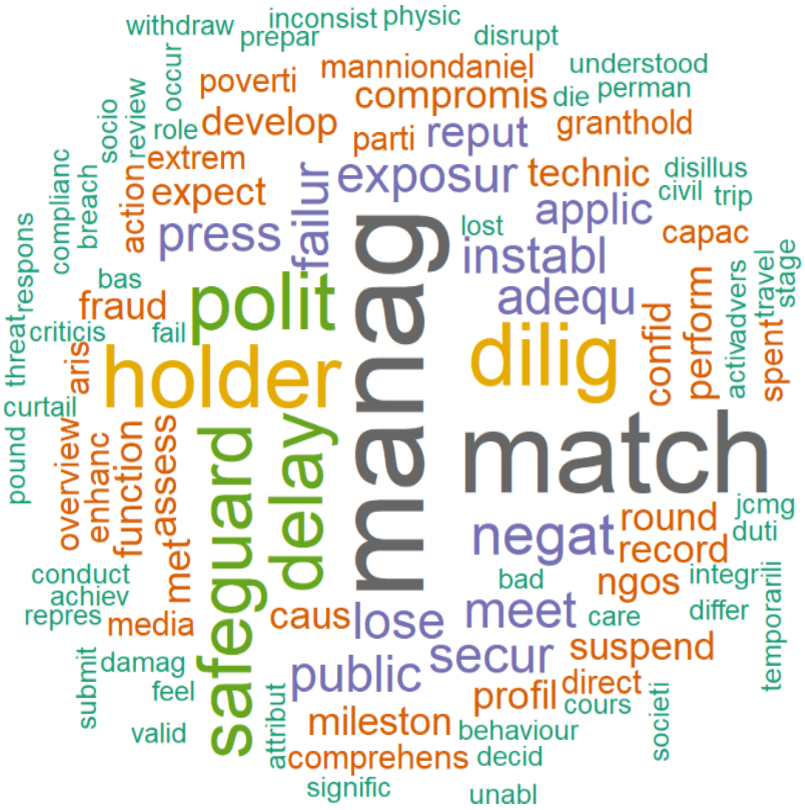


Projects in group

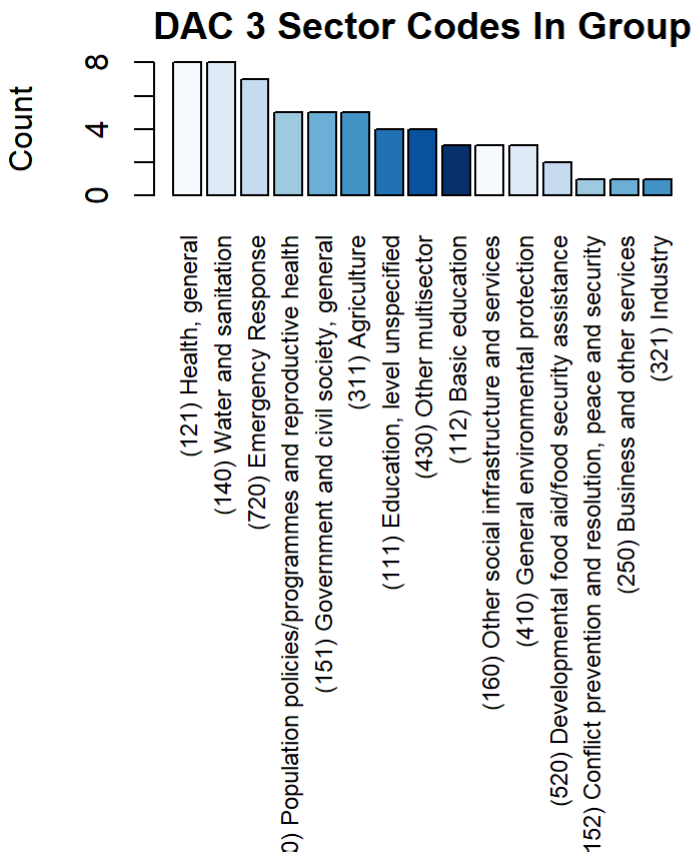
1 project has been omitted for security purposes

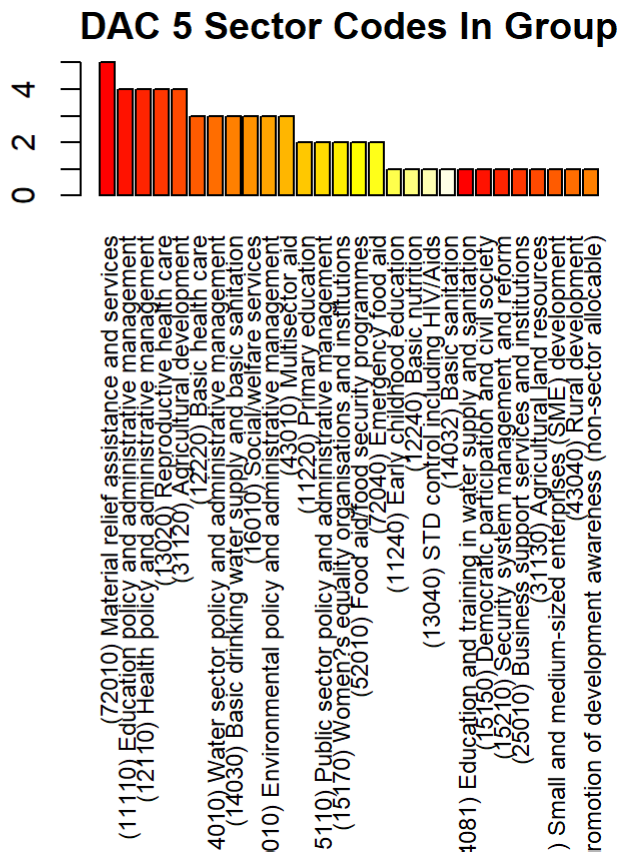
ID	Title	Department
202698	Kenya Market Assistance Programme (MAP)	DFID Kenya (1008)

