Project Success Prediction This is a demonstration of predicting which projects are eventually judged successful by the World Bank, using their data.

Why The reason why I have done this is to start to understand how to apply machine learning to portfolios of projects. Predicting which projects are more likely to succeed would have a high return both in terms of reduced costs and benefits accruing from more successful projects.

What we would be trying to achieve on behalf of a client like the World Bank, or any client with a lot of projects or a lot of work-packages, is to train a machine-learning model on their archive of projects, and which will then predict which projects are most likely to succeed and which to fail.

Data We started with the World Bank because they have opened their data on 12000 projects from the last fifty years. Each project has been given a final class at the end of the Project: Highly Satisfactory, Satisfactory, Unsatisfactory etc. This is what we care about. We can train a model using supervised machine learning to predict this classification for projects that the model has not seen before.

Outcome For projects that the model had not seen, the model achieved a classification accuracy of 89% . i.e. for every 100 predictions it made, it got 89 right.

Conclusion for business: This experiment shows a good level of success. We believe it is worth exploring further the applications of machine learning for advising companies with large portfolios. There are already many tools that can do this, and this is just one project use case of several which are already possible with the new methods. The methods are getting better every year, faster than our ability to exploit them, and I recommend we accelerate efforts to build a community of practice.

How we did this:

- **Extract, transform and load the data:** We brought the 12000 projects in, and tidied up the data. As well as the final status of the project, each project has about 20 features such as date, duration, cost, and the results of different interim project reviews. This is the raw material that the model can use to understand what combinations of features are most predictive of eventual success.
- **Train the model:** We applied the project data to several machine-learning algorithms that learn to classify the projects into Highly Satisfactory, Satisfactory, etc. We compared the models and chose the best performing model
- **Apply the model:** Once trained, the above steps do not need to be re-run. The model can instead be periodically applied by a someone who has had a day or two of training. It will predict for an unseen project whether the project is likely to ultimately be satisfactory.

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. For this library, 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 that we 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 what the reports say.

What we would be trying to achieve: By asking a machine to understand the details of what is in a document library or a project library, we gain an overview of everything in a library or in a portfolio. We expect to be able to see clusters of similar documents. We want to find unusual documents, different to all the other ones. We want to see what topics occur frequently across the library.

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 within the document library.

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 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 chose the most common 500 words within the documents, and the machine kept track of which words were in which documents.
- **Train the model:** 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 "distance" between documents. For topic modelling, an unsupervised machine learning technique called HDA was used
- **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.