The Journal of Business Inquiry 2014, 13, Issue 1 (Special Issue), 27-47
http:www.uvu.edu/woodbury/jbi/articles/
ISSN 2155-4072

A Text Mining Analysis of Religious Texts

By DANIEL MCDONALD[*]

Religious text scholarship explores the meaning of passages and uses critical/rhetorical research methods. In contrast, automated tools that perform shallower but broader quantitative analysis have been created. These tools process entire books and help illuminate relationships between religious texts. We have automatically extracted and categorized noun and verb phrases from nine religious texts: the Book of Mormon, the Greater Holy Assembly, the New Testament, the Old Testament, the Popol Vuh, the Qur'an, the Rig Veda, the Tao Te Ching, and the Torah. The extracted topics were used as input to a Self-Organizing Map (SOM). The map uncovered some interesting relationships.

Keywords: Religious Text Analysis, Self-Organizing Map

JEL Classification: C38

I. Introduction

The research of religious texts is often performed by literary and philosophy scholars and those trained in schools of Divinity. The research uncovers insights into religious passages that go deeper than words or phrases. Such focus on the meaning of scripture passages produces insights into the beliefs and practices of religions. Such research often uses qualitative and critical/rhetorical research methods. Such methodologies are appropriate as passages may have multiple levels of meaning and symbolism and do not fall into the often flattened categories of quantitative research.

In contrast to a deep understanding of scriptural passages is the automated lexical analysis that is used in information retrieval systems of today. Words and phrases are extracted from documents and indexed to facilitate searching. Users of search tools enter a few words and phrases and similar documents are retrieved. While such tools are constantly improving, there is still a large difference between the rich analysis done by religious scholars enabled by critical/rhetorical research methods and that done automatically by computers.

However, computer analysis of text is very fast. As a result, entire books can be processed producing quantitative data that can be analyzed. While the analysis is not as rich or deep, it is broader and can be revealing in its breadth and its quantitative nature. Over the last two decades, automatic text processing has improved in part because of the growth in availability of digitized text. In this research, we look to apply techniques used in the analysis of business and medical texts to religious texts. Our goal is to explore similarities and differences between nine religious texts based on our automatic processing of the text using current methods in text mining.

[*] Assistant Professor of Information Systems and Technology, Utah Valley University, 800 West University Parkway, Orem, UT 84058. E-mail: dmcdonald@uvu.edu.

The author would like to acknowledge those from the Religious Faith and Social and Applied Sciences Conference whose comments helped improve this work.

II. Literature Review

Digitized natural language texts in the form of research, email, web pages, and digitized books have proliferated greatly over the last decade. The growth of the World Wide Web has been a catalyst for such growth. In 2008, Alpert and Hajaj (2008) reported that Google had counted one trillion unique URLs on the Web. The U.S. National Library of Medicine (NLM) posts all medical research abstracts on the Web. The NLM currently reports having 19 million abstracts, with between 2,000-4,000 added daily (Medicine, 2013). Project Guttenberg, which was founded in 1971 by Michael Hart, supplies access (via the Web) to previously published books that are now out of copyright. The project site reports offering 43,000 eBooks for free download (Gutenberg, 2013). Started in 1999, the Internet Sacred Text Archive is a "text archive of electronic texts about religion, mythology, legends and folklore, and occult and esoteric topics" (Hare, 2013). The site is not sponsored by a religion and seeks to support religious tolerance. The sacred text archive includes over 900 references in its bibliography.

For a researcher or executive, staying on top of new publications or synthesizing information relevant for decision making is a challenging task. This deluge of information is referred to as the information overload problem (Bowman *et al.*, 1994). Language processing and analysis techniques have been developed to facilitate gleaning knowledge or at least highlighting relevant information from text. Applying such automated techniques of text analysis to cultural and religious texts creates an opportunity to produce new insights into cultures and religions as well as cultivate tolerance.

A. *Stylometry*

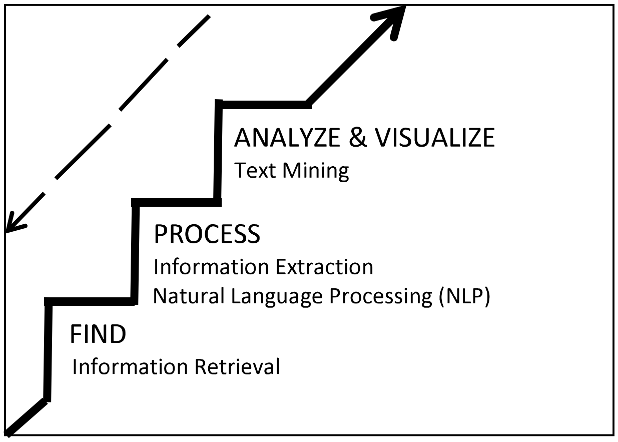
Over the years, applications of text analysis have varied. A common application of text analysis is stylometry. Stylometry does not analyze text content, but rather uses statistics to analyze writing style (Zheng *et al.*, 2006). This research area, also called authorship analysis, has been applied early on to literature such as the Federalist Papers (Mosteller, 1964) and also to the writings of Shakespeare (Hope, 2009). Stylometry has also been applied to email (de Vel *et al.*, 2001), online forums (Zheng *et al.*, 2006), and computer code (Gray *et al.*, 1997). Use of stylometry to analyze religious texts, in particular, the Book of Mormon has also been common (Jockers *et al.*, 2008; Reynolds, 1982 and 1997, Reynolds and Tate, 1982).

B. Text Mining

In contrast to stylometry, a research area called text mining focuses more on text content analysis and seeks to help alleviate the information overload problem (Fan *et al.*, 2006). Text mining includes many computer-automated tools and techniques. Figure 1 shows different focus areas of text mining, which include the find, process, and analyze and visualize areas. Underlying each area is the goal of uncovering relevant and timely information or insights. A starting point in text analysis is to find relevant document(s) to support a task (Salton *et al.*, 1975). This task may be broad research, an information search task, searching email for relevant forensic data, or increasing understanding among cultures and religions. Depending on the user task, the resulting set of documents may be large or small. Higher steps depicted in Figure 1 can provide feedback to lower steps, so the process operates more like a cycle than a one-way linear staircase. For example, finding the correct documents can be facilitated through analysis and visualization.

In the process area of text mining, automated tools can identify proper nouns, identify entity relationships, or extract events (DARPA 1993a, 1993b, 1995, 1998). The structures extracted can inform the "find" task or be used as input to "analyze and visualize" tasks. Finally, in the "analyze and visualize" area, text or extracted structures such as events can be aggregated and displayed using different visual metaphors (Hearst, 1999, and Pirolli *et al.*, 2001). Such analysis and visualization metaphors can assist "find" tasks or be used to help support drawing conclusions or making decisions. Text mining analysis reveals relationships found in single or multiple textual documents. Applying text mining to religious texts can yield cultural and religious insights as content relationships among religious texts are explored.