Sub-cluster 2



DAC 3 and 5 Sectors in Group



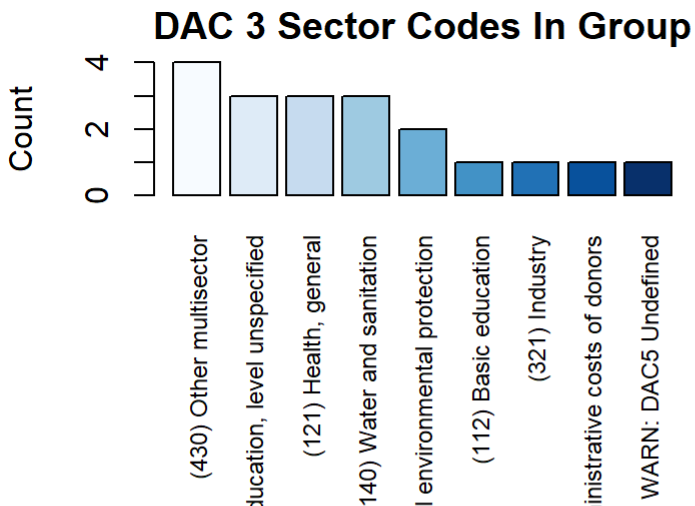


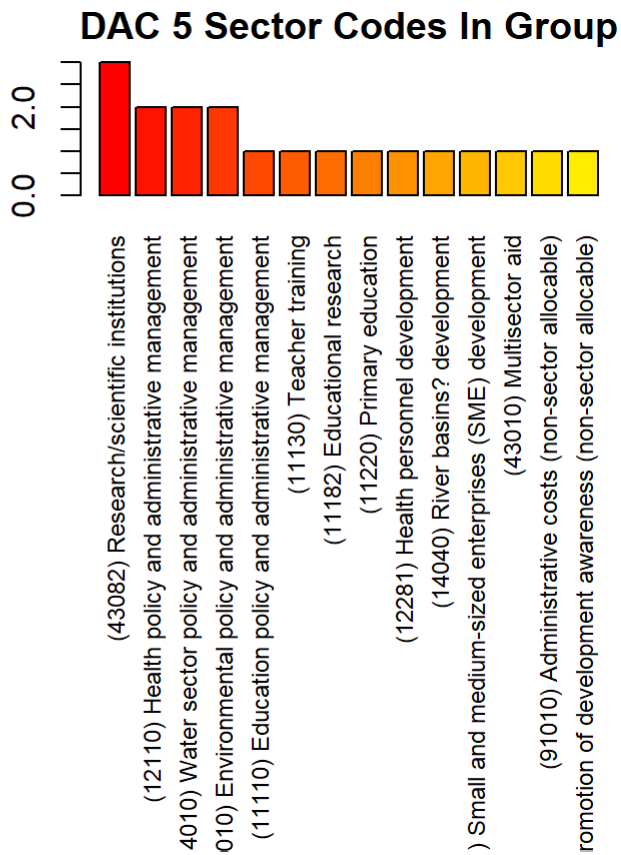
Projects in group

ID	Title	Department
202035	UK Aid Direct Fund	Inclusive Societies Department (1352)
203559	UK Aid Match 2013?2016: giving the public a say in how a portion of the aid budget is spent	Inclusive Societies Department (1352)
205210	UK Aid Match II Fund	Inclusive Societies Department (1352)

Sub-cluster 3

DAC 3 and 5 Sectors in Group





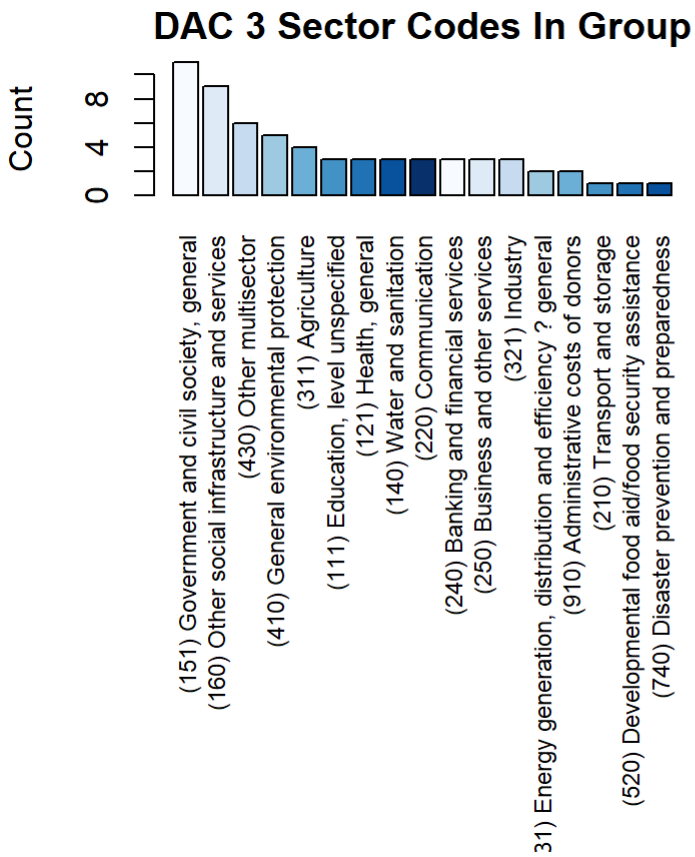
Projects in group

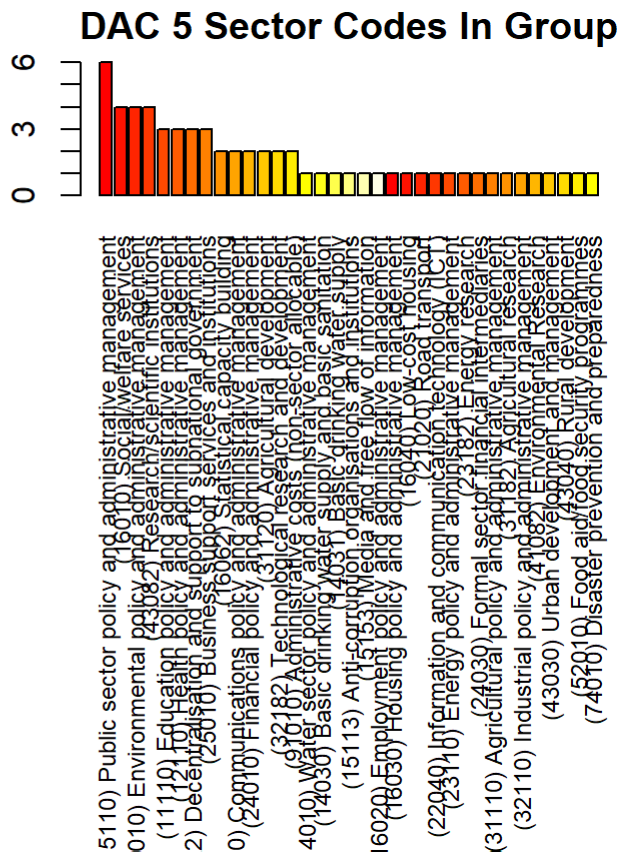
1 project has been omitted for security purposes

ID	Title	Department
200496	Cooperation in International Waters in Africa	Africa Regional Department (1001) DFID Southern Africa (1014)
202208	Health Partnership Scheme (HPS) 2010 to 2019	Human Development Department (1404)
202884	The Water Security Programme	Climate and Environment Department (1519)
203539	Human Development Innovation Fund for Tanzania	DFID Tanzania (1010)
203798	Amplify Open Innovation for Development	Emerging Policy, Innovation & Capability (1400) Private Sector Department (1543)
204277	Ilm Ideas Phase II	DFID Pakistan (1059)
205148	Global Learning for Adaptive Management (GLAM)	Evidence Department (1211)

