Math. Appl. **8** (2019), 151–171
DOI: 10.13164/ma.2019.10

WHAT DO ASIAN AND NON-ASIAN SCRIPTURES HAVE IN COMMON? AN APPLIED STATISTICAL MACHINE LEARNING INQUIRY

PREETI SAH AND ERNEST FOKOUÉ

Abstract. This paper presents a substantially detailed statistical machine learning approach to the analysis of several aspects of sacred texts from both the Asian and Biblical scriptural canons. The corpus herein considered consists of 4 Asian sacred scriptures, namely the Tao Te Ching, the teachings of the Buddha, the Yogasutras of Patanjali, and the Upanishads, and 4 non-Asian sacred texts essentially four books from the Bible, namely the Book of Proverbs, the Book of Wisdom, the Book of Ecclesiastes and the Book of Ecclesiasticus. Standard text mining tools are used, like the creation of Document Term Matrices (DTM) to pre-process raw English translations into word frequencies, and both unsupervised and supervised learning methods are used to answer some foundational questions featuring similarities and dissimilarities within each canon and interesting differences between all the canons considered. Despite the vast disparities between the translators of the original texts, our findings reveal sharp differences between Asian and non Asian scriptures regardless of whether clustering techniques or pattern recognition methods are used. We provide several compelling visualizations to help highlight our striking findings, chief of which are the persistent groupings of the scriptures based on geography.

1. Introduction

Consider a fragment of sacred text like, say $\mathbf{x} \equiv$ *"For the spirit of wisdom is benevolent, and will not acquit the evil speaker from his lips: for God is witness of his reins, and he is a true searcher of his heart, and a hearer of his tongue".* One may be interested is answering the question: *What is the probability that this fragment of scripture came from the Upanishads, or the Yoga Sutra or the Bible?* The same question can be asked in connection with yet another fragment of scripture like $\mathbf{x} \equiv$ *"No manifested form of life can be independent of its source, just as no wave, however mighty, can be independent of the ocean. Nothing moves without that Power. He is the only Doer".* We consider $B = 8$ different sacred scriptural texts from the scriptural canon of 4 different religions, namely Christianity, Hinduism, Taoism and Buddhism, with the finality of using very standard statistical text mining tools to explore questions typically always explored solely via qualitative research means, namely: (a) Are Asian religions all that different after all? (b) What are the foundational common traits among Asian spiritual traditions?

MSC (2010): primary 62F15; secondary 62F07.
Keywords: Asian religious texts, biblical texts, statistical text mining, similarity measures, distances, document term matrix, k-means clustering, partitioning around medoids, unsupervised learning, graphs, k-nearest neighbors, support vector machine, random forest.

(c) Is it possible for a quantitative analysis to help gain objective insights that are not always easily glean by the subjective conclusions of qualitative researchers? (d) How different are Asian scriptural sacred texts different from the Biblical canon? First and foremost, we like to make it clear that our work is purely exploratory at this stage, also based on the rather simplistic and borderline naive assumption known as the Bag Of Words (BOW) assumption that represents a given document solely by the collection of meaningful words that constitute it. Essentially the data for our study is a collection of text documents transformed into mathematical objects known as document term matrices (DTM) or Term Document Matrices (TDM) containing the frequencies of the words appearing in the documents. Specifically, we consider the following: (1) Hinduism (India): Yogasutras, Upanishads (2) Buddhism (Tibet): Four Noble Truth of Buddhism (3) Taoism (China): Tao Te Ching (4) Christianity (Central Asia/America): Book of Proverb, Book of Ecclesiastes, Book of Ecclesiasticus, Book of Wisdom. We crucially normalized/standardized our dictionary by considering the English translations of the original sacred texts. Most of our documents came from the now famous and well known Project Gutenberg, and both the raw texts and the preprocessed Document Term Matrices (DTM) have been made available to anyone who wishes to explore this work any further. Now, there are several challenges with the data: non-uniform structure data in each sacred book, initial preprocessing reveals a large amount of stop word data which can mislead the similarity measures. Throughout this paper, our document analysis assumes that (a) a given chapter in one of the given books is the document, and is the smallest unit of data being used for finding similarity (b) under the Bag Of Words (BOW) assumption, each document is represented by the meaningful words it is made up of. Using the BOW assumption, our basic data structure after pre-processing is the term-document matrix (tdm) also known as the document term matrix (DTM), which can be written in the following $n \times p$ matrix. For our study, we have $B = 8$ different books made up of $n = 590$ different chapters (documents) and a total of $p = 8265$ terms (words). As a result, the preprocessed representation of our corpus is a 590×8265 Document Term Matrix (DTM), $\mathbf{X} \in \{0, 1, \cdots, c^\star\}^{590 \times 8265}$, where $c^\star$ is the largest possible word frequency. Essentially, our gigantic corpus consists of B stacked up DTMs, $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \cdots, \mathbf{X}^{(b)}, \cdots, \mathbf{X}^{(B)}$, where each b^{th} sacred text is considered separately by way of its own DTM $\mathbf{X}^{(b)}$, given by

$$\mathbf{X}^{(b)} = \begin{bmatrix} X^{(b)}_{1,1} & X^{(b)}_{1,2} & \cdots & \cdots & \cdots & \cdots & X^{(b)}_{1,j_b} & \cdots & X^{(b)}_{1,p} \\ \vdots & \vdots & \ddots & \ddots & \cdots & \cdots & \cdots & \cdots & \vdots \\ X^{(b)}_{i_b,1} & X^{(B)}_{i_b,2} & \cdots & \cdots & \cdots & \cdots & X^{(b)}_{i_b,j_b} & \cdots & X^{(b)}_{i_b,p} \\ \vdots & \vdots & \ddots & \ddots & \cdots & \cdots & \cdots & \cdots & \vdots \\ X^{(b)}_{n_b,1} & X^{(b)}_{n_b,2} & \cdots & \cdots & \cdots & \cdots & X^{(b)}_{n_b,j_b} & \cdots & X^{(b)}_{n_b,p} \end{bmatrix}. \tag{1.1}$$

Each column $\mathbf{X}^{(b)}_{\cdot,j_b}$ of $\mathbf{X}^{(b)}$ represents an atomic word like *truth*, *diligent*, *sense*, *power*, *right*. Each row $\mathbf{X}^{(b)}_{i_b,\cdot}$ of $\mathbf{X}^{(b)}$ represents a document, which in our case is what we call an entire chapter of the b^{th} sacred book in our corpus. In most document analysis tasks, the term document matrix $\mathbf{X}^{(b)}$ is typically very sparse,

with 90% of zeros not unusual. Besides, except in rare cases, $\mathbf{X}^{(b)}$ tends to be ultra-high dimensional, meaning that $p \gg n$ as depicted in the matrix, since the number of words tends be far much higher than the number of documents to be text-analyzed. Typically, the entries $X^{(b)}_{i_b,j_b}$ of $\mathbf{X}^{(b)}$ are defined as follows:

$$X^{(b)}_{i_b,j_b} \equiv \text{ Frequency of the } j_b{}^{\text{th}} \text{ word in the } i_b{}^{\text{th}} \text{ chapter of the } b^{\text{th}} \text{ book of the corpus } \mathbf{X}.$$

For instance, each of the fragments below, from each of the $B = 8$ sacred scriptures considered, is pre-processed into word frequencies, and made into a portion of the corresponding $\mathbf{X}^{(b)}$, ready for all the different statistical analyses considered in this paper.