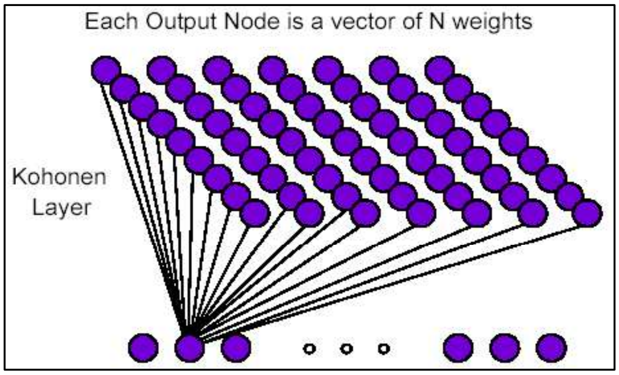
Figure 1: The Text Mining Research Areas



C. Self-Organizing Map

A self-organizing map (SOM) is a two-layered neural network algorithm used for clustering and dimension reduction. The SOM, developed by Teuvo Kohonen, is unsupervised in that it does not require human intervention (Kohonen, 1995). The SOM is similar to multi-dimensional scaling in that it takes a multi-dimensional input layer and maps the inputs to a two-dimensional output layer (Jain and Dubes, 1988). Figure 2 shows the SOM topology. In dealing with text,

Figure 2: The SOM Topology



each input is typically a word or phrase, which we will call a term. The number of inputs corresponds to a union of unique terms from the documents to be clustered. The value for an individual input is either 1 or 0 depending on whether a particular document contains the input term. Each document to be clustered arranged as a set of inputs, is presented as a sequence of 1s and 0s, depending on the existence of the given term in the document. The number of output nodes can vary based on configuration. The topology in Figure 2 shows 63 output nodes. Also shown in Figure 2 is that the network is completely connected meaning each input node is connected to every output node. Each output node contains a vector of weights that correspond to each input node. During the training phase of the algorithm, sets of inputs are presented multiple times in order to tune the connection weights. At the end of a training iteration, an output node is selected with the smallest distance from the input set. The weight vectors of the selected output node along with neighboring output nodes are adjusted to further decrease the distance to the input set. Slowly the difference between the output nodes begins to correspond to the difference between sets of inputs. Similar sets of inputs will be placed closer together on the output map with less similar inputs being placed farther apart (Lin *et al.*, 1999).

The SOM has been used extensively to assist in text mining search tasks. The graphical map has been used to cluster similar documents to support information seeking (Kaski *et al.*, 1993). Lin *et al.* (1991) used the SOM as a retrieval interface for an online bibliographic system. Chen *et al.* (1996) used the SOM to categorize Web pages. Kaski *et al.* (1998) in their WEBSOM system used the SOM for categorizing over one million documents from 85 Usenet newsgroups.

Beyond supporting the finding of documents in a collection, the visualization power of the SOM has also been used to analyze textual content. Orwig *et al.* (1997) used the SOM to classify electronic brainstorming output from an Electronic Meeting System called GroupSystems (Orwig *et al.*, 1997). The SOM was used to facilitate group problem solving. Roussinov and Chen (1999) also analyzed the output of electronic meetings, comparing the output of the SOM along with Ward's clustering algorithm. In a more literary application, Honkela *et al.* (1995) used the SOM to visualize the raw text from 200 Grimm Tales. The map was used to highlight the relationships between words from the stories based on their proximity of placement on the map.

III. Research Gap

The primary research gap we wish to explore in this paper is the lack of automated analysis and visualization of religious text content. Text from non-religious topics has been visualized using the SOM. Examples in research include text from web pages, electronic meetings, online bibliographic systems, and even popular stories. We are not aware of studies using the SOM to visually compare content solely from religious texts.

In addition, we want to address a gap in the way text is represented when used with a self-organizing map. When the self-organizing map is used to cluster documents, the input nodes are 1s and 0s that represent the existence of terms (words or phrases) extracted from the text itself. For example, Honkela *et al.* (1995) originally identified 7,000 unique words from the Grimm Tales and ultimately reduced the vocabulary down to 270 to use as inputs to the SOM, a 96 percent reduction. While such an approach works well to find contextual relationships between the terms that were kept as inputs, reducing the inputs in this fashion greatly hides differences between the documents from which the terms came. The dimensionality of the inputs has been reduced by simply using the words with the highest frequency. In this paper we introduce a technique for reducing dimensionality by preprocessing text into word categories based on the semantic similarity as opposed to eliminating words with lower frequency.

Finally, there is a research gap in the way the SOM is being used. The typical application of the SOM identifies themes, topics, or word contexts across documents. We are more interested in highlighting the similarity and differences between entire books. For example, electronic meetings or web pages were processed to identify themes or topics and the Grimm Tales were processed to identify words with similar contexts across tales. The document boundaries were not as important as the topic or category boundaries that resulted. We are using the SOM to see how closely religious texts cluster to other religious texts and not to identify common themes or contexts among the texts.

IV. Research Questions

The purpose of this research is to use automated techniques to perform quantitative analysis on religious texts. We explore the use of the SOM clustering algorithm to cluster seven varied religious texts based on their topic similarity. Instead of using terms as inputs, each religious text will be preprocessed into topic categories and those categories will serve as inputs to the SOM. Also, different from previous research, the purpose of the SOM clustering will be to analyze the cluster position of entire religious texts as opposed to identifying topic categories from within those texts. We have four primary questions that we want to explore.

1. Can any interesting observations be drawn from the resulting visualization of the sacred texts?
2. Does the location of documents on the output layer of the SOM correspond to word category overlapping among religious texts?
3. Which religious texts actually cluster together?
4. Does the clustering of texts using noun categories vary from the clustering created using verb categories?

V. Experimental Design

The experimental design consisted of selecting the sacred texts to include in the experiment, processing the texts by extracting the noun and verb phrases and placing them in semantic categories, selecting cutoffs for the number of topic categories to include as inputs from each book, and then producing the SOMs.

A. Selecting the Text

The nine religious texts were downloaded from the Internet Sacred Text Archive (Hare, 2013). The books processed included the Book of Mormon (BOM), the Greater Holy Assembly (GHA), the King James Version of the New Testament (KJV-NT), the King James Version of the Old Testament (KJV-OT) the Popol Vuh (PV), the Qur'an (Q), the Rig Veda (RV), the Tao Te Ching (TTC), and the Torah (T). In selecting the books, we aimed to use books that came from different parts of the world and are part of different religious traditions. We chose books from the following four geographic areas:

Americas—The Book of Mormon and the Popol Vuh both have origins on the American continent. The Book of Mormon is the scripture of the Church of Jesus Christ of Latter Day Saints

(LDS). The book was originally published in English in 1830 CE. The book claims to be a translation of ancient writings from inhabitants of the American continent with the translator being the religion's first modern-day prophet Joseph Smith. Joseph Smith reported the original language of the book to be Reformed Egyptian, though there is no surviving original record.

The Popol Vuh is a sacred text of the Quiche Indians in Central and South America. The Popol Vuh was originally written in Quiche. The oldest surviving written account is from 1701 thanks to the Spanish 18th century Dominican Friar Francisco Ximénez. A major reference of the Popol Vuh is the one translated by Adrián Recino into Spanish in 1947, with a subsequent translation into English in 1954 by Delia Goetz & Sylvanus Griswold Morley. It is this translation that we use in our text analysis.

Middle East—The Old Testament, the New Testament, the Greater Holy Assembly, the Torah, and the Qur'an all have origins in the Middle East. The Old Testament and the New Testament are sacred to religions of a Christian denomination. We used the King James Version of the Old Testament and the New Testament also called the Authorized Version published in English in 1611. This version was the third translation to English and was commissioned by King James VI, later known as King James I after the union of the Scottish and English crowns. The King James Version was conceived to fix reported errors in earlier translations and is a popular version used in Christianity today. According to *The Christian Post*, based on Bible sales in the United States in 2013 through September, the King James Version was the second most purchased version of the Bible after the New International Version (NIV) which was first published in 1978 (Menzie, 2013).

The Greater Holy Assembly is a sacred book of Kabbalah originally written in Aramaic and first appeared in Spain in the 13th century. Kabbalah is a school of thought that originated in Judaism and is considered mystical. While its teachings are used by many religious denominations, it is not a religious denomination itself. The Greater Holy Assembly is one of three books from a collection called the Zohar. The other two books of the Zohar are The Lesser Holy Assembly and The Book of Concealed Mystery. The Zohar was published by a Jewish writer named Moses de Leon. De Leon credited the work to Shimon bar Yochai, a 2nd century rabbi who was inspired by the Prophet Elija to write the Zohar. Adherents of Kabbalah claim the Zohar is the concealed part of the Oral Torah. The version we use in our text analysis is the first translation to English. It was translated by S. L. Macgregor Mathers in 1912 from the Latin version translation by Christian Knorr Von Rosenroth in 1684.

