**Project Proposal**

**Application Development – Document analysis of previously classified CIA UFO intelligence**

**DSC 478**

**Mike Messina and Omar Khan**

**Overview**

Since the late 1980s, the CIA had only released about 1,000 pages of classified information regarding unidentified flying objects (UFOs) after a FOIA court case forced the publication of relevant documents. Recently in late December 2020 however due to the Freedom of Information Act, The CIA has declassified upwards of 2,780 documents detailing the agency’s findings on unidentified flying objects (UFOs). The contents of these documents range from information regarding UFOs from private citizen sightings that appeared in the media to internal CIA intelligence communications regarding specific case numbers and research into certain phenomena by divisions within the agency. The CIA has claimed that this is their entire collection but with no way to substantiate that claim, it seems appropriate to try and apply machine learning algorithms to parse the declassified pages looking for any patterns of speech that may indicate additional documents still classified or any signs of the agency attempting to hide or muddy the information included in the declassified documents. Our document analysis also has the possibility of uncovering absolutely nothing and could lean towards confirming that the US Government does not have any hard evidence on the existence of UFOs nor their relation to extraterrestrial beings.

**Analysis Approach**

The documents have already been scanned and converted to searchable files by theblackvault.com who have fought for years with the US government to make classified US agency documents available to the public. Many of the documents are poorly photocopied by hand (most likely on purpose by the CIA) with a number of terms redacted so there is a higher assumed error rate of data collection during the pre-processing stage.

As pure word frequencies are of little importance in this type of analysis, inverse term x document frequency categorization methods will be applied as less used words are assumed to have more importance here than words with the highest frequency in documents talking of specific science and subject matter. As TFxIDF methods can sometimes tune out rarely used words, we’ll also be applying other methods to categorizing the contents of the text so as to not lose the importance of rarer words which could hold value here. As we explore the contents of these documents, we will attempt to cluster the terms and documents in an explainable and simple way using various clustering techniques like KNN and Rocchio methods.

On top of finding pure word frequencies across documents, ensuring we have comprehensive and simple class types to categorize the terms in regard to how they are used in the documents will be important. Once we can analyze the numerical qualities of our terms will we have a better idea of how many and what kind of classes we want to categorize the terms in to. This is important as we could have inaccurate results from applying the Rocchio method if our classes are not well distinguished.

More specifically, sentiment analysis and word embedding will be applied to the CIA documents in an attempt to see if we can create our own readable documents based on the characteristics of the CIA documents in an attempt to uncover what may still be kept hidden by the agency. If we can uncover a common tone and language used by the CIA when producing these records, maybe we can perhaps have a better understanding on what is factual and what is not.

**Data Schema and Size**

The data used for this analysis is contained in 714 .pdf files (~400MB) which contain 2,780 pages of previously classified CIA intelligence regarding UFOs. Number of terms is not yet known.

**Plan for Evaluation**

Our goal for evaluating this dataset is to see if we can identify a clear positive, negative, or neutral tone used by the CIA in an attempt to uncover whether additional information regarding UFOs is being kept hidden or if evidence is being ignored. Perhaps even evaluating if the glorification of nothing is present and the content is all pure speculation.

We will evaluate our application of word embedding by attempting to produce and write our own readable documents based on the content already produced by the CIA. Perhaps we can put pieces to the puzzle of truth together by creating what *should* be there with what *isn’t* there.

**Plan for Work Distribution**

We are planning on working concurrently together through the project so as to minimize user error and maximize the viewpoints of how to build certain models and categorize the data. Multiple models of each algorithm are expected to be produced by each individual with marginal changes so we can have comparisons of different ways of applying the same application to better determine fit and function to our final goal.