Sub-cluster 4

DAC 3 and 5 Sectors in Group





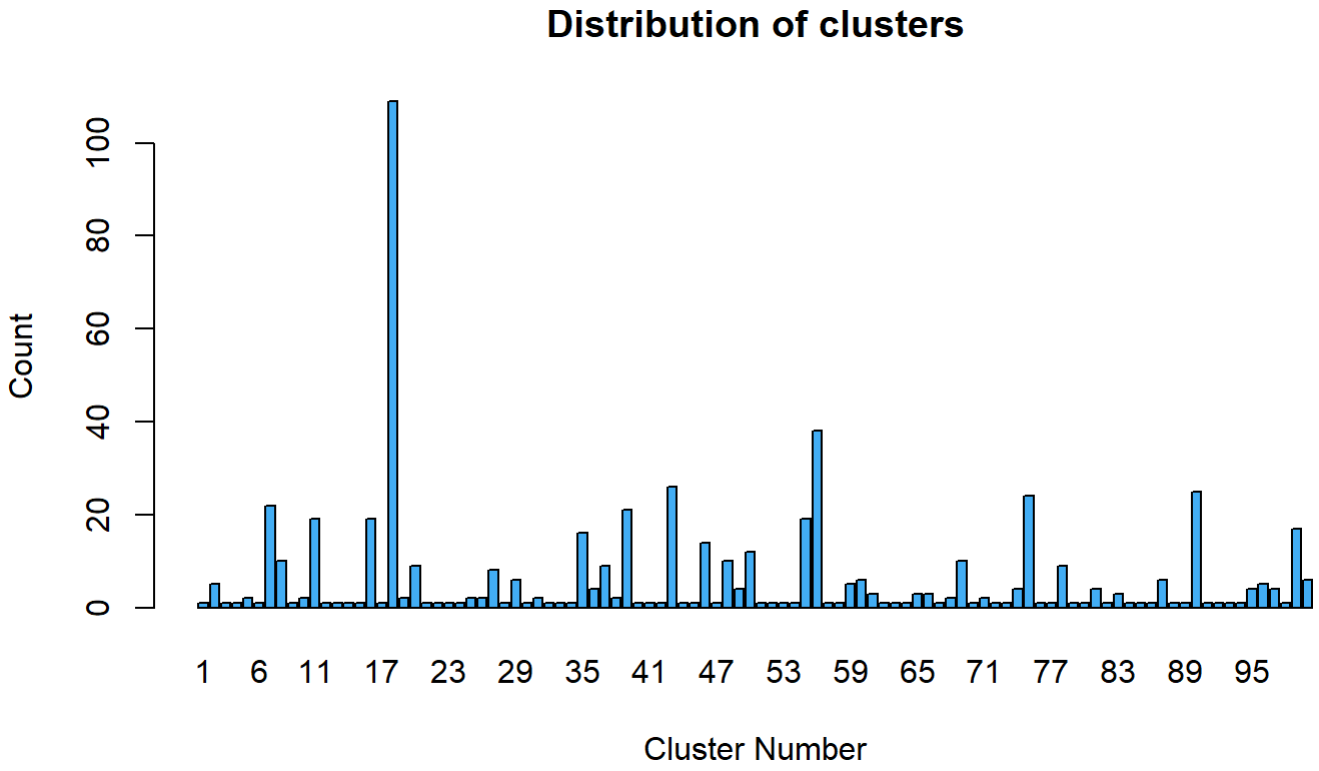
Projects in group

4 projects have been omitted for security purposes

ID	Title	Department
203186	Rural Access Programme 3	DFID Nepal (1058)
203385	Evidence for Development	DFID Nepal (1058)
203452	Support to Palestinians at Risk of Displacement in Israeli Controlled Area C of the West Bank and Gaza	DFID OPTs (1529)
203469	African Risk Capacity (ARC)	Africa Regional Department (1001) Private Sector Department (1543)
203627	Public Sector Performance Programme	DFID Zambia (1015)
203804	M4D - Mobile for Development Strategic Partnership	Private Sector Department (1543) Research Department (1361)
204609	Community Led Infrastructure Finance Facility (CLIFF) Phase 2B	Health and Resilience Dept (1403) Private Sector Department (1543)
204612	Kyrgyz Republic Public Sector Reform Programme	DFID Central Asia (1525)
205189	Sierra Leone Support to the Presidential Delivery Unit Phase II	DFID Sierra Leone (1019)
205275	Better Jobs in Bangladesh	DFID Bangladesh (1054)

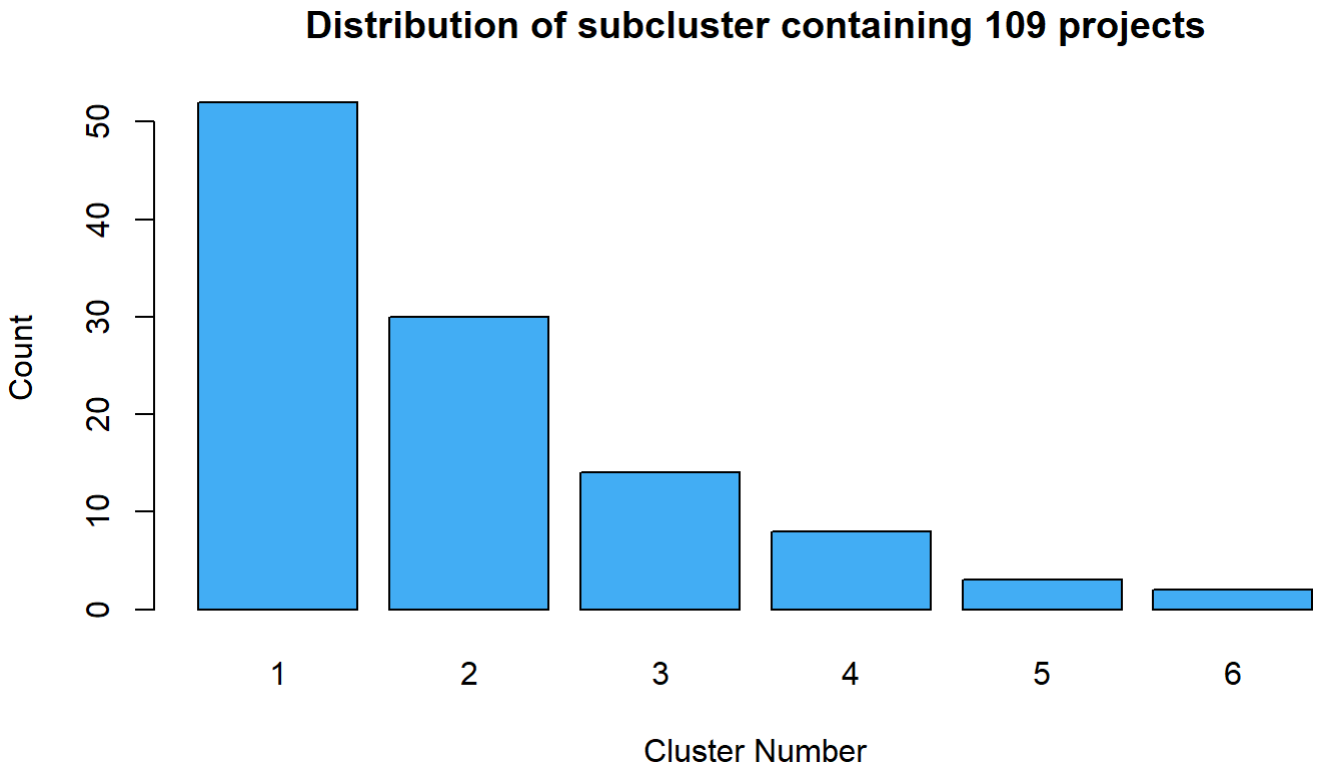
Initial Cluster Distribution

KMeans



Sub-cluster of Largest Cluster

Hierarchical using Title, Purpose and Risk Only



Appendix L – Interview Questions

Interview Guide (semi-structured): **Interviewee:**

1. Interview Introduction

Length: 45-60 minutes

Goal: To evaluate the project groups created by clustering, using the expert knowledge of project managers in DFID.

There are no right or wrong answers, any feedback you have, positive or negative is valuable.

2. Verbal Consent

Are you happy to take part in this interview? Verbal consent obtained Verbal consent NOT obtain

Are you happy for the interview to be recorded? This recording is strictly for reference and will be destroyed upon completion of the project.

Verbal consent obtained Verbal consent NOT obtained

3. Background Information

Warm-up questions

How would you describe an average day as a project manager?

Is knowledge sharing between other project managers across DFID a common practice for you or your team? YES | NO

Yes – Try to remember a recent occasion (within last year) when you shared knowledge/learning with DFID colleagues (or other PMs). Could you briefly tell me about it?

Yes – How did you identify the correct person to contact?

No – Try to remember a recent occasion (within last year) when knowledge sharing with DFID colleagues (or other PMs) would have been useful. Could you briefly tell me about it?

4. Cluster 1 – Malaria Control

Do you think this group makes sense? YES | NO

No – could you explain why?

Yes – could you explain why?

How do you feel about the size of the group?

Suppose you were working on one of these projects, would this group aid you in contacting PMs (or other staff) who could share knowledge?

No – could you explain why?

Yes – could you explain why?