Buddhism: *And what are fabrications? There are these six classes of intention: intention aimed at sights, sounds, aromas, tastes, tactile sensations, ideas. These are called fabrications [1].*

Tao Te Ching: *Heaven and earth do not act from (the impulse of) any wish to be benevolent; they deal with all things as the dogs of grass are dealt with. The sages do not act from (any wish to be) benevolent; they deal with the people as the dogs of grass are dealt with. May not the space between heaven and earth be compared to a bellows? 'Tis emptied, yet it loses not its power; 'Tis moved again, and sends forth air the more. Much speech to swift exhaustion lead we see; Your inner being guard, and keep it free [2].*

Upanishads: *The Brahman once won a victory for the Devas. Through that victory of the Brahman, the Devas became elated. They thought, "This victory is ours. This glory is ours." Brahman here does not mean a personal Deity. There is a Brahma, the first person of the Hindu Trinity; but Brahman is the Absolute, the One without a second, the essence of all. There are different names and forms which represent certain personal aspects of Divinity, such as Brahma the Creator, Vishnu the Preserver and Siva the Transformer; but no one of these can fully represent the Whole. Brahman is the vast ocean of being, on which rise numberless ripples and waves of manifestation. From the smallest atomic form to a Deva or an angel, all spring from that limitless ocean of Brahman, the inexhaustible Source of life. No manifested form of life can be independent of its source, just as no wave, however mighty, can be independent of the ocean. Nothing moves without that Power. He is the only Doer. But the Devas thought: "This victory is ours, this glory is ours." [4].*

Yoga Sutras: *perception of the true nature of things. When the object is not truly perceived, when the observation is inaccurate and faulty, thought or reasoning based on that mistaken perception is of necessity false and unsound [3].* The Book of Proverbs: *Doth not wisdom cry aloud, and prudence put forth her voice? 8:2. Standing in the top of the highest places by the way, in the midst of the paths, 8:3. Beside the gates of the city, in the very doors she speaketh, saying: 8:4. O ye men, to you I call, and my voice is to the sons of men. 8:5. O little ones understand subtlety, and ye unwise, take notice. 8:6. Hear, for I will speak of great things: and my lips shall be opened to preach right things. 8:7. My mouth shall meditate truth, and my lips shall hate wickedness [3].*

The Book of Ecclesiastes: *Speak not any thing rashly, and let not thy heart be hasty to utter a word before God. For God is in heaven, and thou upon earth: therefore let thy words be few. 5:2. Dreams follow many cares: and in many words shall be found folly. 5:3. If thou hast vowed any thing to God, defer not to pay it: for an unfaithful and foolish promise displeaseth him: but whatsoever thou hast vowed, pay it. 5:4. And it is much better not to vow, than after a vow not to perform the things promised. 5:5. Give not thy mouth to cause thy flesh to sin: and say not before the angel: There is no providence: lest God be angry at thy words, and destroy all the works of thy hands. 5:6. Where there are many dreams, there are many vanities, and words without number: but do thou fear God [6].*

The Book of Ecclesiasticus: *Then Nathan the prophet arose in the days of David. 47:2. And as the fat taken away from the flesh, so was David chosen from among the children of Israel. 47:3. He played with lions as with lambs: and with bears he did in like manner as with the lambs of the flock, in his youth. 47:4. Did not he kill the giant, and take away reproach from his people? 47:5. In lifting up his hand, with the stone in the sling he beat down the boasting of Goliath: 47:6. For he called upon the Lord the Almighty, and he gave strength in his right hand, to take away the mighty warrior, and to set up the horn of his nation. 47:7. So in ten thousand did he glorify him, and praised him in the blessings of the Lord, in offering to him a crown of glory: 47:8. For he destroyed the enemies on every side, and extirpated the Philistines the adversaries unto this day: he broke their horn for ever. 47:9. In all his works he gave thanks to the holy one, and to the most High, with words of glory. 47:10. With his whole heart he praised the Lord, and loved God that made him: and he gave him power against his enemies: 47:11. And he set singers before the altar, and by their voices he made sweet melody [7].*

The Book of Wisdom: *Love justice, you that are the judges of the earth. Think of the Lord in goodness, and seek him in simplicity of heart: 1:2. For he is found by them that tempt him not: and he sheweth himself to them that have faith in him. 1:3. For perverse thoughts separate from God: and his power, when it is tried, reproveth the unwise: 1:4. For wisdom will not enter into a malicious soul, nor dwell in a body subject to sins. 1:5. For the Holy Spirit of discipline will flee from the deceitful, and will withdraw himself from thoughts that are without understanding, and he shall not abide when iniquity cometh in. 1:6. For the spirit of wisdom is benevolent, and will not acquit the evil speaker from his lips: for God is witness of his reins, and he is a true searcher of his heart, and a hearer of his tongue [8].*

The whole corpus can then be represented as a complete DTM given by $\mathbf{X}$ in Equation (1.2).

$$\mathbf{X} = \begin{bmatrix} X_{1,1}^{(1)} & X_{1,2}^{(1)} & \cdots & \cdots & \cdots & \cdots & X_{1,j_1}^{(1)} & \cdots & X_{1,p}^{(1)} \\ \vdots & \vdots & \ddots & \ddots & \cdots & \cdots & \cdots & \cdots & \vdots \\ X_{i_1,1}^{(1)} & X_{i_1,2}^{(B)} & \cdots & \cdots & \cdots & \cdots & X_{i_1,j_1}^{(1)} & \cdots & X_{i_1,p}^{(1)} \\ \vdots & \vdots & \ddots & \ddots & \cdots & \cdots & \cdots & \cdots & \vdots \\ X_{n_1,1}^{(1)} & X_{n_1,2}^{(1)} & \cdots & \cdots & \cdots & \cdots & X_{n_B,j_1}^{(1)} & \cdots & X_{n_1,p}^{(1)} \\ \vdots & \vdots & \ddots & \ddots & \cdots & \cdots & \cdots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \cdots & \cdots & \cdots & \cdots & \vdots \\ X_{1,1}^{(b)} & X_{1,2}^{(b)} & \cdots & \cdots & \cdots & \cdots & X_{1,j_b}^{(b)} & \cdots & X_{1,p}^{(b)} \\ \vdots & \vdots & \ddots & \ddots & \cdots & \cdots & \cdots & \cdots & \vdots \\ X_{i_b,1}^{(b)} & X_{i_b,2}^{(B)} & \cdots & \cdots & \cdots & \cdots & X_{i_b,j_b}^{(b)} & \cdots & X_{i_b,p}^{(b)} \\ \vdots & \vdots & \ddots & \ddots & \cdots & \cdots & \cdots & \cdots & \vdots \\ X_{n_b,1}^{(b)} & X_{n_b,2}^{(b)} & \cdots & \cdots & \cdots & \cdots & X_{n_b,j_b}^{(b)} & \cdots & X_{n_b,p}^{(b)} \\ \vdots & \vdots & \ddots & \ddots & \cdots & \cdots & \cdots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \cdots & \cdots & \cdots & \cdots & \vdots \\ X_{1,1}^{(B)} & X_{1,2}^{(B)} & \cdots & \cdots & \cdots & \cdots & X_{1,j_B}^{(B)} & \cdots & X_{1,p}^{(B)} \\ \vdots & \vdots & \ddots & \ddots & \cdots & \cdots & \cdots & \cdots & \vdots \\ X_{i_B,1}^{(B)} & X_{i_B,2}^{(B)} & \cdots & \cdots & \cdots & \cdots & X_{i_B,j_B}^{(B)} & \cdots & X_{i_B,p}^{(B)} \\ \vdots & \vdots & \ddots & \ddots & \cdots & \cdots & \cdots & \cdots & \vdots \\ X_{n_B,1}^{(B)} & X_{n_B,2}^{(B)} & \cdots & \cdots & \cdots & \cdots & X_{n_B,j_B}^{(B)} & \cdots & X_{n_B,p}^{(B)} \end{bmatrix} \tag{1.2}$$

