Beginner Track: Religious Text Analysis (NLP)

Team Members: Oswald Harris, Chaitanya Naphade, Lukas Williams
Team Name: Divine Patterns
Date: April 11, 2021

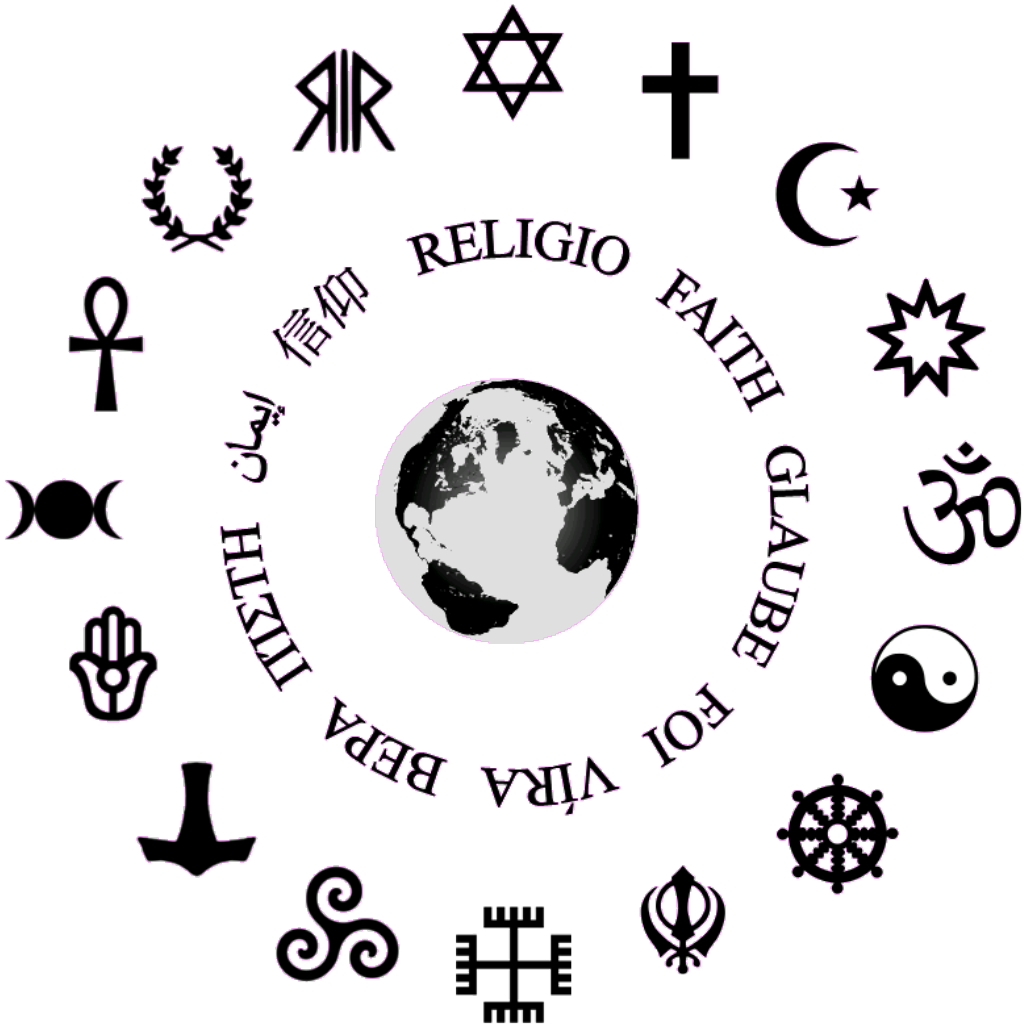


Table of Contents

I. Intro

For this project, our group analyzed the word counts over 8000 words in the following religious texts: Buddhism, Tao Te Ching, Upanishad, Yoga Sutra, the Book of Proverbs, the Book of Ecclesiasticus, and the Book of Wisdom. Along with the data set were prompts asking for: the top 20 words in each book, the top 20 words in all books, a model to determine how wording changed throughout the Old Testament books, and a hypothesis to find similarities in the Buddhism and Tao Te Ching text. Our overall hypothesis for this data set is that books with similar concepts and ideologies can be grouped based on relationships between contextual information in the data.