The Qur'an is a sacred text of Islam. The book contains revelations from God, given to Muhammad through the angel Gabriel. The revelations occurred up to the year 632 CE, which is the year of Muhammad's death. After the death of Muhammad, the first caliph Abu Bakr, with the help of Zayd ibn Thabit collected the book into one volume so it could be preserved. The original language of the Qur'an is Arabic and many argue that it cannot be reproduced in another language. The English version we use was translated from Arabic to English by Mohammed Marmaduke Pickthall in 1930. Pickthall, an Islamic scholar, had converted to Islam in 1917.

The Torah is central to Judaism. The Torah specifically means the first five books of the Tanakh or the Five Books of Moses. In some cases, the term Torah can also include the Oral Torah, which contains interpretations and amplifications. Revelations in the Torah are believed to have been given by God to Moses. The main revelatory event occurred on Mount Sinai. Some believe others revelations occurred at the Tabernacle. We used an English version of the Torah which did not include the Oral Torah. Our English translation was done by the Jewish Publication Society of America (JPS), which was first available in 1917.

India—There are four Vedas that originated in India, the Rig Veda, the Sama Veda, the Yajur Veda, and the Artharva Veda. The Rig Veda is one of the canonical sacred texts of Hinduism. The Rig Veda is a book of hymns composed by rishi that are dedicated to various deities. The content of the Rig Veda is accepted as originating between 1700-1100 BCE, being written in an early Indo-Aryan language. The oldest existing Rig Veda manuscript is from the 14th century CE and is kept at The Benares Sanskrit University. The first translation of the Rig Veda to a Western language was into Latin in 1830. The version we use was the second translation into English and was done by Ralph T.H. Griffith in 1896.

China—The Tao Te Ching originated in China. The Tao Te Ching is fundamental to religious Taoism. The book was written by the sage Lao-Tse (or Laozi) around the 6th century BCE. The oldest excavated text dates back to the 4th century BCE. The book was originally written in Classical Chinese using seal script, an ancient style of Chinese calligraphy. The Tao Te Ching has been translated many times to Western languages. The version we use was translated into English by James Legge in 1891.

We hypothesized that the book topic categories would cluster together based on geographic origins of the text. We supposed that regional speaking and writing influences would tend to make books from similar regions overlap.

B. Processing the Text

After selecting the books for the experiment, the texts were then formatted into XML files. Each XML file was processed by our content tagging algorithm (McDonald *et al.*, 2004). The algorithm tokenizes the document and recognizes sentence boundaries. The tokenizing process separates hyphenated words, adds spaces to punctuation, recognizes abbreviations, and matches parenthesis and quotations. Once tokenized, a document’s words and phrases are tagged using hybrid semantic/syntax tags. The tagging process is aided by the use of an extensive dictionary of approximately one million word-tag entries. After being tagged, the text is processed several times more to combine tags into topic categories. The topic categories are part of a large category hierarchy. Nouns and verbs are placed into different categories. There are just over 3,400 different categories in the hierarchy into which terms (words or phrases) can be assigned. Once a sentence is processed, the instances of terms in topic categories are summed. The result is a list of topic categories with count totals.

In Table 1 we report the total breakdown of terms into nouns, verbs, and other categories between the nine books. Nouns typically made up 17 to 24 percent of terms, while verbs made up 13 to 17 percent of terms. The most common terms from books were neither nouns nor verbs however, which accounted for around 40 percent of terms, but rather prepositions. As shown in the table, the Popol Vuh, the Old Testament, the Torah, and the Rig Veda had a higher percentage of nouns than did the other texts. Subsequently, those four books were offset with a smaller percentage of verbs than the other texts. After combining nouns and verbs together, The Greater Holy Assembly had the lowest combination percentage. The length of the texts vary as well, with the Old Testament being the largest, followed by the Book of Mormon, the New Testament, the Qur’an, the Torah, the Popol Vuh, the Rig Veda, the Greater Holy Assembly, and the Tao Te Ching. Several texts were also very evenly distributed between nouns and verbs. The Qur’an, the Tao Te Ching, and the New Testament all had a similar number of nouns and verbs.

Table 1: Tallies for Content Categories Across Religious Books

	Nouns	%	Verbs	%	Other	Total
Book of Mormon (BOM)	48,613	18.57	42,492	16.23	170,742	261,847
Greater Holy Assembly (GHA)	8,666	18.94	6,347	13.87	30,744	45,757
New Testament (KJV-NT)	33,200	17.73	31,915	17.04	122,151	187,266
Old Testament (KJV-OT)	143,089	**21.53**	98,975	**14.90**	422,405	664,469
Popol Vuh (PV)	18,736	**24.18**	10,935	**14.11**	47,829	77,500
Qur'an (Q)	27,479	16.46	28,443	17.04	110,989	166,911
Rig Veda (RV)	13,018	**20.49**	8,775	**13.81**	41,734	63,527
Tao Te Ching (TTC)	1,931	17.62	1,907	17.40	7,124	10,962
Torah (T)	34,576	**20.90**	23,343	**14.11**	107,521	165,440

C. Selecting Topic Categories as Inputs

A critical experimental design issue was how to decide on the number of topic categories to include as inputs into the analysis. In order to keep dimensionality to a minimum and not overweight rarely occurring topic categories, we decided to select the most commonly occurring topic categories based on term counts. We sorted the topic categories for each book in descending order based on the frequency of the member terms.

When the term running total summed to 85 percent of the total terms for a book, we used that topic category as the cutoff for the cluster inputs. Using this approach, we were able to represent 85 percent of all term instances from a book to place the book on the SOM. At the same time, we kept the total number of noun inputs to 179 and the total number of verb inputs to 86. No individual book had more than 108 noun topic categories or more than 66 verb topic categories. This approach varies from the typical strategy of including actual terms as inputs and thus not being able to represent nearly as many topics from a book. Table 2 shows the total noun and verb topic categories per book and the number of actual topic categories used in order to achieve the 85 percent coverage of book term instances. Being able to represent 85 percent of the terms from a book and at the same time keeping inputs to below 180 is an innovation of our work.

Table 2: Topic Categories Selected as Inputs

	Total Noun Categories	Categories to cover 85% of nouns	% of Total Noun Categories	Total Verb Categories	Categories to cover 85% of verbs	% of Total Verb Categories
Book of Mormon (BOM)	406	84	20.69	220	50	22.73
Greater Holy Assembly (GHA)	331	75	22.66	172	38	22.09
New Testament (KJV-NT)	400	78	19.50	219	46	21.00
Old Testament (KJV-OT)	448	66	14.73	230	50	21.74
Popol Vuh (PV)	435	76	17.47	213	53	24.88
Qur'an (Q)	432	85	19.68	234	47	20.09
Rig Veda (RV)	350	65	18.57	212	66	31.13
Tao Te Ching (TTC)	270	108	40.00	155	47	30.32
Torah (T)	339	59	17.40	200	43	21.50

In most cases, we were able to represent 85 percent of the term instances while using under 23 percent of the topic categories. An exception to that rule was the Tao Te Ching. Forty percent of its noun categories were required to cover 85 percent of the book’s nouns. Similarly, 30 percent of its verb categories were required to cover 85 percent of the book’s verbs. Also, 31 percent of the Rig Veda verb categories were required to cover 85 percent of the books verbs. These results would indicate a higher variety of verb usage in the Tao Te Ching and Rig Veda compared to the other sacred texts. The Tao Te Ching was the shortest text in the study. The greater diversity of topic categories could also relate to the shorter length of the text. Longer books have more opportunity to revisit the same topics. The Rig Veda was also one of the three shortest books.