5. Cluster 2 – Justice

Do you think this group makes sense? YES | NO

No – could you explain why?

Yes – could you explain why?

How do you feel about the size of the group?

Suppose you were working on one of these projects, would this group aid you in contacting PMs (or other staff) who could share knowledge?

No – could you explain why?

Yes – could you explain why?

6. Cluster 3 – Corruption and Government

Do you think this group makes sense? YES | NO

No – could you explain why?

Yes – could you explain why?

How do you feel about the size of the group?

Suppose you were working on one of these projects, would this group aid you in contacting PMs (or other staff) who could share knowledge?

No – could you explain why?

Yes – could you explain why?

7. Cluster 4 – Energy

Do you think this group makes sense? YES | NO

No – could you explain why?

Yes – could you explain why?

How do you feel about the size of the group?

Suppose you were working on one of these projects, would this group aid you in contacting PMs (or other staff) who could share knowledge?

No – could you explain why?

Yes – could you explain why?

8. Cluster 5 – Education and Economy

Do you think this group makes sense? YES | NO

No – could you explain why?

Yes – could you explain why?

How do you feel about the size of the group?

Suppose you were working on one of these projects, would this group aid you in contacting PMs (or other staff) who could share knowledge?

No – could you explain why?

Yes – could you explain why?

9. Cluster 6 – Sanitation

Do you think this group makes sense? YES | NO

No – could you explain why?

Yes – could you explain why?

How do you feel about the size of the group?

Suppose you were working on one of these projects, would this group aid you in contacting PMs (or other staff) who could share knowledge?

No – could you explain why?

Yes – could you explain why?

10. Cluster 7 – Sub-cluster 1: Market

Do you think this group makes sense? YES | NO

No – could you explain why?

Yes – could you explain why?

How do you feel about the size of the group?

Suppose you were working on one of these projects, would this group aid you in contacting PMs (or other staff) who could share knowledge?

No – could you explain why?

Yes – could you explain why?

11. Cluster 8 – Sub-cluster 2: Aid Match

Do you think this group makes sense? YES | NO

No – could you explain why?

Yes – could you explain why?

How do you feel about the size of the group?

Suppose you were working on one of these projects, would this group aid you in contacting PMs (or other staff) who could share knowledge?

No – could you explain why?

Yes – could you explain why?

12. Cluster 9 – Sub-cluster 3: Human Development

Do you think this group makes sense? YES | NO

No – could you explain why?

Yes – could you explain why?

How do you feel about the size of the group?

Suppose you were working on one of these projects, would this group aid you in contacting PMs (or other staff) who could share knowledge?

No – could you explain why?

Yes – could you explain why?

13. Cluster 10 – Sub-cluster 4: Broad

Do you think this group makes sense? YES | NO

No – could you explain why?

Yes – could you explain why?

How do you feel about the size of the group?

Suppose you were working on one of these projects, would this group aid you in contacting PMs (or other staff) who could share knowledge?

No – could you explain why?

Yes – could you explain why?

14. Closing Questions

If a tool which produced these results were rolled out, would you use it?
Do you have any other comments of feedback which we have not yet discussed?

Appendix M – R Markdown Code To Create Presentation

```
---
title: "Clustering DfID Aid Data"
author: "Suzie Beith"
date: "8 August 2018"
output:
  ioslides_presentation: default
  autoplay: 30000
---

<style>

table.rmdtable th {
  padding:0.5em 0.5em;
    font-size:14px;
}

table.rmdtable tr > td    {
  font-size:13px;
  padding: 0.5em 0.5em;
}


tr.header {
  font-size: 14px;
  padding: 0.75em 0.5em;
}

/*article#broad {
  margin-top:-40px;
}*/

article#sub4groups table{
  margin-top:-50px;
}

article#sub4groups table tr th:nth-child(2) {
  width:50%;
}

.footer {
    color: black;
    margin:0 auto;
    font-size:9px;
    font-style:italic;
    text-align:center;
}

.footer-l {
  float:left;
  margin-left:10em;
}

.footer-r {
  float:right;
```

```
  margin-right:10em;
}

.footer-c{
  display:inline-block;
  text-align:center;
}

p.subtext{font-size:12px;}

article#projects-in-group-1 table.rmdtable tr:nth-child(6){
  background-color: #ffdddd;
}

h3{
  font-size: 20px;
  margin-top:-5px;
}

</style>

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = FALSE)
```

Introduction

DFID manages many projects from varying sectors across the world.

How does DFID share knowledge and learning across such a complex organisation?

Can we improve the way knowledge is shared to facilitate learning from each other?

The aim of this project is to create a groups of projects ("clusters") that could help project managers to more effectively identify staff working on similar projects or facing simialr challenges.

During this chat I will:

- Ask some questions about your role

- Show you a sample of the groups created

- Ask for your feedback on these groups

Sample Data

      To create each group the following project attributes were used:

      - Title

      - Purpose

      - DAC 3 sector code text

      - DAC 5 sector code text

      - Risk Description
```

```
111.
112.       ## Malaria Control
113.
114.       ```{r, echo=FALSE, message=FALSE, warning=FALSE}
115.       library(tm)
116.       library(qdap)
117.       library(tidytext)
118.       library(dplyr)
119.       library(tidyr)
120.       library(plyr)
121.       library(ggplot2)
122.       library(quanteda)
123.       library(plyr)
124.       library(stats)
125.       library(wordcloud)
126.       library(doParallel)
127.       library(foreach)
128.       library(parallel)
129.       library(fpc)
130.       library(RColorBrewer)
131.       library(knitr)
132.
133.       # Data containing cluster assignment
134.       clust <- read.csv("C:/Users/s-
    beith/Documents/Clustering_Project/Project_Data/Data_with_clusters/Risk_Data/TPDAC/
    dem-TPDR-k-100-dept.csv", stringsAsFactors = FALSE)
135.
136.       #reorder data by cluster number
137.       clust <- clust[order(clust$cluster),]
138.
139.       #split dataframe based on cluster number
140.       clusters <- split(clust, clust$cluster)
141.       # Create name for each cluster dataframe
142.       cname <- as.character(unique(clust$cluster))
143.       cname <- paste("cluster",cname,sep="")
144.
145.       # Assign name to each cluster dataframe and create
146.       for(i in 1:length(clusters)){
147.         assign(cname[i], clusters[[i]])
148.       }
149.
150.       # stopword list from https://gist.github.com/larsyencken/1440509 plus additi
    onal terms added
151.       filename <- "C:/Users/s-beith/Documents/Clustering_Project/stopwords.txt"
152.       stp_wds <- read.delim(filename, stringsAsFactors = FALSE)
153.
154.       df_dac3 <- read.csv("C:/Users/s-
    beith/Documents/Clustering_Project/Project_Data/Creating_Data/ID_DAC3.csv", strings
    AsFactors = F)
155.       #Remove rows which are (998) Unallocated/unspecified
156.       df_dac3 <- subset(df_dac3, df_dac3$DAC.3.Digit.Sector != "(998) Unallocated/
    unspecified")
157.
158.       #ID + DAC5 Info
159.       df_dac5 <- read.csv("C:/Users/s-
    beith/Documents/Clustering_Project/Project_Data/Creating_Data/ID_DAC5.csv", strings
    AsFactors = F)
160.       # Remove rows which are WARN: DAC Sector Undefined
```