To further clarify, the overall DTM $\mathbf{X}$ is an $n \times p$ matrix, with entries from the frequency set $\{0, 1, 2, c^\star\}$, where $c^\star = \max_{l,m}\{X_{l,m}\} = \text{largest word frequency}$, and

$$\sum_{b=1}^{B} n_b = n_1 + n_2 + \cdots + n_b + \cdots + n_B = n,$$

and $\mathbf{X}^{(b)}$ representing the b^{th} sacred book, the whole $\mathbf{X}$ can be written matrix block form as

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \\ \vdots \\ \mathbf{X}^{(b)} \\ \vdots \\ \mathbf{X}^{(B)} \end{bmatrix} \in \{0, 1, 2, c^\star\}^{n \times p}.$$

It is important to emphasize the fact that the above representation of complex sacred texts with only the frequency of the words clearly loses the subtlety of the deeper meaning often hidden behind the semantics and the idioms that mere words

alone cannot capture. For instance, *Buddhism teaches about the four noble truths. Each of these truths entails a duty: stress is to be comprehended, the origination of stress abandoned, the cessation of stress realized, and the path to the cessation of stress developed. When all of these duties have been fully performed, the mind gains total release [1]. The Tao Te Ching teaches that the Tao is The Way, Not 'Your Way' about. The chapters talk about staying detached, letting go and keeping things simple [2]. The Yoga Sutras of Patanjali contain the essence of wisdom. We think of ourselves as living a purely physical life, in these material bodies of ours. In reality, we have gone far indeed from pure physical life; for ages, our life has been psychical, we have been centered and immersed in the psychic nature [3]. The Upanishads represent the loftiest heights of ancient Indo-Aryan thought and culture. They form the wisdom portion or Gnana-Kanda of the Vedas, as contrasted with the Karma-Kanda or sacrificial portion. In each of the four great Vedas–known as Rik, Yajur, Sama and Atharva–there is a large portion which deals predominantly with rituals and ceremonials, and which has for its aim to show man how by the path of right action he may prepare himself for higher attainment [4]. The Book of Proverbs consists of wise and weighty sentences: regulating the morals of men: and directing them to wisdom and virtue [5]. Book of Ecclesiastes or The Preacher, (in Hebrew, Coheleth,) because in it, Solomon, as an excellent preacher, setteth forth the vanity of the things of this world: to withdraw the hearts and affections of men from such empty toys [6]. Book of Ecclesiasticus gives admirable lessons of all virtues [7]. Book of Wisdom abounds with instructions and exhortations to kings and all magistrates to minister justice in the commonwealth, teaching all kinds of virtues under the general names of justice and wisdom [8].* These tenets of each of the religious certainly cannot be entirely captured by the words alone without the semantics of the writers thereof. We therefore proceed in our analysis, mindful of the inherent limitations of our BOW representation. All the sacred scriptures considered in this paper clearly originated from different geographical locations and at different historical timelines. Some of the most natural questions that arise are: *Are there any similarities between the sacred scriptures in terms what these texts want to teach and how the various quintessential lessons of their canons are organized and taught? Is it possible to go as far as extracting meaningful topical units/entities within and between the sacred scriptures herein considered.* In the paper, we use very standard methods of statistical machine learning to perform both unsupervised and supervised exploration of the answers to the above questions, specifically investigating if and how the sacred scriptures considered are related/connected. Some of the similarity measures considered in [9], such as the Euclidean distance, the Manhattan distance, the Jaccard distance and the Cosine distance are considered are used on the most basic unit, namely the chapter, herein referred to as the document. Several other authors have touched on aspects of text analytics similar to the work explored in this paper, namely [10–14]. Although we did explore topic modelling as well, results were so inconclusive that we decided to postpone it to future work and concentrate on what we deemed more revealing and more useful with the present data. The rest of this paper is organized as follows: Section 2 provides a detailed descriptions of the crucially needed similarity measures used throughout the practical exploration of

our central motivating questions. We present both the within document similarity measures in the form of standard distances and the crucially important between document similarity measures herein created based of the concept of linkage from the hierarchical clustering literature. Section 3 describes the standard unsupervised and supervised machine learning methods used in this paper. Specifically, we use the k-Medoids clustering method known as Partitioning Around Medoids (PAM) and hierarchical clustering to produce the visually compelling dendrogram revealing the relationships between the $B = 8$ sacred scriptures considered. We also present the description of the multiclass classification task between the books along with the simpler binary classification. Section 4 presents all our results and findings in great details, specifically exploring supervised learning techniques like K-Nearest Neighbors, Support Vector Machines and Random Forest with comparisons of predictive performances. Section 5 is dedicated to our conclusion and discussion along with elements of our future work on this fascinating topic.

2. Similarity measures

From the outset of this paper, we clearly stipulated our intention to explore similarities within a given sacred scripture and between a collection of sacred scriptures. The main tool for such an exploration is clearly a measure of similarity or dissimilarity. Many distances and kernels have typically been used in statistics to quantify similarities and dissimilarities between different types of mathematical objects. In this section, we will resort to some of the standard distances used in text analytics like the cosine distance, but we will also content ourselves with the traditional Euclidean and Manhattan distances.

2.1. Similarity measures for between documents/chapters comparisons

For the purposes of this paper, we explore between chapters/documents comparisons using classical/traditional distances like the Euclidean distance and the Manhattan distance, but also a couple of non standard ones. For documents form the a^{th} book and the b^{th} book of the overall corpus,

$$d(\mathbf{X}_l^{(a)}, \mathbf{X}_m^{(b)}) \equiv \text{distance between the } l^{\text{th}}\text{chapter of the } a^{\text{th}} \text{ book } \mathbf{X}^{(a)}, \\ \text{and the } m^{\text{th}}\text{chapter of the } b^{\text{th}}\text{book } \mathbf{X}^{(b)}. \quad (2.1)$$

This dissimilarity measure constitutes the foundation of all our unsupervised comparisons be it within a given sacred book and/or between the chapters of different sacred books. For instance, if one decided to create the distance matrix of all the n chapters of the whole corpus, the matrix created would be the nonnegative $n \times n$ matrix given by

$$\mathbf{D} = \begin{bmatrix} d_{1,1} & d_{1,2} & d_{1,3} & \dots & d_{1,n} \\ d_{2,1} & d_{2,2} & d_{2,3} & \dots & d_{2,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{n,1} & d_{n,2} & d_{n,3} & \dots & d_{n,n} \end{bmatrix} \in \mathbb{R}_+^{n \times n}$$

with $d_{l,m} = d(\mathbf{X}_l^{(a)}, \mathbf{X}_m^{(b)})$ given by Equation (2.1). For the b^{th} sacred book for instance, it might turn out to be interesting to study its internal dynamics, that is, the interplay among its chapters.

$$\mathbf{D}^{(b)} = \begin{bmatrix} d_{1,1}^{(b)} & d_{1,2}^{(b)} & d_{1,3}^{(b)} & \dots & d_{1,n_b}^{(b)} \\ d_{2,1}^{(b)} & d_{2,2}^{(b)} & d_{23} & \dots & d_{2,n_b}^{(b)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{n_b,1}^{(b)} & d_{n_b,2}^{(b)} & d_{n_b,3}^{(b)} & \dots & d_{n_b,n_b}^{(b)} \end{bmatrix} \in \mathbb{R}_+^{n_b \times n_b}$$

where

$$d_{l,m}^{(b)} \equiv d(\mathbf{X}_l^{(b)}, \mathbf{X}_m^{(b)}) \equiv \text{distance between the } l^{\text{th}} \text{ and } m^{\text{th}}\text{chapters of the } b^{\text{th}}\text{book } \mathbf{X}^{(b)}.$$