Once the topic categories for each book were identified, we removed the topics that were common to all nine sacred books. We wanted to include only information that would help differentiate the books. There were 14 noun topic categories that were common to all the books. Those 14 removed categories included animals, date/time references, external body parts, family relationships, emotions, references to a group, types of geography, references to God, internal body parts, plants, positions, the word thing(s), different roles, and uncategorized nouns. There were 16 verb categories that were shared by all 9 sacred books. The removed verb categories included the following: amuse, appear, be, build, characterize, conjecture, directed motion, do, future, get, give, have, message transfer, put, say, and see.

D. Creating the SOM

The SOM was created using free software developed by Cao Thang called Spice-SOM version 2.1 (Thang, 2011). The training of the SOM used a learning rate of .01 and a Sigma of 2 with a Sigma decreasing rate of .01. We ran the training for 1000 iterations. The output map is a 20 neuron by 20 neuron map using a hexagonal topology. Because each neuron on the map has six sides, the hexagonal topology allows for more neighbor neurons. We used a Gaussian neighborhood function to change the weights of the neurons after each iteration.

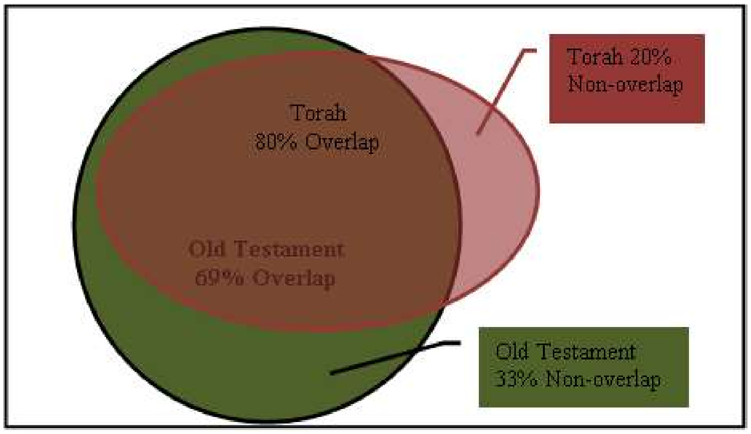
VI. Results

We created maps from two different sets of data. In the first map, only noun topic categories were used as inputs. In the second map, only verb topic categories were used as inputs. We have therefore separated the results of our analysis by noun and verb topic analysis.

A. Noun Topic Analysis

The self-organizing maps are created in part based on total topic overlap between books. Books with more topics in common should appear closer on the map. In order to get an idea of noun topic overlap between books, we organized the data in Table 3. Table 3 shows the percent of all the topics from each row that overlap with each column text. The overlap was calculated by taking the total topic matches between books (rows and columns) and then dividing that total by the number of topics in the first book listed in the comparison (the row). All the books had a different number of topics, though the Greater Holy Assembly and the Popol Vuh topic counts were close. Overlap percentage is thus not a commutative calculation. Because books had a different number of topics, the Torah's overlap percentage with the Old Testament, for example, is different than the Old Testament's overlap percentage with the Torah. Eighty percent of the Torah's topics overlapped with the Old Testament. Sixty-nine percent of the Old Testament's topics, however, overlapped with the Torah. This relationship is shown in Figure 3. While the overlapping categories are the same, the overlapping category count represents a different percentage of the sacred texts in the comparison. The difference in overlap percent results from the difference in topic counts from each book. Thirty-six categories overlapped between the Torah and the Old Testament. The Torah, however, had 45 topic categories (80 percent overlap), while the Old Testament had 52 topic categories (69 percent overlap). As a result, the Torah has a higher overlap percent with the Old Testament.

Figure 3: Percentage of Category Overlap Calculation



As shown in Table 3, the greatest topic overlap percentage (80 percent) takes place between the Old Testament and the Torah. So, on the self-organizing map, we would expect these two books to appear close together. The Book of Mormon also has high overlap (71 percent) with the New Testament and with the Qur'an (64 percent). Again, we would expect these books to appear close together. By looking at the "average overlap percent of other texts" line at the bottom of Table 3, we can see which books should appear near the center of the map. In other words, to what book are most other books related. The sacred books are most similar to the Book of Mormon, followed by the New Testament and the Qur'an. The books show an average of a 59 percent topic overlap with the Book of Mormon, a 57 percent topic overlap with the New Testament, followed by a 55 percent overlap with the Qur'an. In the SOM, we would expect these three books to appear a bit more central on the map, with the other books being the closest to these three books.

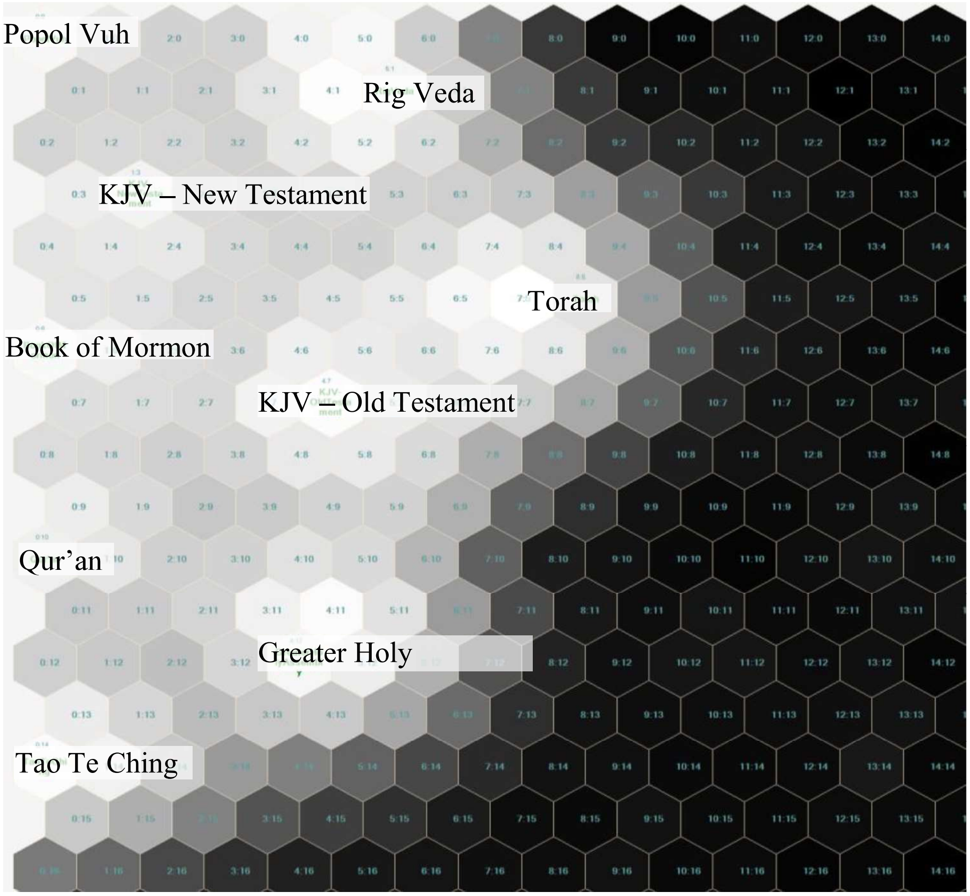
Table 3: Overlap of Noun Categories Between Books