```
161.       df_dac5 <- subset(df_dac5, df_dac5$DAC.5.Digit.Sector != "WARN: DAC Sector U
   ndefined")
162.
163.       #Admin project spend codes
164.       admin <- read.csv("C:/Users/s-
   beith/Documents/Clustering_Project/Project_Data/Creating_Data/Admin_Projects_Spend.
   csv", stringsAsFactors = F)
165.
166.       #Funding type codes (admin resource, admin captial and WARN: Undefined)
167.       admin2 <- read.csv("C:/Users/s-
   beith/Documents/Clustering_Project/Project_Data/Creating_Data/Admin_Funding.csv", s
   tringsAsFactors = F)
168.
169.       # remove admin projects
170.       df_dac3 <- df_dac3[!(df_dac3$Project.ID %in% admin$Project.ID),]
171.       df_dac3 <- df_dac3[!(df_dac3$Project.ID %in% admin2$Project.ID),]
172.
173.       df_dac5 <- df_dac5[!(df_dac5$Project.ID %in% admin$Project.ID),]
174.       df_dac5 <- df_dac5[!(df_dac5$Project.ID %in% admin2$Project.ID),]
175.
176.       preprocess <- function(x){
177.
178.         x <- replace_contraction(x) #replace common contractions
179.
180.         x <- replace_abbreviation(x) #replace common abbreviations
181.
182.         x <- tokens(x, what="word",
183.                     remove_numbers = TRUE, remove_punct = TRUE, #rmv numbs only re
   moves strings made entirly of numbers
184.                     remove_url = TRUE) #tokenise
185.
186.         x <- tokens_tolower(x) #lowercase
187.
188.         x <- tokens_select(x, stp_wds$stopwords, selection = "remove") #stopword r
   emoval
189.
190.         x <- tokens_wordstem(x, language = "english") #stemming
191.
192.       }
193.
194.       clean_data <- function(data){
195.         project_text <- paste(data$Project.Title, data$Purpose, data$Risk_list)
196.
197.         #clean data
198.         clean_txt <- preprocess(project_text)
199.         data$clean_text <- clean_txt
200.
201.         text_corpus <- VCorpus(VectorSource(data[,9]))
202.
203.         #remove numbers, special characters and punctuation missed by quanteda
204.         text_corpus <- tm_map(text_corpus, removeNumbers) # removal of numbers
205.         toSpace <- content_transformer(function(x, pattern) gsub(pattern, " ", x))

206.         text_corpus <- tm_map(text_corpus, toSpace, "/|@|//|$|:|:)|*|&|!|?|_|-
   |#|") # replace special characters by space
207.         text_corpus <- tm_map(text_corpus,removePunctuation)
208.
209.         # create DTM - default weighting is term frequency (tf)
210.         dtm_tf <- DocumentTermMatrix(text_corpus)
```

```
    row.names(dtm_tf) <- data$Project.ID

    m <- as.matrix(dtm_tf)
    m <- t(m)
  }

  # clean risk data a
  clean_risk <- function(cl_risk){

    #clean data
    clean_risk <- preprocess(cl_risk[,2])
    cl_risk$clean_risk <- clean_risk

    term_count <- freq_terms(clean_risk, 10)

    # Plot term_count
    plot(term_count, main = "Top Risk Terms")
  }

  top_terms <- function(data){
    # create word cloud of top terms
    v <- sort(rowSums(data),decreasing=TRUE)
    d <- data.frame(word = names(v),freq=v)
    head(d, 10)

    set.seed(100)
    wordcloud(words = d$word, freq = d$freq, min.freq = 1,
              max.words=100, random.order=FALSE, rot.per=0.35,
              colors=brewer.pal(8, "Dark2"))

  }


  # create dataframes for comparison of sectors
  compareD3 <- function(data){
    compare <- merge(data, df_dac3, by="Project.ID")
  }

  compareD5 <- function(data){
    compare2 <- merge(data, df_dac5, by = "Project.ID")
  }

  # set data to cluster being worked on
  data <- cluster60

  # create df for risk text only
  cl_risk <- data[,c(1,6)]


  # create dataframes used for creating dac3 and 5 plots
  compare <- compareD3(data)
  compare2 <- compareD5(data)


  # clean data
  data <- clean_data(data)
  # plot top terms
  top_terms(data)
  # clean risk
```

```
        # cl_risk <- clean_risk(cl_risk)


        ```
        <div class="footer">
        <p class="footer-1">Wordcloud: Top terms accross all features</p>
        </div>

        ## DAC 3 and 5 Sectors in Group

        ```{r, echo=FALSE, message=FALSE, warning=FALSE, fig.width=4}

        dac3 <- function(data){
          # compare <- merge(data, df_dac3, by="Project.ID")
          sectorD3 <- table(compare$DAC.3.Digit.Sector)
          sectorD3 <- sort(sectorD3, decreasing = T)
          par(mar=c(17,5,2,1))
          barplot(sectorD3,col = blues9, las = 3, cex.names=0.75, ylab = "Count", ma
    in = "DAC 3 Sector Codes In Group")
        }

        dac5 <- function(data){
          # compare2 <- merge(data, df_dac5, by = "Project.ID")
          sectorD5 <- table(compare2$DAC.5.Digit.Sector)
          sectorD5 <- sort(sectorD5, decreasing = T)
          par(mar=c(17,5,2,1))
          barplot(sectorD5 ,col = heat.colors(20), las = 3, cex.names=0.75, main = "
    DAC 5 Sector Codes In Group")
        }

        dac3(data)
        dac5(data)

        ```

        ## Projects in group | 3 projects have been omitted for security purposes

        ```{r, echo=FALSE, message=FALSE, warning=FALSE}
        library(DT)
        library(devtools)
        library(xtable)

        c60 <- cluster60 [,c(1,2,8)]
        names(c60) <- c("ID", "Title", "Department")
        c60 <- c60[ which(c60$ID!=c(204005,300191,300249)), ]
        kable(c60, row.names = F, align = "l")


        ```

        ## Justice

        ```{r, echo=FALSE, message=FALSE, warning=FALSE}

        # set data to cluster being worked on
        data <- cluster81
        # create df for risk text only
        cl_risk <- data[,c(1,6)]
```

```
        # create dataframes used for creating dac3 and 5 plots
        compare <- compareD3(data)
        compare2 <- compareD5(data)
        # clean data
        data <- clean_data(data)
        # plot top terms
        top_terms(data)
        # plot top 10 risk terms
        # cl_risk <- clean_risk(cl_risk)

        ```
        <div class="footer">
        <p class="footer-1">Wordcloud: Top terms accross all features</p>
        </div>

        ## DAC 3 and 5 Sectors in Group

        ```{r, echo=FALSE, message=FALSE, warning=FALSE, fig.width=4}

        dac3(data)
        dac5(data)

        ```

        ## Projects in group | 2 projects have been omitted for security purposes

        ```{r, echo=FALSE, message=FALSE, warning=FALSE}

        c81 <- cluster81 [,c(1,2,8)]
        names(c81) <- c("ID", "Title", "Department")
        c81 <- c81[ which(c81$ID!=c(202559,300255)), ]
        kable(c81, row.names = F, align = "l")

        ```

        ## Corruption and Government

        ```{r, echo=FALSE, message=FALSE, warning=FALSE}

        # set data to cluster being worked on
        data <- cluster48
        # create df for risk text only
        cl_risk <- data[,c(1,6)]
        # create dataframes used for creating dac3 and 5 plots
        compare <- compareD3(data)
        compare2 <- compareD5(data)
        # clean data
        data <- clean_data(data)
        # plot top terms
        top_terms(data)
        # plot top 10 risk terms
        # cl_risk <- clean_risk(cl_risk)