As far as the distances considered in this paper are concerned, we first used the Euclidean distance,

$$d(\mathbf{X}_l^{(a)}, \mathbf{X}_m^{(b)}) = \left(\sum_{j=1}^{p} (X_{l,j}^{(a)} - X_{m,j}^{(b)})^2 \right)^{\frac{1}{2}}.$$

Then we used the Manhattan distance, namely

$$d(\mathbf{X}_l^{(a)}, \mathbf{X}_m^{(b)}) = \sum_{j=1}^{p} |X_{l,j}^{(a)} - X_{m,j}^{(b)}|.$$

Since the so-called Cosine similarity measure, which is the normalized inner product between two documents, has been found to be very successful in text analytics, we also explored it on our corpus. For simplicity, we herein formulate it for two documents as

$$d(\mathbf{X}_l^{(a)}, \mathbf{X}_m^{(b)}) \equiv d(\mathbf{x}_l, \mathbf{x}_m) = \frac{\mathbf{x}_l^\top \mathbf{x}_m}{(\mathbf{x}_l^\top \mathbf{x}_l)^{\frac{1}{2}} (\mathbf{x}_m^\top \mathbf{x}_m)^{\frac{1}{2}}}.$$

Finally, we considered the Jaccard similarity measure, sometimes defined as the relative cardinality of the intersection between two documents. Specifically, the Jaccard coefficient between two documents $\mathbf{X}_l^{(a)}$ and $\mathbf{X}_m^{(b)}$ is calculated using the formula:

$$\mathsf{sim}(\mathbf{X}_l^{(a)}, \mathbf{X}_m^{(b)}) \equiv \mathsf{sim}(\mathbf{x}_l, \mathbf{x}_m) = \frac{\displaystyle\sum_{j=1}^{p} \min\{\mathrm{x}_{lj}, \mathrm{x}_{mj}\}}{\displaystyle\sum_{k=1}^{p} \max\{\mathrm{x}_{lk}, \mathrm{x}_{mk}\}},$$

from which the corresponding Jaccard distance between those two chapters is then defined as:

$$d(\mathbf{X}_l^{(a)}, \mathbf{X}_m^{(b)}) \equiv d(\mathbf{x}_l, \mathbf{x}_m) = 1 - \mathsf{sim}(\mathbf{x}_l, \mathbf{x}_m).$$

2.2. Similarity measures for comparisons between entire sacred scriptures

The central and indeed overarching question in this research was triggered by the desired to explore and find the similarities and differences (if any) between different spiritual traditions. Operating on the assumption that the core of a spiritual tradition is contained in its main sacred scripture, it makes to consider, define and develop a valid concept of a distance between two different sacred scriptures (books). We herein explored the comparisons between sacred books using distances between sets akin to the linkages used in hierarchical clustering. For brevity and simplicity, we considered only 4 different scenarios, starting with a distance between books adapted from the so-called single or minimum linkage. Specifically, our first distance between books is given by

$$\Delta(\mathbf{X}^{(a)}, \mathbf{X}^{(b)}) = \min_{\substack{l\in[n_a]\\ m\in[n_b]}} \big\{d(\mathbf{X}_l^{(a)}, \mathbf{X}_m^{(b)})\big\},$$

which intuitively means the distance between two books $\mathbf{X}^{(a)}$ and $\mathbf{X}^{(b)}$ becomes the smallest of the $n_a \times n_b$ distances between their respective chapters. The second between books distance borrows from the so-called maximal or complete linkage, and is defined as

$$\Delta(\mathbf{X}^{(a)}, \mathbf{X}^{(b)}) = \max_{\substack{l\in[n_a]\\ m\in[n_b]}} \big\{d(\mathbf{X}_l^{(a)}, \mathbf{X}_m^{(b)})\big\},$$

which intuitively means the distance between two books $\mathbf{X}^{(a)}$ and $\mathbf{X}^{(b)}$ becomes the largest of the $n_a \times n_b$ distances between their respective chapters. This distance tends to produce very good tessellations of books for the mere reason that by traversing the whole book it most likely captures more subtleties. The third distance here is simply the so-called average distance whose intuition is clear from its definition, namely

$$\Delta(\mathbf{X}^{(a)}, \mathbf{X}^{(b)}) = \operatorname*{mean}_{\substack{l\in[n_a]\\ m\in[n_b]}} \big\{d(\mathbf{X}_l^{(a)}, \mathbf{X}_m^{(b)})\big\}.$$

Very interestingly, our fourth distance, is the median whose intuition is also just as clear as the intuition underlying the mean defined earlier.

$$\Delta(\mathbf{X}^{(a)}, \mathbf{X}^{(b)}) = \operatorname*{median}_{\substack{l\in[n_a]\\ m\in[n_b]}} \big\{d(\mathbf{X}_l^{(a)}, \mathbf{X}_m^{(b)})\big\}. \tag{2.2}$$

It turns out that this median distance between books ends up producing far more revealing and intuitively appealing tessellations. Regardless of the distance used, the main mathematical object for our unsupervised learning purposes turns out to be the matrix of distances

$$\boldsymbol{\Delta} = \begin{bmatrix} \Delta_{11} & \Delta_{12} & \Delta_{13} & \dots & \Delta_{1B} \\ \Delta_{21} & \Delta_{22} & \Delta_{23} & \dots & \Delta_{2B} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \Delta_{B1} & \Delta_{B2} & \Delta_{B3} & \dots & \Delta_{BB} \end{bmatrix} \in \mathbb{R}_+^{B\times B},$$

where

$$\Delta_{ab} = \Delta(\mathbf{X}^{(a)}, \mathbf{X}^{(b)}) \equiv \text{distance between the } a^{\text{th}} \text{ sacred book and the } b^{\text{th}} \text{ sacred book.}$$

The matrix $\mathbf{\Delta}$ ends up playing a central role in the clustering algorithms, and in the generation of all the compelling visualizations like the heatmaps, the matrix plots and the beautiful graph plots that revealed the geographical connections between the sacred scriptures. Our future work will consider non-linkage inspired distances between books, like the Frobenius distance,

$$\Delta(\mathbf{X}^{(a)}, \mathbf{X}^{(b)}) = \|\mathbf{X}^{(a)} - \mathbf{X}^{(b)}\|_F^2$$

or even the Kullback–Leibler divergence.

3. Unsupervised and supervised learning on sacred scriptures

At this point, having defined both within books distances and the between books distances, along with the representation of the raw text data in the form of the corresponding document term matrices, we are now in the position to perform both informal statistical analyses like exploratory data analyses through summaries and plots, but also formal statistical analyses via model building and even a bit of inference. One of the analyses we initially intended to carry out is topic modelling, and indeed we did perform a fair amount of it. However, we do not present our topic modelling findings in this paper because those were so inconclusive that we postponed the deeper analysis of the same to our future work.

3.1. Unsupervised knowledge discovery from sacred scriptures

As indicated earlier, one of the most interesting questions one may seek to answer in the presence of a collection of documents dealing with the different sacred texts: *are there any similarities among the various sacred scriptures? And if so, how does one go about identifying, discovering, extracting and revealing those?* In this section, as far as the relationships among entire sacred scriptures are concerned, we herein briefly describe the way in which we use cluster analysis to tackle and answer that overarching question. Specifically, if we anticipate k groups of sacred texts, and denote by $P_k = G_1 \cup \cdots \cup G_k$, the partitioning of the B sacred scriptures into k groups/clusters, then we seek the optimum clustering.

$$P_k^* = \underset{P_k}{\operatorname{argmin}} \left\{ \sum_{c=1}^{k} \sum_{b=1}^{B} z_{bc} \Delta(\mathbf{X}^{(b)}, \mathbf{X}_\star^{(c)}) \right\},$$

where $z_{bc} = \mathbb{1}(\mathbf{X}^{(b)} \in G_c)$ and $\Delta(\cdot, \cdot)$ is one of the distances between sacred scriptures defined earlier. The implementation of this method is standard and readily found in R.