TAG SOURCE	BOM	GHA	KJV-NT	KJV-OT	PV	Q	RV	TTC	T	AVG
Book of Mormon (BOM)	100%	39%	***71%***	51%	43%	***64%***	41%	49%	44%	50%
Greater Holy Assembly (GHA)	44%	100%	54%	39%	44%	51%	41%	46%	31%	44%
New Testament (KJV-NT)	***78%***	52%	100%	53%	44%	***63%***	44%	50%	50%	54%
Old Testament (KJV-OT)	***69%***	46%	65%	100%	54%	***62%***	50%	50%	***69%***	58%
Popol Vuh (PV)	48%	44%	45%	45%	100%	44%	39%	42%	47%	44%
Qur'an (Q)	***63%***	44%	56%	45%	38%	100%	45%	46%	35%	47%
Rig Veda (RV)	57%	49%	55%	51%	47%	***63%***	100%	49%	37%	51%
Tao Te Ching (TTC)	40%	33%	38%	31%	31%	39%	29%	100%	27%	33%
Torah (T)	***69%***	42%	***71%***	***80%***	***64%***	56%	42%	51%	100%	59%
Average Overlap % of Other Texts	***59%***	43%	***57%***	49%	46%	***55%***	41%	48%	43%	49%

Figure 4 shows the actual SOM produced based on the noun topics. Based on the overlap calculations from Table 3, the sacred texts on average were most similar to the Book of Mormon (BOM), followed by the New Testament (KJV-NT), the Qur'an (Q), and then the Old Testament (KJV-OT). The SOM is consistent with our overlap calculations in its placement of those four sacred texts in the center area of the SOM. Specifically, it placed the Book of Mormon vertically right in the middle of the SOM, with four books above and four books below it. The Book of Mormon was most similar to the New Testament, which appears right above it and the Qur'an, which appears right below it, followed by the Old Testament, which appears next to it on the right.

The greatest topic overlap between any two books from Table 3 occurred between the Old Testament and the Torah. This close relationship did manifest itself on the SOM by placing the two books right next to each other. The Torah was next most similar to the New Testament and then the Book of Mormon, which were also placed in a cluster close to the Torah.

Figure 4: Self Organizing Map of Nine Religious Texts Using 179 Noun Category Inputs



The proximity on the SOM shown in Figure 4 between the Rig Veda and the Torah requires some explanation. Based on the overlap percentages shown in Table 3, the Rig Veda is tied with the Greater Holy Assembly for being the least similar to the Torah. On the SOM, however, the Torah is placed close to the Rig Veda. A partial explanation for this is that the Rig Veda shares a unique topic category with the Torah while the Greater Holy Assembly does not. The topic uniquely shared between the Torah and Rig Veda is called "container" and includes words from the Torah including sack(s), basket, laver, and sheath. The "container" words from the Rig Veda include draught(s), jars, chalice(s), and beakers.

The book that on average is the least similar to other books is the Tao Te Ching, showing an average of 33 percent topic overlap with other books. On the SOM, the Tao Te Ching was placed at the very bottom, which is consistent with its limited overlap. While still relatively low, the Tao Te Ching has the most topic overlap with the Book of Mormon and then the Qur'an. On the SOM, the Book of Mormon and the Qur'an appear relatively close to the Tao Te Ching, but so does the Greater Holy Assembly, which shares fewer topic categories. Again the explanation lies in the number of unique topic categories shared between the Tao Te Ching and the Greater Holy Assembly. The three categories uniquely shared between the books include the "form", "issue", and "purpose" topic categories. The "form" category includes largely the term form with different modifiers. The "issue" category includes variations of issue, but also matter and account. The "purpose" category includes variations of the words reason and purpose.

B. Verb Topic Analysis

Table 4 shows the overlap in verb topic categories. Similar to Table 3, the overlap percentage calculation is not commutative. Overall, Table 4 reveals a slightly higher overlap percentage of verb topics (50 percent) compared to the noun topic overlap percentages (49 percent from Table 3). Also, the verb categories show a much greater range of overlap percentages. Noun overlap ranges from 29 percent to 80 percent. Average verb overlap (Table 4), on the other hand, ranges from 17 percent to 96 percent. There are more verb topics shared among all sacred texts than there are noun categories that are shared. We removed 16 verb categories that are shared by all the books in our sample compared to the 14 noun categories that are shared by all the books. Table 5 lists topics that overlap between books and includes some example terms that are members of each topic. The nouns and verb topics are listed in alphabetic order. Related to the verbs having more topic overlap is the lack of topic diversity in the verb topics. The verb-based SOM needed only 86 topics to cover 85 percent of all verbs, while the noun model required 179 different topics. Religious documents repeat verbs more frequently than they do nouns. The verbs to be, to build, to do, and to have all appear frequently in all documents.

Another possible factor that impacts the verb topic overlap is simply the lower number of verbs in the documents compared to nouns. While the ratio of nouns to verbs is similar in three books as shown in Table 1, there are more noun topics than verb topics in 8 of the 9 books. While verbs clearly play different roles than nouns, the greater noun frequency may have had an impact on the number of noun categories present.

Table 4: Overlap of Verb Categories Between Books

TAG SOURCE	BOM	GHA	KJV-NT	KJV-OT	PV	Q	RV	TTC	T	AVG
Book of Mormon (BOM)	100%	30%	***70%***	***70%***	58%	52%	61%	45%	55%	55%
Greater Holy Assembly (GHA)	48%	100%	43%	38%	52%	24%	33%	33%	38%	39%
New Testament (KJV-NT)	***79%***	31%	100%	***66%***	***66%***	62%	59%	59%	52%	59%
Old Testament (KJV-OT)	***70%***	24%	58%	100%	55%	55%	***67%***	42%	***76%***	56%
Popol Vuh (PV)	53%	31%	53%	50%	100%	42%	53%	50%	44%	47%
Qur'an (Q)	57%	17%	60%	60%	50%	100%	50%	***67%***	43%	50%
Rig Veda (RV)	41%	14%	35%	45%	39%	31%	100%	31%	37%	34%
Tao Te Ching (TTC)	50%	23%	57%	47%	60%	***67%***	50%	100%	37%	49%
Torah (T)	***69%***	31%	58%	***96%***	62%	50%	***69%***	42%	100%	60%
Average Overlap % of Other Texts	**58%**	25%	54%	**59%**	55%	48%	55%	46%	48%	50%

Also interesting in Table 4 is the similarities and differences in religious book centrality when using verb topics compared to noun topics. When comparing overlap using noun topics, the sample of religious books was most similar to the Book of Mormon and the Qur'an. However, when comparing topic overlap using verb categories, the sample of books was most similar to the Old Testament and the Book of Mormon, followed by the Popol Vuh and the Rig Veda. In other words, these four books had their verb topics most shared by the other sample religious books. Only the Book of Mormon has high numbers in both noun and verb overlap categories. The average overlap percentage numbers to the Old Testament are a bit misleading. The average is high largely due to its overlap with the Torah. As shown in Table 4, the greatest similarity of all the texts is the Torah's similarity to the Old Testament, the same as with our noun analysis.