To properly visualize our results, we created a Tableau dashboard. The link can be found here: https://public.tableau.com/profile/oswald.harris?fbclid=IwAR3D7ZizfSwl15x7Laub3tjzOYBdUtD3FEXnCEJmj5jZtSzvk1XK-9OC_KI#!/

The link to our GitHub repository can be found here: https://github.com/cnaphade/datahacks2021

II. Data Cleaning

In our data cleaning stage, we removed words from the data set that were not real words (such as the letter 's', and the letters 'nt'). The most likely reason for these being in the data set is due to how the texts were tokenized. If the text were split along all punctuation, these letters would appear as words even though they are contractions. Further evidence of unreliable tokenization are combined words such as 'neitherpainfulnorpleasant'.

We also removed words that occurred zero times in the books, as only words that are in the texts should be kept. Finally, we removed words which occurred less than five times. The reasoning behind this is that such rare words should not influence our analysis.

The word counts for each text were standardized so we can analyze their distance from the mean, making it more convenient to see common and rare words.

III. Analysis

a) Top 20 Words

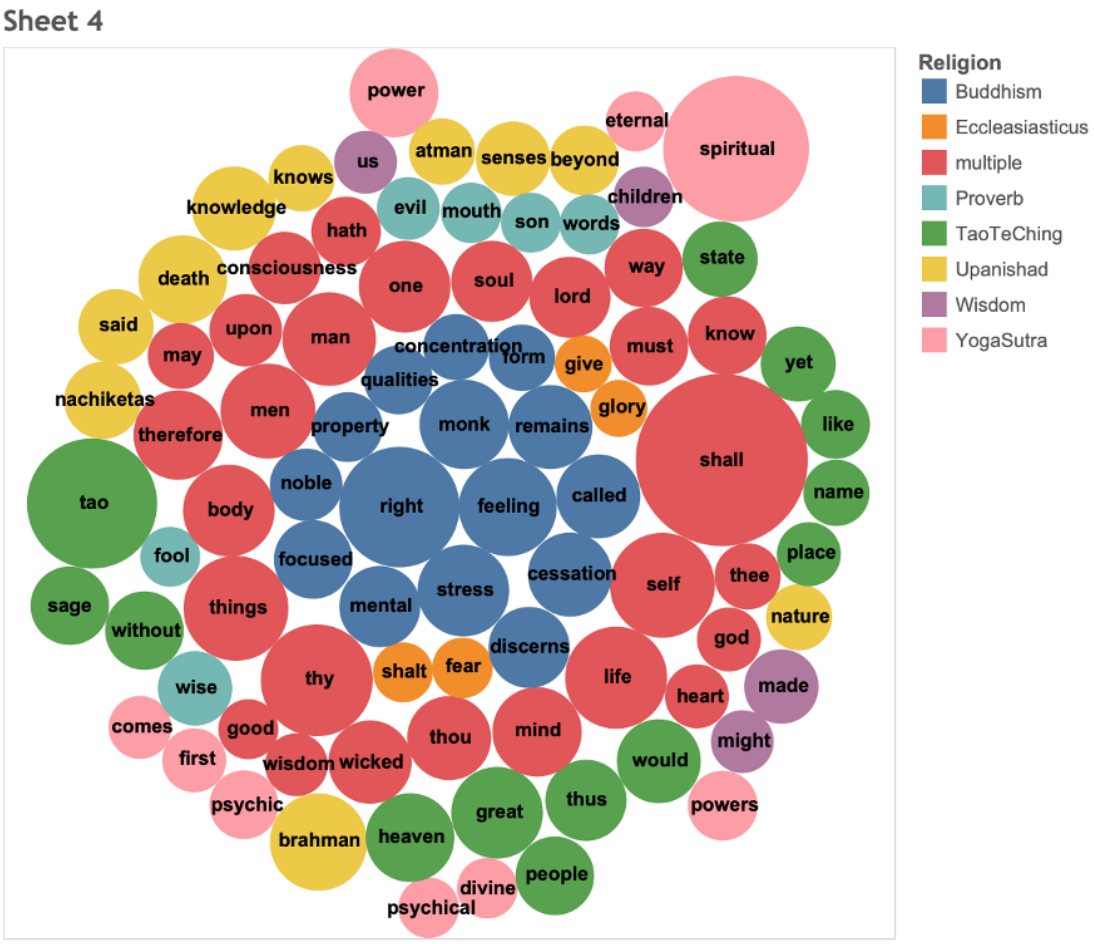
This is a table of the top 20 most common words in each text. It is sorted in descending order of frequency. As expected, the three books taken from the Old Testament have many words in common, such as: shall, man, god, lord, wisdom, etc.