The merit of this portion of the work resides in the proper formulation and the intuitive definition of appropriate distances that proved to be of great practical use. We do not herein perform any within book clustering of chapters, although it is readily doable. We deemed the between books clustering far more interesting and far more appealing.

3.2. Classification of new randomly selected sacred scripture chapters

As stated from the beginning, another question that naturally arises from a corpus of documents like ours with a wide variety of origins is: *For any given document can we predict which sacred text it belongs to?* Religious debates are typically replete with a flurry of quotations in the form of fragments of scriptures. It is not uncommon to encounter those fragments of scriptures in all manners of spiritual books and even textbooks written by scientists and mathematicians. Let $\mathbf{x} \equiv$ *"For the spirit of wisdom is benevolent, and will not acquit the evil speaker from his lips: for God is witness of his reins, and he is a true searcher of his heart, and a hearer of his tongue".* One may be interested is answering the question: *What is the probability that this fragment of scripture come from the Upanishads, or the Yoga Sutra or the Bible?* The same question can be asked in connection with yet another fragment of scripture like $\mathbf{x} \equiv$ *"No manifested form of life can be independent of its source, just as no wave, however mighty, can be independent of the ocean. Nothing moves without that Power. He is the only Doer".* The sacred scripture from which a document/chapter/fragment originated is traced in a supervised learning manner with a variable Y from the set of labels of all the books considered here namely $\mathcal{Y} = \{s_1, s_2, \cdots, s_8\}$ where

s_1 is Book 1 referring to the Teachings of the Buddha;
s_2 is Book 2 referring to the Tao Te Ching;
s_3 is Book 3 referring to the Upanishads;
s_4 is Book 4 referring to the Yoga Sutras of Patanjali;
s_5 is Book 5 referring to the Book of Proverbs;
s_6 is Book 6 referring to the Book of Ecclesiastes;
s_7 is Book 7 referring to the Book of Ecclesiasticus;
s_8 is Book 8 referring to the Book of Wisdom.

Let $\mathbf{x} \in \{0, 1, \cdots, c^\star\}^{8565}$ be brand new 8265-dimensional vector of word frequencies corresponding to a chapter of sacred scripture from one of the $B = 8$ sacred texts considered. An interesting question to ask and explore is the following: what is the probability that the sacred fragment represented by $\mathbf{x}$ comes from the b^{th} book of our corpus?

$$\text{Find } \pi_b(\mathbf{x}) \equiv \mathbb{P}[Y = s_b|\mathbf{x}].$$

While it is interesting to formulate this classification problem in terms of the posterior probability of class membership, most methods regard the task as a function estimation problem, namely:

To find functions $f : \mathcal{X} \to \mathcal{Y}$, such that for every document or fragment of document represented by $\mathbf{x} \in \{0, 1, 2, \cdots, c^\star\}^p$, we seek $f(\mathbf{x}) = \mathsf{class}(\mathbf{x}) \equiv$ *sacred scripture from which* $\mathbf{x}$ *was taken.*

Ideally, we would like to build the Bayes classifier,

$$f(\mathbf{x}) = \underset{b \in [B]}{\operatorname{argmax}} \big\{\mathbb{P}[Y = s_b|\mathbf{x}]\big\}.$$

From an even more empirical perspective, one could seek to constructed a classifier of fragments of sacred scriptures, maybe for the pure purpose of scriptural authentication. In such a case, one would seek to construct a classifier $\widehat{f}$ from the data, such that given the fragment of scripture $\mathbf{x}$, the predicted class or more specifically

the predicted sacred scripture $\widehat{f}(\mathbf{x})$ from which $\mathbf{x}$ comes is found. Three supervised learning methods are used here, namely: (a) k Nearest Neighbors classifier; (b) Support Vector Machine Classifier and (c) Random Forest Classifier. The k Nearest Neighbors classifier for instance has a prediction function of the form

$$\widehat{f}_n^{(\texttt{kNN})}(\mathbf{x}) = \underset{b\in\mathcal{Y}}{\operatorname{argmax}}\left\{\frac{1}{k}\sum_{i=1}^{n}\mathbb{1}(y_i=b)\mathbb{1}(\mathbf{x}_i\in\mathcal{V}_k(\mathbf{x}))\right\}.$$

The other learning machines can be looked up in the literature.

4. Results

Upon segmentation of those books whose partitioning into chapters was not obvious, we ended with a total of $n = 590$ documents/chapters for the whole corpus. Using text mining tools from R, we ended up with a total of $p = 8265$ individual words, the foundational building blocks of our data under our Bag Of Words (BOW) assumption.

4.1. Within book exploratory analysis

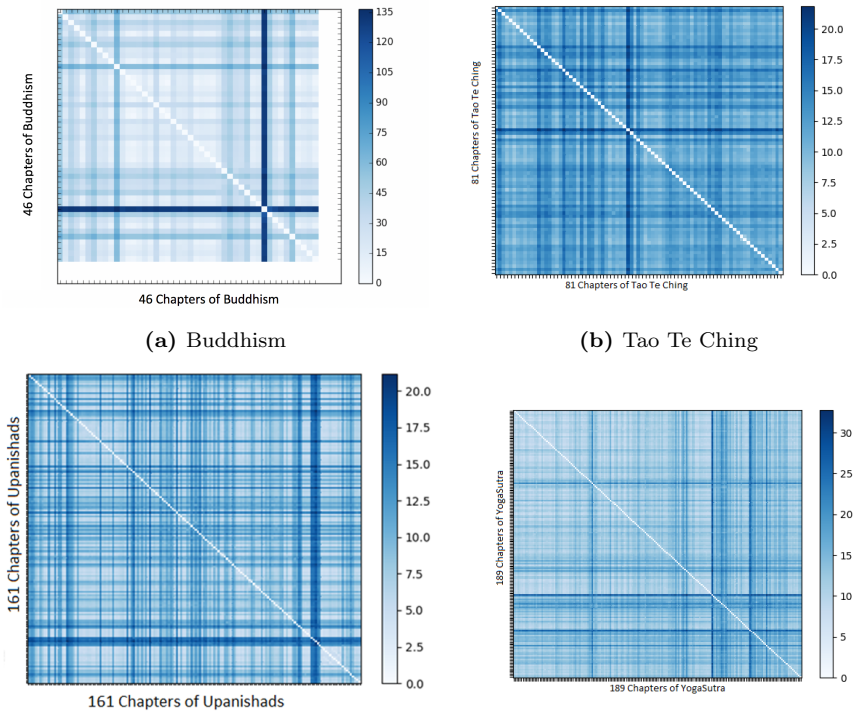
Clearly, it makes total sense, upon constructing the individual distance matrix of any given sacred book, to consider visualizing the internal dynamics of the similarities between its chapters. This subsection concentrates of the use of heatmaps to provide compelling visualizing of each of the sacred book considered. Figures 1a, 1b, 1c and 1d show the Euclidean distances between chapters within the Asian scriptures considered here. Among all the sacred scriptures, chapters within the teachings of the Buddha reveal the highest level of internal consistency as shown in Figure 1a.

With the $B = 8$, considered, we are now in the position to describe the computational results we obtained from our various analyses. A large number of the results are omitted due to space constraints and only the most striking and most revealing ones are presented.

Figures 2a, 2b, 2c and 2d show the Euclidean distance between chapters within the Bible canon of scripture, and like before this helps find the most similar and indeed the most dissimilar chapters within the same book.