Table 5: Topic Categories Common Across All Sample Religious Books

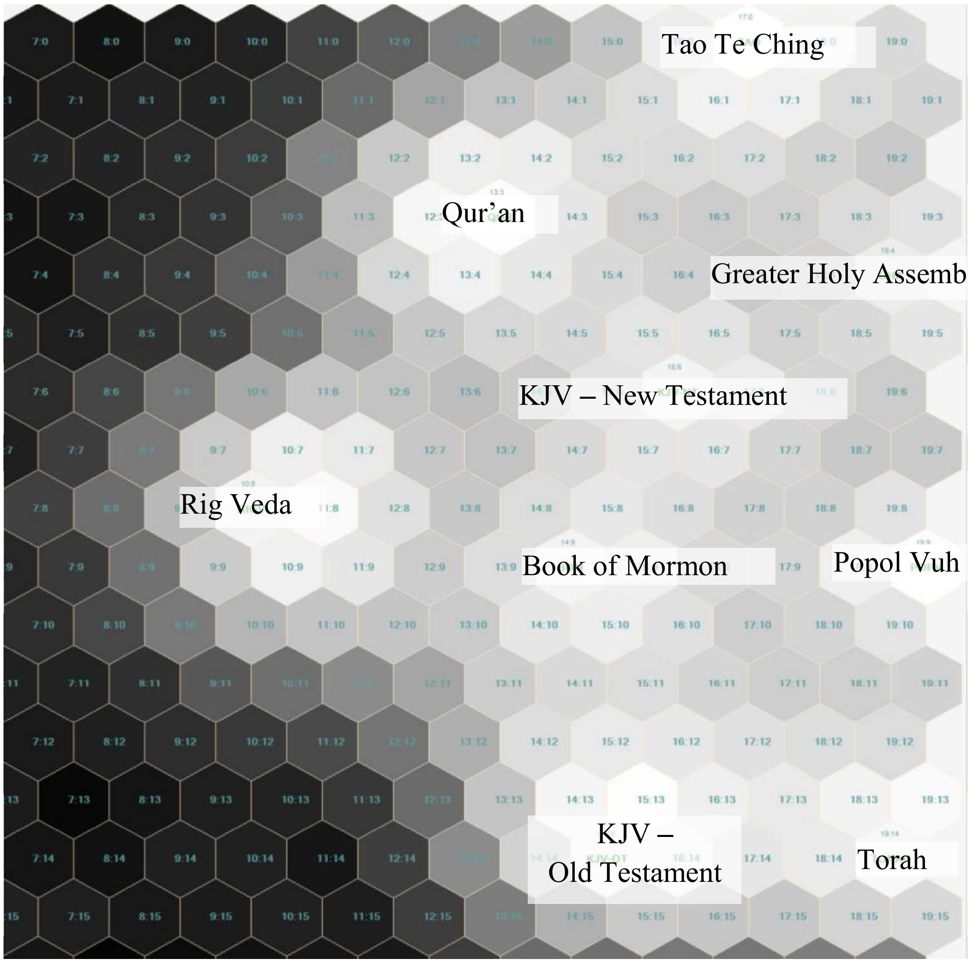
Noun Category	Examples
Animals	flocks, sheep, creatures, beast(s), dragon, fish(es), cattle, oxen, bird(s), deer, jaguar, kine, dogs, rhinoceros, tiger, bullock, lamb(s), swine
Date/Time	first year, night, day, morrow, today, time, Sabbath, hour, dawn
Earth	earth, world
Emotions	joy, fear, pride, anger, wrath, love, laughter, bliss, pleasure, happiness
External body parts	hand(s), face, eye(s), hair, beard, mouth, flesh, body(ies), ear(s), arm(s)
Family relationships	father(s), son(s), daughter(s), children, brother, grandmother, mother
Geography	land, wilderness, vineyard, waters, field, town(s), sky, garden, sea
God	Lord, God, Allah, Christ, Tetragrammaton, Ancient of Days, Elohim, Jesus
Group	mankind, branches, secret combinations, members, cast, council
Internal body parts	heart, bowels, womb, brain, throat, organs, bones, kidneys
Plants	tree(s), timbers, thorn(s), briers, gourd(s), grass, crops, flower, cedars, roses, lilies, reed
Positions	left, side, ground, midst, top, center, corner, right side
Roles	brethern, king, man, inferiors, disciples, lords, hero(s), sage, woman, children, boys, messenger, folk
Thing	thing(s), all things, everything, sacred things, remnant

Verb Category	Examples
Amuse	concerning, sought to, offend, tried to, gladden, inspired, try to, trieth, shame, sought, tempt, gladdening, pleased with, gladden, dazzle, afflict
Appear	shall come, came, arose, cometh, appear
Be	is, were, was, are, will be, shall be
Build	make, made, making, will make, would make, build
Characterize	remember, establish, regarding, entered into, used, describing, praised, ascribe, hearld, regards, praised, were numbered, choose
Conjecture	know, knew, knoweth, means, deny, let, observe
Directed motion	go, enter into, ascendeth, departed, arrived, flee, fell, escape, fall upon, descendeth, going, go, depart
Do	did, do, doeth, had done, was done
Future	shall, will, shalt
Get	called, gather, found, save, buy, invoke, call earn, reach
Give	pass, give, gave, giveth, pass
Have	have, had, hath, having
Message transfer	teach, tell, told, preach, ask, translated, teaches, translates, shew, show, ministering, present, to minister
Put	put, set, placed, arranged, inserted at, mounted
Say	said unto, say unto, saith, says, declare, claim, saying
See	saw, see, seeing, seest, heard, hear, taste

The next highest similarity is the New Testament's similarity to the Book of Mormon.

Figure 5 shows the Self-Organizing Map created based on the 86 verb topic inputs. According to Table 4, we should most likely see the Old Testament, the Book of Mormon, the Popol Vuh, and the Rig Veda across the center of the map. This predicted outcome plays out, but only in part. The Rig Veda, the Book of Mormon, and the Popol Vuh do appear in the middle of the SOM. However, the Old Testament joins the Torah at the bottom of the SOM. The Old Testament's high average overlap was largely a result of its high similarity to the Torah and the Book of Mormon despite its lower overlap with the Tao Te Ching and the Greater Holy Assembly. Because the range of its overlap percentage was so great, it was pushed out of the center towards the books with which it had the most overlap.

Figure 5: Self Organizing Map of Nine Religious Texts Using 86 Verb Category Inputs



As far as verb topics go, the sample religious texts have the least overlap with the Greater Holy Assembly. The sample religious books have, on average, only a 25 percent overlap with the Greater Holy Assembly. On the SOM, however, the Greater Holy Assembly is on a side, but not in the corner. The placement may in part result because the Greater Holy Assembly has its greatest overlap with the Popol Vuh (52 percent) and the Book of Mormon (48 percent). Because the Popol Vuh and the Book of Mormon are central on the map, the Greater Holy Assembly was placed a bit more central. The Tao Te Ching, which is in a more isolated position than the Greater Holy Assembly on the SOM has the next lowest average overlap percent at 46 percent.

The New Testament has a very central position on the SOM. The New Testament has its highest overlap with the Book of Mormon (79 percent), the Popol Vuh (66 percent), the Old Testament (66 percent), and the Qur'an (62 percent). Being similar to both the Qur'an and the Old Testament gives it central position on the SOM, but also having consistent overlap with many books accounts for its central position as well.

VII. Discussion

We started our analysis with four primary research questions. In this section, we address each research question and elaborate specifically regarding the religious texts that clustered together.

A. Quality of Observations From Maps

The first research question was whether any interesting observations could be drawn from the resulting analysis and visualization of the sacred texts. Automatic text analysis is very common, but applying those same text mining techniques to religious text is not common. While the criteria of "interesting" is admittedly subjective, the maps produced by the text-processing algorithms have supplied evidence of being useful. For example, the noun topic map shown in Figure 4 places the Book of Mormon in the center of the SOM. This placement is a result in part of having strong overlap with the New Testament, the Old Testament, and the Torah. The overlap makes sense as the Book of Mormon shares an account of a group that left Jerusalem around 600 BCE that lived the Law of Moses. The group traveled to the American continent. About half-way through the book, Christ visits and ministers to the people of the American continent. The stories in the Book of Mormon overlap similar time frames as the Old Testament, the Torah, and the New Testament, so it makes sense that they would share a lot of noun topic overlap.

