MIDCON Data Services, is an independent records and information management company headquartered in Oklahoma City. They have hired me to find a creative solution on how to manage and govern electronic records. Companies today are creating petabytes of new data, mostly electronic, each year. This has become a cost and logistics problem for many corporations because most do not have an enforceable document management policy for electronic data. This is not because of negligence, but rather ineffective tools on how to govern and regulate the knowledge that is shared electronically today.  
  
One key problem with electronic data is that it can be stored anywhere (personal drive, public drive, usb, email server, etc.) and “titled” in any random way. One solution is to use a common repository like SharePoint, and then enforce a common naming procedure. This solution will work fine for company procedures and official policies, but it is unrealistic to think that solution is scalable.  
  
My idea to solve this problem is to create a document classification engine that will assign “categories” or “tags” to a document regardless of document location or filename. I feel that a tool like this would give flexibility to the user, but also the governing power to apply best practices from physical records management (picture library systems). If this tool is efficient enough, then it could also be used in Discovery work for the professional services industry.  
  
To start, MIDCON will provide physical documents that belong to one of three categories: ‘Operations’, ‘Accounting’, or ‘Legal’. I will use in-house software and equipment to digitize and OCR each record, and then save the output text file to classified folders.

In my code you will see these steps: load data, preprocessing, train/test split, algorithm modeling, and finally result analysis. The preprocessing removes stopwords, whitespace, and reduced inflected words to their stem. The corpus is then sent to a pandas data frame and split into equal halves - training and test sets. To concisely test different combinations of Transformers and Classifiers I used Pipeline and GridSearchCV.

After analysis it was found that both the TfidfTransformer and HashVectorizer performed equally, while the SGDClassifier was superior to SVM. All three combinations provided results over 99% recall and accuracy, which is remarkably successful. To figure out why this is I looked at the top features, frequency of the words, and part of speech in the corpus. In the top 3 words I see an operations token (item), legal token (storage), and accounting token (2017). The accounting term was an oversight on my part since many of the financials will be forward looking and the 2017 term will be often repeated in accounting documents.

The supervised approach has proven to be successful in identifying the three trained classes. The model could be applied as-is or I could add more classifications to make it more commercially operational. It would also be interesting to try a new method, an unsupervised approach, and let the model extract the topics. However, the current model does address the proposed problem, which is to give a records manager a tool to identify and classify documents into subgroups.