It bears recalling that the data used in this paper came mainly from Project Gutenberg. For clarity though, we herein give a slightly more detailed description of the raw data.

- Yogasutras: Project Gutenberg's Yoga Sutras of Patanjali, by Charles Johnston;
- Upanishads: Project Gutenberg's EBook of The Upanishads, by Swami Paramananda;
- Four Noble Truth of Buddhism: `https://www.accesstoinsight.org/lib/study/truths.html`;
- Tao Te Ching: Tao Te Ching - Translated by J. Legge;
- Book of Proverb: Project Gutenberg EBook The Bible, Douay-Rheims, Book 22: Proverbs;



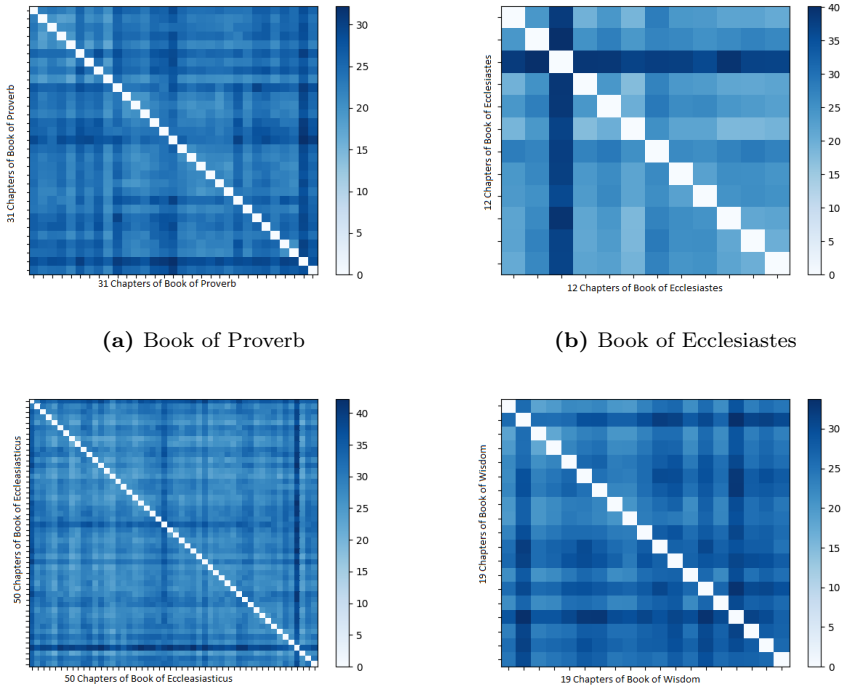
(a) Buddhism **(b)** Tao Te Ching

(c) Upanishads **(d)** Yogasutra

Figure 1. Euclidean distance between different chapters of Asian Religious scriptures.

- Book of Ecclesiastes: Project Gutenberg EBook The Bible, Douay-Rheims, Book 23: Ecclesiastes;
- Book of Ecclesiasticus: Project Gutenberg EBook The Bible, Douay-Rheims, Book 26: Ecclesiasticus;
- Book of Wisdom: Project Gutenberg EBook The Bible, Douay-Rheims, Book 25: Wisdom.

The strength of similarities among different sacred scriptures can be found by visualizing the clustering results obtained using the medianized Euclidean distance. In all the following plots, each node is the network graph represents a sacred scripture, and the strength between two sacred scriptures is proportional to the width and brightness of the corresponding edge. The number of clusters, k in this case, is made to vary from two to seven, and each figure represents groups of similarities for different values of k. [Nodes : Bdd = Buddhism / Tao = TaoTeChing/ Upd = Upanishad/ Yoga = YogaSutra/ Prv = Proverb/ Ecc = Ecclesiastes/ Ecs = Ecclesiasticus/ Wsd = Wisdom]



(a) Book of Proverb **(b)** Book of Ecclesiastes

(c) Book of Ecclesiasticus **(d)** Book of Wisdom

Figure 2. Euclidean distance between different chapters of Bible texts.

4.2. Practical between books comparisons

Among all the between books distances considered and defined earlier, the median defined in Equation (2.2), ended up producing the most revealing knowledge discovery. Moreover, the core distance whose median was computed turned out to be the Euclidean distance. Figure 3 shows the heatmap (matrix plot) corresponding to the medianized Euclidean distance of Equation (2.2). Among all the scriptures considered, the distance is minimum between the Upanishads and the Tao Te Ching.

4.3. Multicategorical and binary classification of fragments of sacred texts

For each of the learning machines considered here, we build them using their internal optimality criteria, and then compare their predictive performances using the average test error $\texttt{AVTE}(\cdot)$, namely

$$\texttt{AVTE}(\widehat{f}) = \frac{1}{R}\sum_{r=1}^{R}\left\{\frac{1}{m}\sum_{t=1}^{m}\ell(\mathrm{y}_{i_t}^{(r)}, \widehat{f}_r(\mathbf{x}_{i_t}^{(r)}))\right\},$$

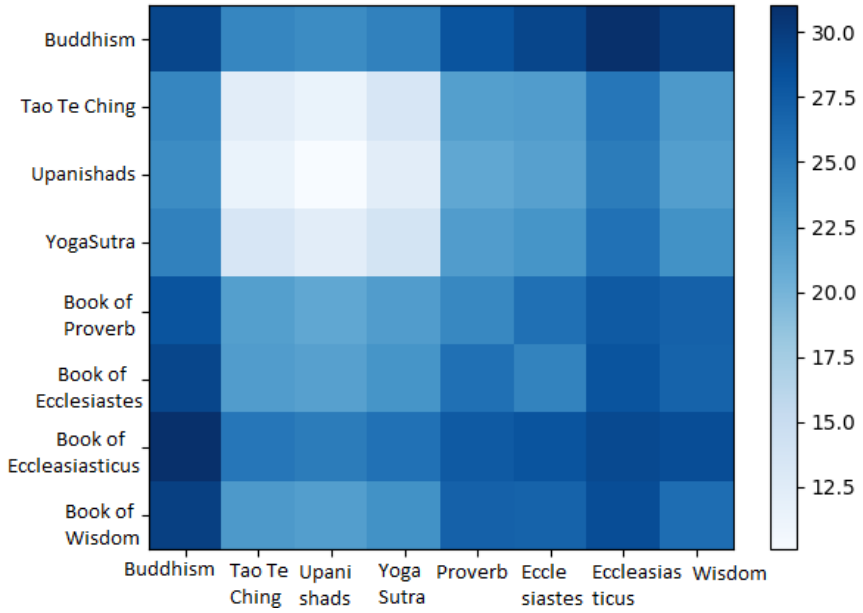
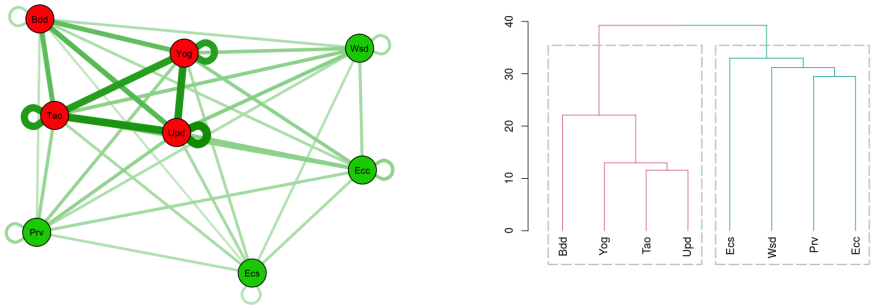


Figure 3. Pictorial view of the medianized Euclidean distances among the 8 sacred scriptures studied.



(a) Graph representation of the 8 sacred scriptures.

(b) Dendrogram of the clustering result.