B. Document Placement on the SOM and Tag Overlap

The second question was whether the location of documents on the output layer of the SOM corresponds to word category overlapping among religious texts. The short answer is yes, we were able to track the placement of books on the SOM (both on the noun and verb maps) to the overlap percentages listed in Tables 3 and 4. The SOM, however, was more sophisticated than just a summary of topic overlaps. The SOM took into account additional factors in addition to total tag overlap. When books shared topics with just one other book, then that unique overlap was valued more than an overlap that was widely shared by the books. In other words, not all topics provided equal amounts of information into the similarity between books. For example, we were able to explain why the Rig Veda was placed next to the Torah on the noun SOM, despite having such a low tag overlap percentage. Also, average tag overlap to a book does not necessarily result in a book being centrally located on the SOM. The Old Testament, for example, has the highest average tag overlap of verb tags, but it appears at the bottom of the SOM in Figure 5. The SOM considers overlap relationships of every book and not just averages that can be skewed by very large overlaps by two books (like the Old Testament and the Torah).

C. What Texts Actually Cluster Together

The third research question was which religious texts actually cluster together. In the SOM from the noun inputs and in the SOM from the verb inputs, the Book of Mormon clustered closely with the New Testament, the Old Testament, and the Torah. Also, if books clustered with the Old Testament, they typically clustered with the Torah, due to the Torah's high overlap with the Old Testament (96 percent). For example, the Rig Veda had its highest verb overlap with the Torah (69 percent) and its second highest overlap was with the Old Testament (67 percent). The Popol Vuh had a high overlap with the Torah (62 percent), but also a high overlap with the Old Testament (55 percent). We had expected books to cluster based on their geographic origins and this result did not occur. While it is interesting to observe the clusters of books, we want to further explore the actual words that are shared between books.

Shared topic categories can reveal differences between texts in addition to analyzing topics that were not shared. Table 6 shows very common topics shared among the texts, yet differences abound. For example, the Popol Vuh, the Tao Te Ching, and the Greater Holy Assembly include more references to female family roles like grandmother, mother and daughter. The Qur'an, the New Testament, the Book of Mormon, and the Rig Veda reference male family roles like father(s), son(s), and brother more. Also, the Greater Holy Assembly and the Tao Te Ching reference mouth frequently, while the other five books reference hand(s) a lot. Of the common emotions, joy and fear seem to be the most common. Instead of joy, though, the Tao Te Ching talks of happiness and the Qur'an talks of pleasure. The kinds of animals in the books are also interesting. Sheep is a common in the Book of Mormon and the New Testament, but not in other books. Cows are common in the Rig Veda and the Qur'an, but not in other books. The term beast(s) is common in the Greater Holy Assembly, the New Testament, the Old Testament, and the Torah.

Table 6: Common Noun Categories with Most Common Examples

Books	Family Relationships	External Body Parts	Emotions	Virtues	Animals
Book of Mormon	father(s), son, children	hand(s), face, eyes	joy, fear, pride	strength, righteousness	flocks, lamb, sheep
Greater Holy Assembly	son, father, daughter	hair, beard, mouth	wrath, anger, fear	wisdom	creatures, beasts, dragons
KJV – New Testament	son, father, brother	hand(s), flesh, body	love, joy, fear	righteousness	beast, sheep, lamb
KJV – Old Testament	son, father, wife, daughter	hand(s), eyes, mouth	fear, anger, wrath	righteousness, trust, honour, hope	beast(s), sheep, cattle, oxen
Popol Vuh	father, grandmother, sons	face(s), eyes, hand	joy, laughter, pity	strength, rank	bird(s), jaguar, deer
Rig Veda	son, children, father(s)	eye(s), hand(s), body	joy, love, bliss	strength	kine/cow, horse, cattle
Qur'an	father(s), son brother	hand(s), eyes, ears	fear, wrath, pleasure	good, beneficent, forgiveness	cattle, creature, birds, beast
Tao Te Ching	mother, children, family	arms, mouth, body, eyes	happiness, fear, dislike	skill, gentleness	creatures, dog, rhinoceros
Torah	children, son(s), fathers	hand(s), flesh, eyes	guilt, anger, fear, wrath	grace, justice, righteousness	beast(s), cattle, bullock

In addition to topic categories that are shared by religious texts, most books have a few topics that are unique to it. Examples from the unique topics are listed in Table 7.

Table 7: Unique Noun Examples (total unique categories in parenthesis)

Book of Mormon (4)	categories: cause, manner, military, record ex: cause, manner, ways, armies, record(s)
Greater Holy Assembly (11)	categories: +, =, access, concession, disposition, greater god, introduction, jewelry, numbers, scripture, shape ex: openings, conformations, dispositions, Macroprosopus, Microprosopus, crown(s), number, Psalms, curls
KJV-New Testament (2)	categories: water vehicle, motivation ex: ship(s), boat, sake, reason, temptation
KJV-Old Testament (0)	category: shared metal with Torah ex: gold, brass, iron
Popol Vuh (11)	categories: ball, culture, demon, hispanic deity, insect, language, maya deity, maya demon, order, phrase, version, ex: ball, language, lenguas, Quiche, Maya, Spanish, Mexicans, wasps, louse, ants, bumblebees, stone, title, tribe(s)
Qur'an (10)	categories: discover, evidence, hell, human creation, job, likeness, punishment, ready, religion, success ex: hell, duty, thy duty, tidings, portent(s), woe, burden, harm, aware, scripture, religion, revelation
Rig Veda (5)	categories: assistance, attack, drink, hindu deity, natural disaster, ex: aid, help, battle, juice, milk, drink, Indra, Agni, Soma, Maruts, flood(s), tempest, quake
Tao Te Ching (22)	categories: activity, attitude, behavior, case, complexity, condition, confusion, difficult, existence, govword, idea, impact, individual, vessel, method, movement, nobody, performance, point, rest, state, and value ex: activity, favour, dignity, conduct, complications, condition, disorder, difficulty, existence, government, idea, degree, self, vessel, method, movement, no one, show, display, point, rest, state, superiority
Torah (0)	category: shared metal with the Old Testament ex: gold, brass, iron

The Book of Mormon uniquely has references to record keeping, armies, and “ways” or manners of doing things. The Greater Holy Assembly has more references to jewelry, conformations, dispositions, Macroprosopus and Microprosopus, and numerous mathematical symbols. The New Testament uniquely has more references to water vehicles such as ships and boats. The Old Testament did not have any unique word categories, but shared the metal category uniquely with the Torah. Both books had many instances of gold, brass, and iron. The Popol Vuh has unique references to balls for playing sports, different languages, cultures, tribes, and numerous insects. Insects from the Popol Vuh include wasps, ants, bumblebees, and lice. The Qur’an has a greater number of references to Hell, duties, tidings, portents, burdens, and the term revelation. The Rig Veda has a greater number of references to helping and giving aid and to battles. The Rig Veda also has more references to beverages such as juice, milk and drinks. The

Rig Veda includes references to Hindu deity, such as Indra, Agni, Soma, and Maruts. Finally, the Rig Veda had more references to natural disaster words like floods, tempest, and quake. The Tao Te Ching included many unique categories. Some of its unique terms tended to be abstract in nature like existence, idea, self, vessel, rest, state, and superiority.

D. Clustering with Noun Topics Versus Clustering with Verb Topics

Our final research question was whether the clustering of texts using noun topics varies from the clustering of texts using verb topics. The answer is yes, the clustering does vary between SOMs. The main difference is between the centrality of the Rig Veda and the Popol Vuh in the verb SOM compared to their lack of centrality in the noun SOM. Consistent between the noun and verb SOMs, however, is the common relationship between the Book of Mormon, the New Testament, the Old Testament, and the Torah. While the placement on the maps of the group of four is not identical, the connection remains strong in both SOMs. Also, the Tao Te Ching remains on the edges of both SOMs.