	Buddhism	TaoTeChing	Upanishad	YogaSutra	Proverb	Eccleasiasticus	Wisdom
0	right	tao	one	spiritual	shall	shall	shall
1	feeling	things	self	man	man	thy	things
2	one	one	mind	life	thy	man	thy
3	stress	men	brahman	consciousness	thou	thou	god
4	body	great	man	power	wicked	god	thou
5	monk	therefore	death	one	lord	hath	wisdom
6	mind	heaven	knowledge	mind	wise	thee	man
7	remains	would	know	soul	hath	lord	upon
8	called	thus	must	things	heart	things	made
9	cessation	without	nachiketas	self	thee	upon	hath
10	mental	people	said	powers	way	wisdom	thee
11	discerns	sage	senses	psychic	evil	heart	men
12	focused	know	beyond	may	wisdom	good	lord
13	way	yet	atman	first	mouth	men	us
14	consciousness	state	nature	must	soul	fear	life
15	noble	way	knows	comes	son	soul	therefore
16	property	like	therefore	psychical	words	one	good
17	qualities	may	heart	divine	good	shalt	wicked
18	concentration	place	god	body	fool	glory	might
19	fabrications	name	body	eternal	things	give	children

The following is a table of the 20 most common words across all of the books. It is sorted in descending order. There is a wide variety from which these words come from. Some are more common in the Old Testament texts, and some are more common in the Eastern texts. For example, ‘shall’ was common in former while ‘mind’ was common in Buddhism.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Words	shall	man	thy	one	things	thou	god	life	hath	spiritual	lord	mind	thee	heart	soul	wisdom	men	upon	good	way

This Bubble Chart shows the top 20 words of each book. The size of the bubble represents the frequency of the word in the text. It is color-coded by which book the word is in the top 20 of.



b) Old Testament Analysis

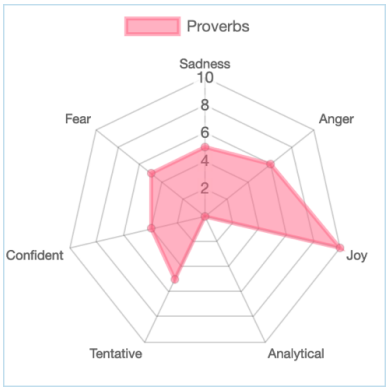
The Book of Proverbs began to be written in 700 BC, and the book of Wisdom was written in the mid first century BC. This is a span of over 500 years, and many societal and philosophical advances occurred. We would expect these to influence and change the common sentiment expressed in religious and philosophical texts of the time.

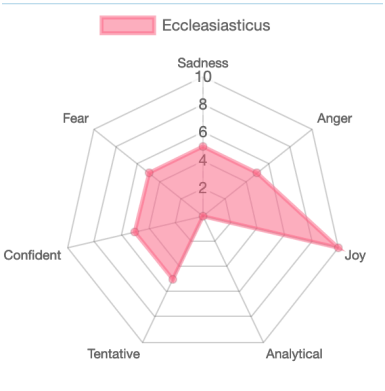
In order to determine the sentiment of the text, we used IBM Watson's Tone Analysis API in Python. This API has seven different sentiments it can predict: sadness, joy, anger, analytical, confident, tentative, and fear.

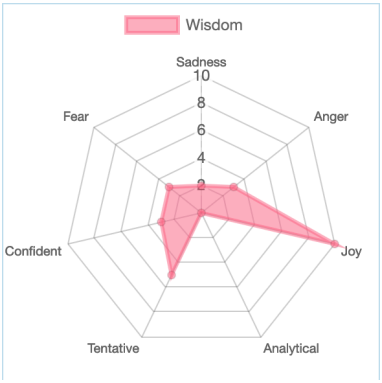
We ran IBM Watson's Tone Analysis on the words to obtain the sentiment for each word. The values of all words for each sentiment were added together to find sentiment values for each text (e.g., the standardized values of words classified as sadness were combined to find a sadness value). These sentiment values were scaled from 0 to 10 to find relative sentiment values.