Figure 4. 2 Medoids clustering of the 8 sacred scriptures.

where $\widehat{f}_r(\cdot)$ is the r-th realization of the estimator $\widehat{f}(\cdot)$ built using the training portion of the split of $\mathcal{D}$ into training set and test set, and $\left(\mathbf{x}_{i_t}^{(r)}, \mathrm{y}_{i_t}^{(r)}\right)$ is the t-th observation from the test set at the r-th random replication of the split of $\mathcal{D}$. For our purposes, we performed $R = 100$ stratified stochastic splits of the $n = 590$

observations into 70% training samples and 30% test samples. As indicated before, we used the most recent versions of each learning machine available in R. All the predictive performances are given in the subsequent tables and comparative boxplots. Tables 1, 2 and 3 are just single instances of confusion matrices herein given to provide a reader with a sense of how the learning machines tended to fare.

Table 1. Example Confusion matrix yield by the kNN learning machine. It can clearly be seen that the 8-category classification task is very challenging for the kNN classifier. Here, we see that the accuracy for kNN is very low, specifically $\texttt{ACCURACY}(\widehat{f}_n^{(kNN)}) = 0.365$.

	Ecc	Ecs	Prv	Wsd	Bdd	Tao	Upd	Yog
Ecc	**1**	0	0	0	0	0	0	0
Ecs	0	**0**	0	0	0	0	0	0
Prv	5	0	**6**	0	0	0	0	0
Wsd	2	0	0	**1**	0	0	0	0
Bdd	0	0	0	0	**2**	0	0	0
Tao	0	0	0	0	1	**0**	0	0
Upd	5	3	2	3	11	24	**49**	51
Yog	2	1	1	2	0	0	0	**6**

Table 2. Example Confusion matrix yield by the SVM learning machine. It can clearly be seen that the 8-category classification task is just as challenging for the SVM classifier as it was for kNN. Here, we see that the accuracy for SVM is very low, specifically $\texttt{ACCURACY}(\widehat{f}_n^{(SVM)}) = 0.393$.

	Ecc	Ecs	Prv	Wsd	Bud	Tao	Upd	Yog
Ecc	**11**	3	7	2	0	0	0	0
Ecs	0	**0**	0	0	0	0	0	0
Prv	0	0	**0**	0	0	0	0	0
Wsd	0	0	0	**0**	0	0	0	0
Bdd	0	0	0	0	**2**	0	0	0
Tao	0	0	0	0	0	**0**	0	0
Upd	0	0	0	0	0	0	**0**	0
Yog	4	1	2	4	12	24	49	**57**

Clearly, out of the three supervised learning machines considered, Random Forest yields by far the best predictive performance. The Upanishads and the Yoga Sutras of Patanjali have the largest number of chapters in the corpus and Random Forest is able to accurately predict most of them. kNN fails to identify majority chapters from the Yoga sutras of Patanjali. SVM fails to identify the majority of chapters from the Upanishads and tends to incorrectly predict them as coming from the Yoga Sutras of Patanjali. From the Random Forest confusion matrix, we can see that the Upanishads and the Tao Te Ching are highly similar as the majority of the Upanishads chapters which are not predicted correctly are predicted to belong to Tao Te Ching.

Table 3. Example Confusion matrix yield by the Random Forest learning machine. It can clearly be seen that the Random Forest Classifier substantially outperformed both kNN and SVM on this 8-category classification task. Here, we see that the accuracy for Random Forest is far better than its previous counterparts, specifically $\texttt{ACCURACY}(\widehat{f}_n^{(RF)}) = 0.764$, virtually twice as accurate. Random forest is clearly relatively far better here than kNN and SVM, although still not spectacular in the absolute.

	Ecc	Ecs	Prv	Wsd	Bdd	Tao	Upd	Yog
Ecc	**15**	3	3	3	0	0	0	0
Ecs	0	**0**	0	0	0	0	0	0
Prv	0	0	**6**	0	0	0	0	0
Wsd	0	0	0	**0**	0	0	0	0
Bdd	0	0	0	0	**6**	0	0	0
Tao	0	0	0	1	0	**14**	0	0
Upd	0	0	0	0	5	10	**44**	6
Yog	0	1	0	2	3	0	5	**51**

To further investigate the predictive performances of the supervised learning machines considered, we also ran a binary version of the classification task.

Table 4. Example of confusion matrix generated by kNN. Here, $\texttt{ACCURACY}(\widehat{f}_n^{(kNN)}) = 0.915$.

	Asian	Biblical
Asian	**143**	15
Biblical	0	**19**

Table 5. Example of confusion matrix generated by SVM. Here, $\texttt{ACCURACY}(\widehat{f}_n^{(SVM)}) = 0.971$.

	Asian	Biblical
Asian	**142**	4
Biblical	1	**30**

Table 6. Example of confusion matrix generated by Random Forest. Here, $\texttt{ACCURACY}(\widehat{f}_n^{(RF)}) = 0.994$

	Asian	Biblical
Asian	**143**	1
Biblical	0	**33**

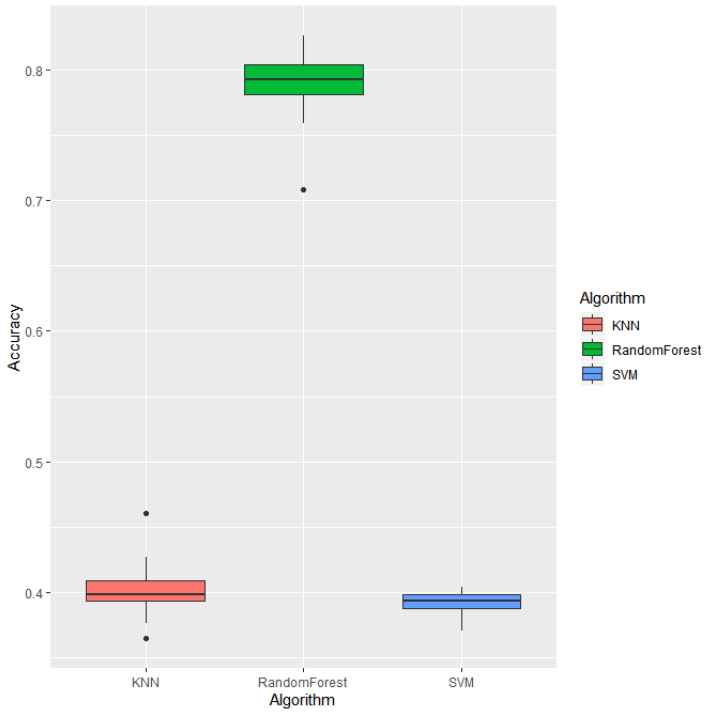


Figure 5. Comparative boxplots of the predictive performances of the learning machines on the 8-category classification task. Although Random Forest exhibits a relatively far better predictive performance, it is clear that 8-category classification task appears very difficult.

5. Conclusion and discussion

We have explored in reasonably great details a quantitative analysis of a carefully built corpus of influential sacred scriptures from both the Asian Religions canon and the Biblical canon. Although we operated throughout our analysis on the rather limited and limiting and borderline naive assumption of a Bag Of Words (BOW) representation of complex spiritual documents, it is very encouraging to see that preliminary findings herein presented are both plausible and refreshingly consistently with several aspects of religions. The most striking finding is that despite the radical difference between all the translators of the documents explored, the statistical machine learning analysis seems to recover the geographical consistency of the sacred scriptures studies, with all the Asian religions clustering together strongly in unsupervised learning, and the binary classification finding it easy to separate fragment of texts from the two main groups/originations. Even the limitation of extreme sparsity inherent in the corpus does not appear to affect the ability of the statistical methods when it comes to clearly identifying the source of a random fragment of sacred scripture from the corpus of interest.