        ```
        <div class="footer">
        <p class="footer-1">Wordcloud: Top terms accross all features</p>
        </div>

```
```

````
386.
387.       ## DAC 3 and 5 Sectors in Group
388.
389.       ```{r, echo=FALSE, message=FALSE, warning=FALSE, fig.width=4}
390.
391.       dac3(data)
392.       dac5(data)
393.
394.       ```
395.
396.       ## Projects in group | 2 projects have been omitted for security purposes
397.
398.       ```{r, echo=FALSE, message=FALSE, warning=FALSE}
399.
400.       c48 <- cluster48 [,c(1,2,8)]
401.       names(c48) <- c("ID", "Title", "Department")
402.       c48 <- c48[ which(c48$ID!=c(203311,300291)), ]
403.       kable(c48, row.names = F, align = "l")
404.
405.
406.       ```
407.
408.       ## Energy
409.
410.       ```{r, echo=FALSE, message=FALSE, warning=FALSE}
411.
412.       # set data to cluster being worked on
413.       data <- cluster78
414.       # create df for risk text only
415.       cl_risk <- data[,c(1,6)]
416.       # create dataframes used for creating dac3 and 5 plots
417.       compare <- compareD3(data)
418.       compare2 <- compareD5(data)
419.       # clean data
420.       data <- clean_data(data)
421.       # plot top terms
422.       top_terms(data)
423.       # plot top 10 risk terms
424.       # cl_risk <- clean_risk(cl_risk)
425.
426.       ```
427.       <div class="footer">
428.       <p class="footer-1">Wordcloud: Top terms accross all features</p>
429.       </div>
430.
431.
432.       ## DAC 3 and 5 Sectors in Group
433.
434.       ```{r, echo=FALSE, message=FALSE, warning=FALSE, fig.width=4}
435.
436.       dac3(data)
437.       dac5(data)
438.
439.       ```
440.
441.       ## Projects in group
442.
443.       ```{r, echo=FALSE, message=FALSE, warning=FALSE}
444.
````

```
c78 <- cluster78 [,c(1,2,8)]
names(c78) <- c("ID", "Title", "Department")
kable(c78, row.names = F, align = "l")


```

Education and Economy

```{r, echo=FALSE, message=FALSE, warning=FALSE}

# set data to cluster being worked on
data <- cluster20
# create df for risk text only
cl_risk <- data[,c(1,6)]
# create dataframes used for creating dac3 and 5 plots
compare <- compareD3(data)
compare2 <- compareD5(data)
# clean data
data <- clean_data(data)
# plot top terms
top_terms(data)
# plot top 10 risk terms
# cl_risk <- clean_risk(cl_risk)

```

Wordcloud: Top terms accross all features

DAC 3 and 5 Sectors in Group

```{r, echo=FALSE, message=FALSE, warning=FALSE, fig.width=4}

dac3(data)
dac5(data)

```

Projects in group | 2 projects have been omitted for security purposes

```{r, echo=FALSE, message=FALSE, warning=FALSE}

c20 <- cluster20 [,c(1,2,8)]
names(c20) <- c("ID", "Title", "Department")
c20 <- c20[ which(c20$ID!=c(205144,300453)), ]
kable(c20, row.names = F, align = "l")


```

Sanitation

```{r, echo=FALSE, message=FALSE, warning=FALSE}

# set data to cluster being worked on
data <- cluster87
# create df for risk text only
```

```
        cl_risk <- data[,c(1,6)]
        # create dataframes used for creating dac3 and 5 plots
        compare <- compareD3(data)
        compare2 <- compareD5(data)
        # clean data
        data <- clean_data(data)
        # plot top terms
        top_terms(data)
        # plot top 10 risk terms
        # cl_risk <- clean_risk(cl_risk)

        ```
        <div class="footer">
        <p class="footer-1">Wordcloud: Top terms accross all features</p>
        </div>

        ## DAC 3 and 5 Sectors in Group

        ```{r, echo=FALSE, message=FALSE, warning=FALSE, fig.width=4}

        dac3(data)
        dac5(data)

        ```

        ## Projects in group | 2 projects have been omitted for security purposes

        ```{r, echo=FALSE, message=FALSE, warning=FALSE}

        c87 <- cluster87 [,c(1,2,8)]
        names(c87) <- c("ID", "Title", "Department")
        c87 <- c87[ which(c87$ID!=c(300331,300353)), ]
        kable(c87, row.names = F, align = "l")


        ```

        ## Broad Topics {#broad}

        <div>
        <p class="subtext">Large cluster containing 109 projects. Further clustering
    was performed. A sample of subclusters are included, however, some are too large t
   o be viewed in a presentation.</p>
        </div>

        ```{r, echo=FALSE, message=FALSE, warning=FALSE}

        # set data to cluster being worked on
        data <- cluster18
        # create df for risk text only
        cl_risk <- data[,c(1,6)]
        # create dataframes used for creating dac3 and 5 plots
        compare <- compareD3(data)
        compare2 <- compareD5(data)
        # clean data
        data <- clean_data(data)
        # plot top terms
```

```
        top_terms(data)
        # plot top 10 risk terms
        # cl_risk <- clean_risk(cl_risk)

        ```
        <div class="footer">
        <p class="footer-1">Wordcloud: Top terms accross all features</p>
        </div>

        ## Subcluster 1


        ```{r, echo=FALSE, message=FALSE, warning=FALSE}

        # Data containing cluster assignment
        sub_clust <- read.csv("C:/Users/s-beith/Documents/Clustering_Project/Project_Data/Data_with_clusters/Risk_Data/TPDAC/subcluster/TPRD/hclust_c18_TPR-dept.csv", stringsAsFactors = FALSE)

        #reorder data by cluster number
        sub_clust <- sub_clust[order(sub_clust$cluster),]

        #split dataframe based on cluster number
        sub_clusters <- split(sub_clust, sub_clust$cluster)
        # Create name for each cluster dataframe
        cname <- as.character(unique(sub_clust$cluster))
        cname <- paste("sub_cluster",cname,sep="")

        # Assign name to each cluster dataframe and create
        for(i in 1:length(sub_clusters)){
          assign(cname[i], sub_clusters[[i]])
        }

        # set data to cluster being worked on
        data <- sub_cluster6
        # create df for risk text only
        cl_risk <- data[,c(2,6)]
        # create dataframes used for creating dac3 and 5 plots
        compare <- compareD3(data)
        compare2 <- compareD5(data)
        # clean data
        data <- clean_data(data)
        # plot top terms
        top_terms(data)
        # plot top 10 risk terms
        # cl_risk <- clean_risk(cl_risk)

        ```
        <div class="footer">
        <p class="footer-1">Wordcloud: Top terms accross all features</p>
        </div>

        ## DAC 3 and 5 Sectors in Group

        ```{r, echo=FALSE, message=FALSE, warning=FALSE, fig.width=4}

        dac3(data)
```

````
        dac5(data)

        ```

        ## Projects in group | 1 project has been omitted for security purposes

        ```{r, echo=FALSE, message=FALSE, warning=FALSE}

        sc6 <- sub_cluster6 [,c(1,2,8)]
        names(sc6) <- c("ID", "Title", "Department")
        sc6 <- sc6[ which(sc6$ID!=203719), ]
        kable(sc6, row.names = F, align = "l")


        ```

        ## Subcluster 2

        ```{r, echo=FALSE, message=FALSE, warning=FALSE}
        # set data to cluster being worked on
        data <- sub_cluster5
        # create df for risk text only
        cl_risk <- data[,c(2,6)]
        # create dataframes used for creating dac3 and 5 plots
        compare <- compareD3(data)
        compare2 <- compareD5(data)
        # clean data
        data <- clean_data(data)
        # plot top terms
        top_terms(data)
        # plot top 10 risk terms
        # cl_risk <- clean_risk(cl_risk)

        ```
        <div class="footer">
        <p class="footer-1">Wordcloud: Top terms accross all features</p>
        </div>


        ## DAC 3 and 5 Sectors in Group

        ```{r, echo=FALSE, message=FALSE, warning=FALSE, fig.width=4}

        dac3(data)
        dac5(data)

        ```

        ## Projects in group

        ```{r, echo=FALSE, message=FALSE, warning=FALSE}

        sc5 <- sub_cluster5 [,c(1,2,8)]
        names(sc5) <- c("ID", "Title", "Department")
        # sc5 <- sc5[ which(sc5$ID!=300272), ]
        kable(sc5, row.names = F, align = "l")


        ```
````
````