Joy was found to be the most common sentiment in all three books. Sadness, anger, and fear decreased between the books. The Book of Proverbs had the most negative sentiment, The Book of Ecclesiasticus had lesser, and the Book of Wisdom had the least. Based on this, we fail to reject our hypothesis.

The results are visualized as radar charts. This type of chart was used because it allows for easy comparisons between groups. These results can be seen below.

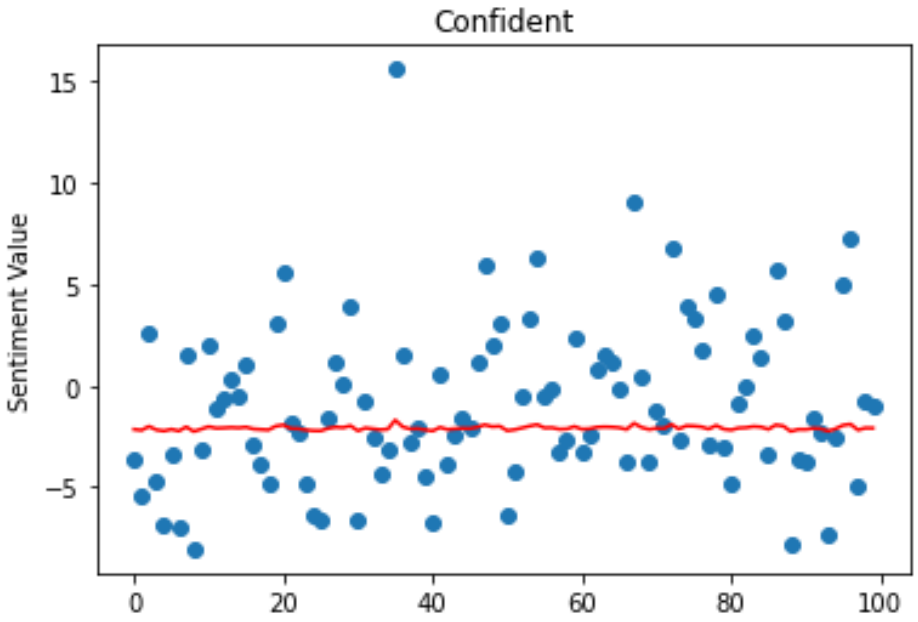




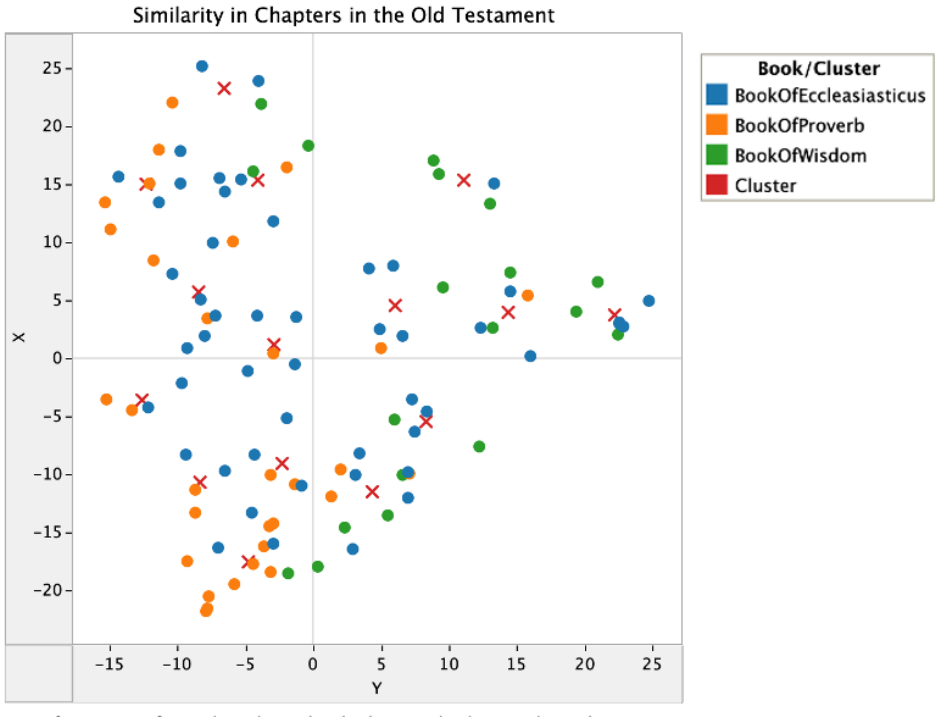


The difference between the Book of Proverbs and the Book of Ecclesiasticus lies mainly in the expression of anger and confidence. There is a stark difference between these two and the Book of Wisdom as there is very little negative sentiment in the Book of Wisdom.

Additionally, we attempted to use a linear regression model to predict how sentiment changes over time. In order to have enough data points, we ran this model on the chapters of the books rather than the books as a whole; however, this model