David Kalamarides

Final Project Initial Analysis

**Initial Description:**

Rather than limit myself to one text, I would like to create an extensive corpus of ancient religious documents on which I will train a machine learning classifier using sk-learn. This classifier can then be used to examine anonymous epistles, books, and writings both in the bible and the apocrypha. The end goals will be to find similarities between writings to identify the authors of these texts as well as confirm if the Pentateuch is indeed by a singular author or a variety.

**Amended Description:**

After initial analysis, I discovered that my attempts to identify authors were misguided. To accurately do so, I would need the original text and examine the author’s handwriting style. Instead, what I as classifying was the subject matter of various texts. Therefore, rather than restart the text analysis, I have re-examined the end result. I am now seeking to cluster bible passages based on similarity and in doing so identifying key themes throughout the bible as well as which passages best represent each.

**Process:**

**Part I: Passage Clustering**

* Corpus text is taken from every passage of the complete World English Bible
* This text is processed through a Count Vectorizer and a Tfidf Transformer
* The sparse matrixes are then analyzed by a K-Means Clusterer

**Part II: Key passages Identification**

* Passages are separated into clusters
* Key Central Passages: those with the highest cosine similarity rating to other passages in the clusters
* Key Linking Passages: those with the highest cosine similarity between multiple clusters *(still need to do)*

**Potential Changes:**

* K-means algorithm does function well but requires a specified number of clusters. This can be resolved by optimizing the K-means clusters or using another soft clustering method
* I began with a Tfidf Transformer to avoid highly used words, but it may be better to use a comprehensive list of stop words to avoid removing ubiquitous themes in the Bible