VIII. Limitations and Future Work

A limitation of our current work is in our selection of texts to analyze. We aimed to process a sample of religious texts from various religious traditions and geographic areas as opposed to being completely comprehensive. For example, we included only one of the four Vedas from Hinduism. We included only one of the three books of the Zohar from Kabbalah. We included only the Torah from Judaism specifically and not the rabbinic commentaries on the Torah which are part of the Oral Torah. With so many different religious texts available in the world, analyzing only nine is a limitation of our work.

An additional limitation of our work is our creation of the noun tag hierarchy. While its creation was informed by current existing tag hierarchies (Miller, 1995 as well as Sekine and Nobata, 2003), our current noun hierarchy has not been evaluated and we cannot make any claims to its generalizability. The verb category hierarchy however uses the verb categories published by Levin (1993).

Finally, we recognize that there are variations in different translations of the sacred texts we have used. We used only one translation of each work. Future work could include clustering different translations of the same texts to see how similar the books appear on the self-organizing map.

Also, in future work, we want to expand the number of books we include in our analysis. We also want to reduce the number of topics that get used as inputs to see how that changes the clustering on the SOM. For example, we want to explore what would happen if we included topics that accounted for only 20 percent of the nouns and verbs as opposed to the 85 percent noun and verb coverage we have in this research.

While in this paper, we focused more on how religious texts clustered together given our inputs to the SOM, in future research, we want to explore deeper how the religious texts differ. With the inputs to the noun SOM, 36 percent of the topics occurred in only one religious text. Of the verb inputs, 35 percent of verb topics occurred in only one religious text. To better understand the unique contribution of each book, we want to explore more deeply where the texts differ given the topic categories.

References

Alpert, Jesse, and Nissan Hajaj. 2008. "We Knew the Web Was Big..." *Google Official Blog.*

Bowman, C. Mic, Peter B. Danzig, Udi Manber, and Michael F. Schwartz. 1994. "Scalable Internet Resource Discovery: Research Problems and Approaches," *Communications of the ACM*, 37(8): 98-107.

Chen, Hsinchun, Chris Schuffels, and Richard Orwig. 1996. "Internet Categorization and Search: A Self-Organizing Approach." *Journal of Visual Communication and Image Representation*, 7(1): 88-102.

DARPA. 1993a. "Proceedings," *Tipster Text Program (Phase I).* Fredricksburg, Virginia: Morgan Kaufmann.

DARPA. 1993b. "Proceedings," *Fifth Message Understanding Conference (MUC-5).* Baltimore, MD: Morgan Kaufmann.

DARPA. 1995. "Proceedings," *Sixth Message Understanding Conference (MUC-6).* San Francisco, California: Morgan Kaufmann.

DARPA. 1998. "Proceedings," *Seventh Message Understanding Conference (MUC-7).* Washington, D.C.: Morgan Kaufmann.

de Vel, O., A. Anderson, M. Corney, and G Mohay. 2001. "Mining E-Mail Content for Author Identification Forensics," *ACM SIGMOD Record,* 30(4): 55-64.

Fan, Weiguo, Linda Wallace, Stephanie Rich, and Zhongju Zhang. 2006. "Tapping the Power of Text Mining." *Communications of the ACM*, 49(9): 76-82.

Gray, Andrew, Philip Sallis, and Stephen MacDonnell. 1997. "Software Forensics: Extended Authorship Analysis Techniques to Computer Programs," in *Proceedings of the 3rd Biennual Conference on the International Association of Forensic Linguists.*

Gutenberg, Project. 2013. "Free Ebooks - Project Gutenberg".

Hare, John Bruno. 2013. "Internet Sacred Text Archive".

Hearst, Marti A. 1999. "User Interfaces and Visualization," in *Modern Information Retrieval,* eds. Ricardo Baeza-Yates and Berthier Ribeiro-Neto, 257-339. New York: ACM Press.

Honkela, Timo, Ville Pulkki, and Teuvo Kohonen. 1995. "Contextual Relations of Words in Grimm Tales Analyzed by Self-Organizing Map," in *Proceedings of the International Conference on Artificial Neural Networks (ICANN-95),* eds. F. Fogelman-Soulié and P. Gallinari, 3-7, Nanterre, France.

Hope, Warren. 2009. *The Shakespeare Controversy: An Analysis of the Authorship Theories.* 2nd ed. McFarland & Co.

Jain, Anil K., and Richard C. Dubes. 1988. *Algorithms for Clustering Data*. Upper Saddle River, NJ: Prentice Hall Inc.

Jockers, Matthew L., Daniela M. Witten, and Craig S. Criddle. 2008. "Reassessing Authorship of the *Book of Mormon* Using Delta and Nearest Shrunken Centroid Classification." *Literary & Linguistic Computing*, 23(4): 465-91.

Kaski, Samuel, Timo Honkela, Krista Lagus, and Teuvo Kohonen. 1993. "Creating an Order in Digital Libraries with Self-Organizing Maps," *International Conference on Neural Networks (ICANN).* London, UK: Springer, 974-77.

Kaski, Samuel, Timo Honkela, Krista Lagus, and Teuvo Kohonen. 1998. "WEBSOM—Self-Organizing Maps of Document Collections." *Neurocomputing*, 21(1): 101-17.

Kohonen, Teuvo. 1995. *Self-Organizing Maps*. Berlin: Springer-Verlag.

Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation.* Chicago: The University of Chicago Press.

Lin, Chienting, Hsinchun Chen, and Jay F. Nunamaker. 1999. "Verifying the Proximity Hypothesis for Self-Organizing Maps," in *Proceedings of the 32nd Hawaii International Conference on Systems Sciences.* Hawaii.

Lin, Xia, Dagobert Soergel, and Gary Marchionini. 1991. "A Self-Organizing Semantic Map for Information Retrieval," in *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* 262-69.

McDonald, Daniel, Hsinchun Chen, Hua Su, and Byron B. Marshall. 2004. "Extracting Gene Pathway Relations Using a Hybrid Grammar: The Arizona Relation Parser." *Bioinformatics.* 20(18).

Medicine, U.S. National Library of. 2013. "Fact Sheet Medline".

Menzie, Nicola. 2013. "Top Bible Translations Remain NIV, KJV, and NKJV," *The Christian Post:* September 19.

Miller, George A. 1995. "WordNet: A Lexical Database for English." *Communications of the ACM*, 38(11): 39-41.

Mosteller, Frederick. 1964. *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. Springer.

Orwig, Richard E., Hsinchun Chen, and Jay F. Nunamaker. 1997. "A Graphical, Self-Organizing Approach to Classifying Electronic Meeting Output." *Journal of the American Society of Information Science*, 48(2): 157-70.

Pirolli, Peter, Stuart K. Card, and Mija Van Der Wege. 2001. "Visual Information Foraging in a Focus + Context Visualization," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* 506-13.

Reynolds, Noel B. 1982. *Book of Mormon Authorship*. Maxwell Institute.

Reynolds, Noel B. ed. 1997. *Book of Mormon Authorship Revisited: The Evidence for Ancient Origins*. Maxwell Institute.

Reynolds, Noel B., and Charles D. Tate eds. 1982. *Book of Mormon Authorship: New Light on Ancient Origins*. Bookcraft.

Roussinov, Dimitri G., and Hsinchun Chen. 1999. "Document Clustering for Electronic Meetings: An Experimental Comparison of Two Techniques." *Decision Support Systems*, 27(1-2): 67-79.

Salton, Gerald, Anita Wong, and Chung-Shu Yang. 1975. "A Vector Space Model for Automatic Indexing." *Communications of the ACM*, 18(11): 613-20.

Sekine, Satoshi, and Chikashi Nobata. 2003. "Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy," *Proceedings of the Language Resources and Evaluation.*

Thang, Cao. 2011. "Spice-Som".

Zheng, Rong, Jiexun Li, Hsichun Chen, and Zan Huang. 2006. "A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques." *Journal of the American Society for Information Science and Technology*, 57(3): 378-93.