was unsuccessful. The sentiment between chapters had too much variability compared to the books as a whole, so the model created a regression line with very little slope.



Below is a scatter plot containing each chapter in the texts taken from the Old Testament, along with crosses representing cluster centers. The dots are color coded based on what book they are taken from. This helps our analysis because it shows that chapters from different books may be more similar than other chapters from the same book.



c) Buddhism and Tao Te Ching Analysis

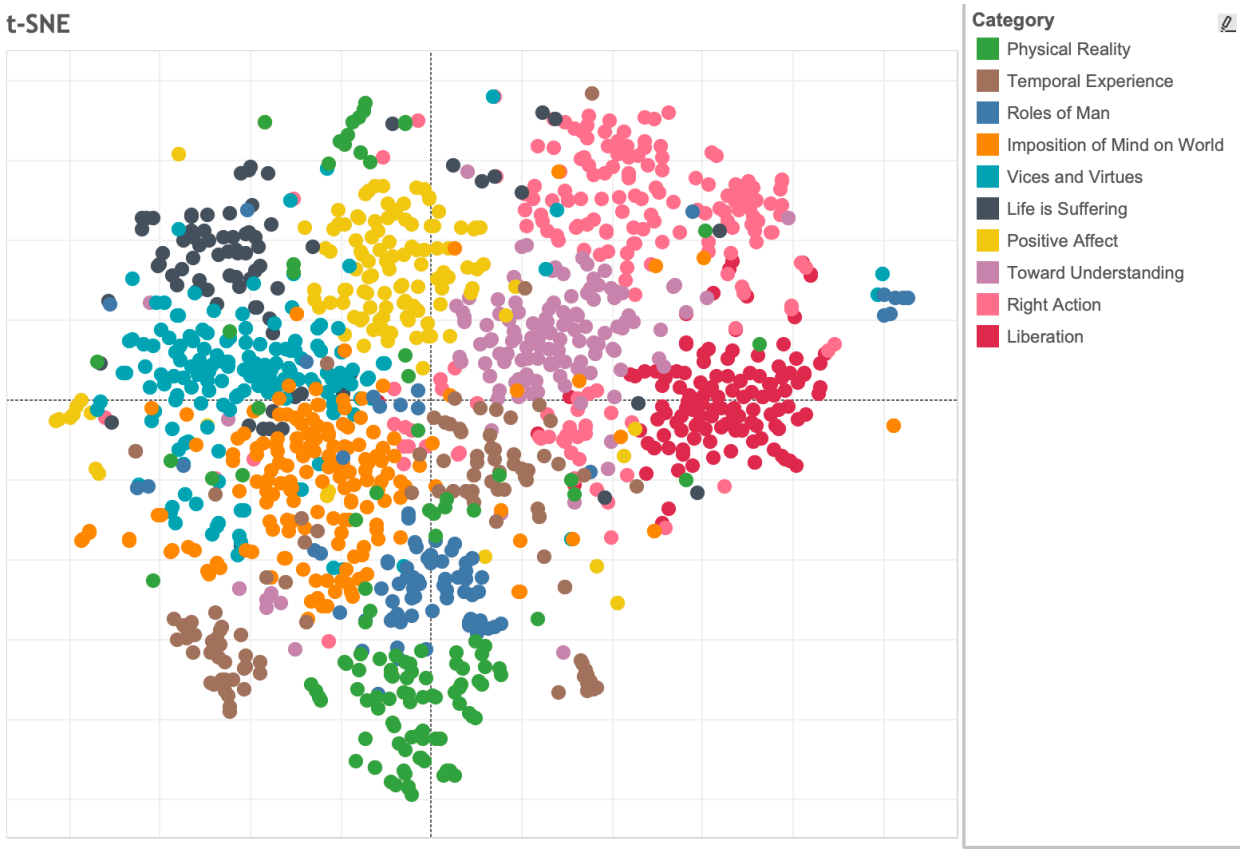
Buddhism and Taoism are both beautiful and complex philosophical schools of thought that originated in the Eastern World. To find similarities between the Tao Te Ching and the Book of Buddhism requires an approach that delves deep into the structure and meaning of these philosophical systems. If we were to find similarities based solely on textual analysis on frequency of words, we would be unable to find any substantial correlations between the ideas.