```
677.
678.       ## Subcluster 3
679.
680.       ```{r, echo=FALSE, message=FALSE, warning=FALSE}
681.       # set data to cluster being worked on
682.       data <- sub_cluster4
683.       # create df for risk text only
684.       cl_risk <- data[,c(2,6)]
685.       # create dataframes used for creating dac3 and 5 plots
686.       compare <- compareD3(data)
687.       compare2 <- compareD5(data)
688.       # clean data
689.       data <- clean_data(data)
690.       # plot top terms
691.       top_terms(data)
692.       # plot top 10 risk terms
693.       # cl_risk <- clean_risk(cl_risk)
694.
695.       ```
696.       <div class="footer">
697.       <p class="footer-1">Wordcloud: Top terms accross all features</p>
698.       </div>
699.
700.
701.       ## DAC 3 and 5 Sectors in Group
702.
703.       ```{r, echo=FALSE, message=FALSE, warning=FALSE, fig.width=4}
704.
705.       dac3(data)
706.       dac5(data)
707.
708.       ```
709.
710.       ## Projects in group | 1 project has been omitted for security purposes
711.
712.       ```{r, echo=FALSE, message=FALSE, warning=FALSE}
713.
714.       sc4 <- sub_cluster4 [,c(1,2,8)]
715.       names(sc4) <- c("ID", "Title", "Department")
716.       sc4 <- sc4[ which(sc4$ID!=300272), ]
717.       kable(sc4, row.names = F, align = "l")
718.
719.
720.       ```
721.
722.       ## Subcluster 4
723.
724.       ```{r, echo=FALSE, message=FALSE, warning=FALSE}
725.       # set data to cluster being worked on
726.       data <- sub_cluster3
727.       # create df for risk text only
728.       cl_risk <- data[,c(2,6)]
729.       # create dataframes used for creating dac3 and 5 plots
730.       compare <- compareD3(data)
731.       compare2 <- compareD5(data)
732.       # clean data
733.       data <- clean_data(data)
734.       # plot top terms
735.       top_terms(data)
```

```
        # plot top 10 risk terms
        # cl_risk <- clean_risk(cl_risk)

        ```
        <div class="footer">
        <p class="footer-1">Wordcloud: Top terms accross all features</p>
        </div>

        ## DAC 3 and 5 Sectors in Group

        ```{r, echo=FALSE, message=FALSE, warning=FALSE, fig.width=4}

        dac3(data)
        dac5(data)

        ```

        ## Projects in group| 4 projects have been omitted for security purposes {#sub4groups}

        ```{r, echo=FALSE, message=FALSE, warning=FALSE}

        sc3 <- sub_cluster3 [,c(1,2,8)]
        names(sc3) <- c("ID", "Title", "Department")
         sc3 <- sc3[ c(2:11), ]
        kable(sc3, row.names = F, align = "l")


        ```

        ## Initial Cluster Distribution | KMeans

        ```{r, echo=FALSE, message=FALSE, warning=FALSE}
        size <- table(clust$cluster)
        barplot(size, col = "#42adf4", xlab = "Cluster Number", ylab = "Count",
                main = "Distribution of clusters")

        ```

        ## Subcluster of Largest Cluster | Hierarchical using Title, Purpose and Risk Only

        ```{r, echo=FALSE, message=FALSE, warning=FALSE}
        size <- table(sub_clust$cluster)
        barplot(size, col = "#42adf4", xlab = "Cluster Number", ylab = "Count",
                main = "Distribution of subcluster containing 109 projects")

        ```
```
```

Bibliography

1. Pathak M. Beginning data science with R. Springer International Publishing; 2014. p. 137
2. Langer D. datasciencedojo/IntroToTextAnalyticsWithR [Internet]. GitHub. 2017 [cited 28 May 2018]. Available from: https://github.com/datasciencedojo/IntroToTextAnalyticsWithR
3. de Gemmis M, Lops P, Musto C, Narducci F, Semeraro G. Semantics-Aware Content-Based Recommender Systems. In: Ricci F, Rokach L, Shapira B, ed. by. Recommender Systems Handbook. 2nd ed. New York: Springer Science+Business Media; 2018. p. 119 - 146.
4. Kumar V, Sridhar R. Unsupervised Topic Modelling for Short Texts Using Distributed Representations of Words. New Jersey. AT&T Labs.
5. Dumais S. Latent Semantic Analysis. Annual Review of Science and Technology. 38th ed. 2018. p. 188 - 230.

6. Kim H, Kim H, Cho S. Bag-of-concepts: Comprehending document representation through clustering words in distributed representation. Neurocomputing. 2017;266:336-352.
7. Le Q. Mikolov T. Distributed Representations of Sentences and Documents. California. Google Inc. 2014
8. Mikolov T. Chen K. Corrado G. J Dean. Efficient Estimation of Word Representations in Vector Space. California. Google Inc. 2013
9. Pennington J, Socher R, Manning C. GloVe: Global Vectors for Word Representation. California. Computer Science Department Stanford. 2014
10. Swamynathan M. Mastering Machine Learning with Python in Six Steps. Berkeley, CA: Apress; 2017.
11. Sodhani S. Notes for GloVe paper [Internet]. Gist. 2018 [cited 6 June 2018]. Available from: https://gist.github.com/shagunsodhani/efea5a42d17e0fcf18374df8e3e4b3e8
12. Pennington J. GloVe: Global Vectors for Word Representation [Internet]. Nlp.stanford.edu. 2014 [cited 19 July 2018]. Available from: https://nlp.stanford.edu/projects/glove/
13. Venables W, Ripley B. Modern applied statistics with S. 4th ed. Springer; 2002. p.315
14. Guides E. Assessing clustering tendency: A vital issue – Unsupervised Machine Learning [Internet]. R-bloggers. 2016 [cited 1 June 2018]. Available from: https://www.r-bloggers.com/assessing-clustering-tendency-a-vital-issue-unsupervised-machine-learning/
15. Kassambara A. Assessing Clustering Tendency: Essentials - Articles - STHDA [Internet]. Sthda.com. 2017 [cited 25 May 2018]. Available from: http://www.sthda.com/english/articles/29-cluster-validation-essentials/95-assessing-clustering-tendency-essentials/
16. Xu R, Wunsch D. Cluster Analysis. Clustering. 1st ed. Wiley-IEEE; 2009. p. 5.
17. Xu R, Wunsch D. Partitional Clustering. Clustering 1st ed. Wiley-IEEE; 2009. p. 64.
18. Venables W, Ripley B. Modern applied statistics with S. 4th ed. Springer; 2002. p. 316
19. Gordon A. Classification. Boca Raton: Chapman & Hall/CRC; 1999. p. 41-53
20. Xu R, Wunsch D. Cluster Analysis. Clustering. 1st ed. Wiley-IEEE; 2009. p. 34.
21. Zhao Y, Karypis G, Fayyad U. Hierarchical Clustering Algorithms for Document Datasets. Data Mining and Knowledge Discovery. 2005;10(2):141-168.
22. Xu R, Wunsch D. Cluster Analysis. Clustering. 1st ed. Wiley-IEEE; 2009. p. 37.
23. Steinbach M, Karypis G, Kumar V. A Comparison of Document Clustering Techniques. 2000;.
24. Venables W, Ripley B. Modern applied statistics with S. 4th ed. Springer; 2002. P318
25. Xu R, Wunsch D. Partitional Clustering. Clustering. 1st ed. Wiley-IEEE; 2009. p. 68.
26. Xu R, Wunsch D. Partitional Clustering. Clustering. 1st ed. Wiley-IEEE; 2009. p. 69.
27. Venables W, Ripley B. Modern applied statistics with S. 4th ed. Springer; 2002. p. 320
28. Sonagaral D, Badheka S. Comparison of Basic Clustering Algorithms. International Journal of Computer Science and Mobile Computing. 2014;3(10). p. 58-61.
29. HUANG Z. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery. 1998;2. p. 283-304.
30. Suyal H, Panwar A, Singh Negi A. Text Clustering Algorithms: A Review. International Journal of Computer Applications. 2014;96(24). p. 36-40.
31. K-means and K-medoids [Internet]. Math.le.ac.uk. 2018 [cited 23 May 2018]. Available from: http://www.math.le.ac.uk/people/ag153/homepage/KmeansKmedoids/Kmeans_Kmedoids.html

