**Document clustering and Topic modelling for a client project library:**  From the Pcubed SharePoint, we selected a large client, FORD, where we have done several different engagements, leaving a sizeable document trail. We machine-read all the different words in each document, and uncovered clusters of similar documents, document outliers, and the main topics covered.

**Why do this?** This is a first, simple, project-management application of what is called Natural language processing. Machine-learning now allows us to can analyse words as much as we can numbers. This allows us to work with a client to understand whether what is being worked on within project libraries is the same as what Management thinks it is, or the same as what status reports say.

**What we would be trying to achieve**: By asking a machine to understand the details of what is in this SharePoint, we gain an overview of every word written about a client. We want to see clusters of similar documents, which is another way of saying: we want to see what detailed sub-folder structure should we apply, to reflect what is actually in the documents. We also want to find unusual documents, different to all the other ones. We also want to see what topics occur frequently across all documents.

**Data**: I took 422 documents from SharePoint that sit within the various FORD engagements and the GETRAG engagement, which is a supplier to FORD: all the Word documents and pdf files. This can also be applied to Excel, text and PowerPoint files, but it takes more pre-processing. There were German and English text, which is clear in the results.

**Outcome:**  We found three large main document clusters, and three smaller document clusters. We found several document outliers that look different to all the other documents and would be worth special review. When we knew which clusters were of most interest to the client, then we would analyse that cluster in the same way. The model also highlighted the top ten themes across the documentation, and of these, we noticed two themes (1. China and 2. Supplier management) which would be first topics we would explore with a client, if we were checking the health and balance of the work represented by all the documents.

**Conclusion for our business:** As we get familiar with these techniques as a business, we expect to be able to scale up ten thousand and then to a hundred thousand documents. We could apply it to all project documents at a client and so understand their whole portfolio. We will be able to combine machine-reading with our own assessment of the client’s business. It is also a natural first step towards being able to write the first draft of a proposal automatically, and towards having chat-bots that can support customers based upon our own project body of knowledge.

**How we did this….**

**Extract, transform and load the data:** We turned the 422 documents into 422 text documents. We counted how many different words of every type there are in each document.

**Run Unsupervised learning:** For document clusters, we identified how similar each document is to each other, and how different, depending on the count of these words per document. For example, two documents that each include the words “risk, issue, challenge, delay” are likely to be similar documents. In addition, an unsupervised machine learning technique called Hierarchical Dirichlet Analysis was applied, which generates a probabilistic model of what the most common topics or themes are across the documents.

**Apply the model:** For document clusters, we visualised each cluster and inspected which documents are in which cluster. Some documents can be seen as not within any cluster- and these are the document outliers. For topic modelling, we looked at each topic to see which documents and which words are captured per topic. We tuned the model until we got the number of clusters and topics that made sense for a first pass from a human perspective. This would then be worked with the client and machine together, down to as much detail as needed in the areas of interest to the client.