**DSC 478**

**Mike Messina and M. Omar Khan**

**Group Project**

**Application Development – Document Analysis of Previously Classified CIA UFO Intelligence Documents**

**Executive Summary:**

***Project Goals:***

The goal of this project was to determine if there is any discernable pattern within the recently declassified documents provided by the CIA regarding Unidentified Flying Objects (“UFOs”). Specifically, there were a total of 2,780 documents that were recently released, each classified by the CIA as they contained potentially sensitive information. Our project was designed to discover if there were any patterns amongst the documents such that they could be classified into categories of similarity. The result would describe those documents that are most likely going to provide pertinent information related to the CIAs intelligence on UFOs. Our assumption is that if the majority of documents end up together in one particular grouping (class) then those documents are the vital documents containing a propensity of the meaningful information. Documents belonging to other classes would have some similarity amongst themselves, but would be of a lower likelihood of containing pertinent information on the CIAs intelligence on UFOs.

***Methodology:***

*Scrubbing Data:*

A number of methods were used in conducting this analysis. All documents obtained were hand scanned and of questionable quality, however we proceeded by extracting the text from each document. We initially wanted to conduct a review of all 2,780 documents, but revised our analysis down to 670 documents as that was more manageable for conducting an analysis of this scale. The documents selected for analysis were done haphazardly to limit any information loss. We replaced known bad characters (i.e. punctuation marks, etc.) that show up in documents as a sort of pre-scrubbing step. Subsequently, a document term matrix was created from the extracted text using the Count Vectorizer module from sklearn. This formed the basis of our data that would then be used to conduct our analysis. Once the dataframe was created we started to filter out those terms we knew were of meaningless value (i.e. did not contain English words, or were a series of numbers that corresponded to no real value). Our final document term matrix contained a total of 7,681 terms across the 670 documents.

*TFxIDF:*

We transformed our data based on term frequency and inverse document frequency, or TFxIDF weights. This allows us to identify those words that are truly “key” terms in the documents. Once this was accomplished we began the clustering process.

*Kmeans Clustering:*

For our clustering methodology we used Kmeans clustering, with a cluster size of 5. We felt that 5 clusters provided enough interpretability without causing too much over fitting of the data. Had we chosen a higher number, for this particular project, we could have ended up with a cluster that contained a large number of pertinent documents that would be ignored because it did not contain a majority of the documents. This initial run through showed us that across all 5 clusters there was indeed one cluster (Cluster 4) that contained a majority of the documents as predicted! We conducted an analysis of the centroids and we identified which terms appeared to have the smallest distances to their centroids. We conducted an analysis of the clustering to obtain the Silhouette Values to determine how well matched each document was to its particular cluster, and noted a mean of 0.78 which indicated that the documents are well matched to their clusters.

*KNN Classification on TFxIDF Matrix:*

We next wanted to conduct a classification exercise, using the 5 nearest neighbors, based on the data we had thus far. We split the data 80/20 between training and test data. Our goal was to see if we could build a classifier that would help identify a documents class and therefore predict whether or not it included pertinent information on CIA intelligence. Our initial review identified that the majority of documents did fall within one specific class. Again this is in line with our prediction from the outset.

*PCA, Kmeans Clustering and KNN Classification:*

Having completed the work thus far, we wanted to verify our results through a separate methodology. Rather than using TFxIDF weights, we conducted a Principal Component Analysis and determined that 4 Principal Components attributed 83% of the variance in our data. Based on this information we conducted another Kmeans clustering exercise, again using 5 clusters, and determined once again that the majority of documents fell into one cluster. We conducted another analysis of the Silhouette Values and noted a mean of 0.89 which exceeded the performance of the analysis done on TFxIDF weights. We next conducted KNN classification again looking for the five nearest neighbors. We used the same 80/20 split as before, and noted the same results of perfect accuracy for our classifier.

***Results:***

We had stated that at the outset our prediction had been that amongst the entirety of all the documents released by the CIA on UFOs only a portion of that documentation would contain any pertinent information. Our goal was to identify those documents specifically and present them as targets for potential future analysis. We conducted Kmeans Clustering and KNN Classification on both TFxIDF and PCA approaches. The end result was inline with our predictions, a large majority of the sampled documents fell into one cluster. Specifically of the 670 documents that were analyzed, a total of 610 fell into one cluster and as such should be deemed relevant for review. A future analysis should be done on the 610 identified documents.