32. Kassambara A. CLARA - Clustering Large Applications - Articles - STHDA [Internet]. Sthda.com. 2017 [cited 25 May 2018]. Available from: http://www.sthda.com/english/articles/27-partitioning-clustering-essentials/89-clara-clustering-large-applications/
33. Kassambara A. Determining The Optimal Number Of Clusters: 3 Must Know Methods - Articles - STHDA [Internet]. Sthda.com. 2018 [cited 7 June 2018]. Available from: http://www.sthda.com/english/articles/29-cluster-validation-essentials/96-determining-the-optimal-number-of-clusters-3-must-know-methods/
34. Rousseeuw P. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 1987;20 p. 53-65.
35. Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust: AnRPackage for Determining the Relevant Number of Clusters in a Data Set. Journal of Statistical Software. 2014;61(6).
36. Pelleg D, Moore A. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. 2002.
37. Srivastava A, Kumar S. Dynamic Reconfiguration of robot software component in real time distributed system using clustering techniques. Procedia Computer Science. 2018;125. p. 754-761.
38. Niebles J, Li F. Lecture 13: k-means and mean-shift clustering [Internet]. Vision.stanford.edu. 2016 [cited 7 June 2018]. Available from: http://vision.stanford.edu/teaching/cs131_fall1617/lectures/lecture13_kmeans_mean_shift_cs131_2016
39. Jarad. Document clustering using Mean Shift [Internet]. Stack Overflow. 2018 [cited 10 June 2018]. Available from: https://stackoverflow.com/questions/46183760/document-clustering-using-mean-shift
40. Hahsler M, Piekenbrock M, Doran D. dbscan: Fast Density-based Clustering with R [Internet]. Cran.r-project.org. 2018 [cited 4 June 2018]. Available from: https://cran.r-project.org/web/packages/dbscan/vignettes/dbscan.pdf
41. Nagaraju S, Kashyap M, Bhattachraya M. An effective density based approach to detect complex data clusters using notion of neighborhood difference. International Journal of Automation and Computing. 2016;14(1). p. 57-67.
42. Xu R, Wunsch D. Cluster Validity. Clustering. 1st ed. Wiley-IEEE; 2009. p. 263.
43. Kovács F, Legány C, Babos A. Cluster Validity Measurement Techniques. 2018.
44. L. Kurgan, P. Musilek, A survey of knowledge discovery and data mining process models, The Knowledge Engineering Review 21 (1) (2006). p. 1–24.
45. Kobayashi V, Mol S, Berkers H, Kismihók G, Den Hartog D. Text Mining in Organizational Research. Organizational Research Methods. 2017;21(3). p. 733-765.
46. Powell V, Lehe L. Principal Component Analysis explained visually [Internet]. Explained Visually. 2018 [cited 10 June 2018]. Available from: http://setosa.io/ev/principal-component-analysis/
47. Castro-Schilo L. Text analysis in the social sciences: A new spectrum of possibilities [Internet]. JMP User Community. 2016 [cited 29 May 2018]. Available from: https://community.jmp.com/t5/JMP-Blog/Text-analysis-in-the-social-sciences-A-new-spectrum-of/ba-p/45139
48. Abdi H, Williams L. Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics. 2010;2(4). p. 433-459.
49. Piech C, Ng A. CS221 [Internet]. Stanford.edu. 2013 [cited 2 June 2018]. Available from: http://stanford.edu/~cpiech/cs221/handouts/kmeans.html
50. AIC vs. BIC [Internet]. The Methodology Center. 2018 [cited 4 June 2018]. Available from: https://methodology.psu.edu/AIC-vs-BIC

51. Rendón E, Abundez I, Arizmendi A, Quiroz E. Internal versus External cluster validation indexes. INTERNATIONAL JOURNAL OF COMPUTERS AND COMMUNICATIONS. 2011;5(1):27-34.
52. Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, Junjie Wu, Sen Wu. Understanding and Enhancement of Internal Clustering Validation Measures. IEEE Transactions on Cybernetics. 2013;43(3). p. 982-994.
53. Denscombe M. The Good Research Guide. Maidenhead: McGraw-Hill Education; 2014. p. 186
54. Denscombe M. The Good Research Guide. Maidenhead: McGraw-Hill Education; 2014. p.192
55. Denscombe M. The Good Research Guide. Maidenhead: McGraw-Hill Education; 2014. p. 202-203
56. Osei-Bryson K. Towards supporting expert evaluation of clustering results using a data mining process model. Information Sciences. 2010;180(3). p. 414-431.
57. Kim Y. Weighted order-dependent clustering and visualization of web navigation patterns. Decision Support Systems. 2007;43(4). p. 1630-1645.
58. C. Shearer. The CRISP-DM Model: The New Blueprint For Data Mining. Journal of Data Warehousing. 2000;5(4). p. 13–22.
59. Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C et al. CRISP-DM 1.0 - Step-by-step data mining guide. 2000.
60. Hahsler M. kNNdist: Calculate and plot the k-Nearest Neighbor Distance in dbscan: Density Based Clustering of Applications with Noise (DBSCAN) and Related Algorithms [Internet]. Rdrr.io. 2018 [cited 12 June 2018]. Available from: https://rdrr.io/cran/dbscan/man/kNNdist.html
61. Thirumuruganathan S. Introduction To Mean Shift Algorithm [Internet]. God, Your Book Is Great !!. 2010 [cited 18 June 2018]. Available from: https://saravananthirumuruganathan.wordpress.com/2010/04/01/introduction-to-mean-shift-algorithm/
62. Liu Y, Li Z, Xiong H, Gao X, Wu J. Understanding of Internal Clustering Validation Measures. 2010 IEEE International Conference on Data Mining. 2010;.
63. Denscombe M. The Good Research Guide. Maidenhead: McGraw-Hill Education; 2014.
64. Wilson C. Interview techniques for UX practitioners. Waltham, MA: Morgan Kaufmann; 2014.
65. Galili T. K-means Clustering (from 'R in Action') [Internet]. R-statistics blog. 2013 [cited 15 June 2018]. Available from: https://www.r-statistics.com/2013/08/k-means-clustering-from-r-in-action/
66. Development Tracker [Internet]. Devtracker.dfid.gov.uk. 2018 [cited 10 September 2018]. Available from: https://devtracker.dfid.gov.uk/
67. Baumeister R. The Causes and Consequences of 'Irrational' Conducts. In: Brocas I, Carrillo J, ed. by. The Psychology of Economic Decisions. 1st ed. Oxford University Press; 2003. p. 1-15.
68. Baker S, Edwards R. How many qualitative interviews is enough?. National Centre for Research Methods Review Paper. 2018;
69. Gorden R. Basic interviewing skills. Prospect Heights, Il.: Waveland Press; 1998. p. 1-13.