One of the motivations of the study was to hopefully identify and extract those quintessential elements that transcend the origin of sacred scriptures, and that

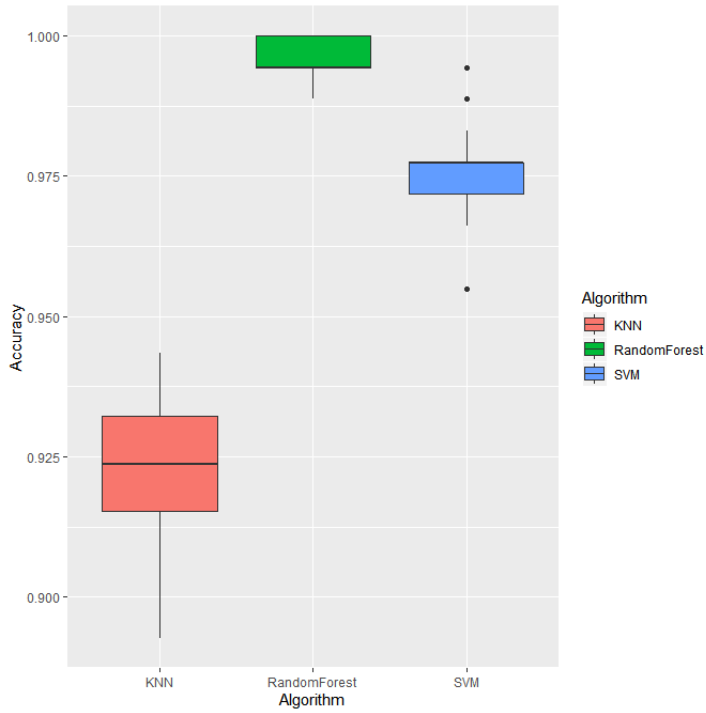


Figure 6. Comparative boxplots of the predictive performances of the learning machines on the binary classification task. This shows clearly that the binary classification version of the underlying task is far easier.

are common to all religions. We attempted that task via topic modelling but in vain: the extracted/revealed topics appeared to be inconclusive, maybe due to the extreme sparseness of the corpus, or the sheer weakness and inadequacy of the bag of words assumptions. An extension of the present work intends to consider a more realistic representation of documents, perhaps one taking the semantics of the documents into account maybe via the use of hidden Markov models as used in language analysis, or other natural language processing tools for representation of documents.

It would be very interesting to consider enriching the corpus further with other religious traditions from around the world, by incorporating the African canon of scriptures, the treasures hidden behind the religions that Originated in South America, more religions of Asia, religions of Europe before Christianity, more books of the Bible, books Judaism, the Holy Qu'ran, and emerging religions. In our present study, despite the naive assumption of the Bag of Words, there seems to be a kind of morphism between the scriptures of religious traditions and the corresponding geographic location. It would be interesting to explore further and find out is scriptures alone appear to consistent with geography up to a mere rotation of maps.

The poor predictive performances noticed in the 8 class pattern recognition and what it might mean. Refinement of classification is harder and maybe sparseness hurt more in 8-category classification than in binary classification. One task, the latter, is clearly harder than the other (the former). The linguistics of Asian canon of religious belief appear clearly different from the one used in the Bible, and this sharp linguistic difference could be the origin of the sharp differences between the two traditions. What happens were linguistic concordance and synonyms are used across traditions to create a normalization of the language used and concentrate on the deep meaning of texts.

We also intend to revisit the topic modelling part of this research more thoroughly both for completeness but also with the hope of finding the hypothesized unity of purpose of all religions.

It is also important to highlight the imbalanced classification aspect of the data as one of the potential sources of the corresponding degenerate confusion matrices, and most likely the source of the poor predictive performances of the learning machines. Throughout this paper, we have operated on the main assumptions that the concepts taught are captured through the words, which means that we have missed the depth of the semantics throughout this work. Yet, there seems to be a persistent and consistent pattern. A good question then is: what happens when the semantics are factored in? Do we still see the same persistent patterns? For instance, do we have enough evidence under such a borderline naive assumption, despite the persistent results, to conclude that Asians do indeed approach spirituality completely differently?

Finally, and in keeping with our strong awareness of the limitations of our assumptions, we view this work as being at the very least a conversation starter, perhaps with qualitative researchers whose findings we would really like to compare to ours, and hopefully embark on a vaster and maybe more conclusive research.

References

[1] T. Bhikkhu, *The four noble truths: A study guide*, 2013, `https://www.accesstoinsight.org/lib/study/truths.html` (accessed 25 March, 2018).

[2] S. Teo, *Three things to learn from Tao Te Ching – The Way, Not 'Your Way'*, Tao Te Ching, `http://tao-in-you.com/three-things-about-tao-te-ching/` (accessed 25 March, 2018).

[3] C. Johnston, *The Yoga Sutras of Patanjali*, 2010, `http://www.gutenberg.org/files/2526/2526.txt` (accessed 25 March, 2018).

[4] S. Paramananda, *The Upanishads*, 2014, `http://www.gutenberg.org/cache/epub/3283/pg3283.txt` (accessed 25 March, 2018).

[5] *The Bible, Douay–Rheims, Book 22: Proverbs The Challoner Revision*, 2005, `http://www.gutenberg.org/cache/epub/8322/pg8322.txt` (accessed 25 March, 2018).

[6] *The Bible, Douay–Rheims, Book 23: Ecclesiastes The Challoner Revision*, 2005, `http://www.gutenberg.org/cache/epub/8323/pg8323.txt` (accessed 25 March, 2018).

[7] *The Bible, Douay–Rheims, Book 26: Ecclesiasticus The Challoner Revision*, 2005, `http://www.gutenberg.org/cache/epub/8326/pg8326.txt` (accessed 25 March, 2018).

[8] *The Bible, Douay–Rheims, Book 25: Wisdom The Challoner Revision*, 2005, `http://www.gutenberg.org/cache/epub/8325/pg8325.txt` (accessed 25 March, 2018).

[9] S. H. M. Qahl, *An Automatic Similarity Detection Engine Between Sacred Texts Using Text Mining and Similarity and Measures*, thesis, Rochester Institute of Technology, 2014.

[10] B. Larsen and C. Aone, *Fast and effective text mining using linear-time document clustering*, in: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '99), ACM, New York, NY, USA, 1999, pp. 16–22.
[11] R. Arun, V. Suresh, C. E. Veni Madhavan and M. N. Narasimha Murthy, *On finding the natural number of topics with latent Dirichlet allocation: Some observations*, in: M. J. Zaki, J. Xu Yu, B. Ravindran and V. Pudi (eds.), Advances in Knowledge Discovery and Data Mining, Springer, Berlin, Heidelberg, 2010, pp. 391–402.
[12] J. Cao, T. Xia, J. Li, Y. Zhang and S. Tang, *A density-based method for adaptive LDA model selection*, Neurocomputing **72** (2009), 1775–1781.
[13] R. Deveaud, É. SanJuan and P. Bellot, *Accurate and effective latent concept modeling for ad hoc information retrieval*, Revue des Sciences et Technologies de l'Information – Série Document Numérique, Lavoisier, 2014, pp. 61–84.
[14] T. L. Griffiths and M. Steyvers, *Finding scientific topics*, Proceedings of the National Academy of Sciences **101** (2004), 5228–5235.

Preeti Sah, College of Computing and Information Sciences, Rochester Institute of Technology, 85 Lomb Memorial Drive, Rochester, New York 14623, USA
e-mail: ks3911@rit.edu

Ernest Fokoué, School of Mathematical Sciences, Rochester Institute of Technology, 98 Lomb Memorial Drive, Rochester, New York 14623, USA
e-mail: epfeqa@rit.edu