Hence, our approach has been to extract the meaning behind the words used in these texts to map conceptual similarities in these two philosophies. This method involves working with the concepts behind the words rather than simply looking at the words themselves.

We hypothesize that the words in these texts represent two very similar paths toward attainment of salvation. Our goal was to use a similarity measure based on the concepts that the words represent to visualize and discover novel insights about the associations between the two texts. We used GenSim, an open-source library that allows users to represent text as semantic vectors to find semantically related documents. To further our rate of progress, we used FastText pretrained word vectors learned on Common Crawl data of over 2 million words that contains word-embeddings we could use for our dataset.

We updated the model with our data and used the word vectors it provided for us to visualize the relationships between the meaning of the words using K-Means Clustering, PCA and t-SNE. We had to use dimensionality reduction techniques to visualize a 300-dimensional vector (that represents its relation to every other word) on a 2D graph! We assumed that the reduction would cause a loss in the final analysis of our data. However, what we found was truly phenomenal.

After our initial visualization of the data, we chose to categorize it using conceptual ideas as labels. The visualization makes a natural progression through the path to enlightenment that is common to both philosophies.



On Tableau, if you hover over the points on the plot, you can see which words they represent, the philosophical text they belong to, and the conceptual category expressed in them. The natural progression of the path to enlightenment begins at the very bottom in the category of Physical Reality, with self-consciousness of the body and the experience of the natural world. It proceeds upward toward the experience of temporal phenomena and an identification with social roles. From there, it evolves into an intellect capable of organizing and imposing itself on the world to create civilization. With the development of a sharp mind comes a realization of good and evil, with character traits expressing vices and virtues. The most interesting observation of all is that the transitions between the categories seamlessly tie them together. Hovering over the data points near the boundaries of the conceptual categories reveals an almost

divine interconnectedness. The evolution of these ideas then continues to the first Nobel Truth of Buddhism, all life is suffering. The Tao Te Ching contains almost equal instances of these ideas in our graph. We can see the non-dualistic elements of both philosophies shining through in this category, with words like pleasure, longing, and craving grouped with pain, sorrow, and despair, reflecting the second Noble Truth of Buddhism (the cause of suffering is desire).

One drawback of using clustering to create the categories is that the boundaries sometimes overlap, with words most similar to one label jumping over to another. In our graph, this leads to a gradient of similarity between the points on the boundaries. Furthermore, dimensionality reduction with PCA and t-SNE is bound to lose information about the data. However, we believe that we were able to see concrete trends and associations regardless.

We encourage you to further explore our visualization on Tableau. What will you discover?

IV. Proposal

One possible way we can improve our analysis is to use a tone analyzer with more prediction options. Only having seven options for sentiment is limiting and does not properly express the range of human emotions in a text. Another avenue of future work would include using other classification algorithms for different insights on our data. Furthermore, having more text to work with is the best way to improve our analysis. There were only 8000 words in the data set, but the English language has over 170,000 different words, so this data set only scratched the surface on what is possible to analyze. A potential application that could extend the capabilities of the semantic clustering with the t-SNE visualization we created is using it on more philosophical texts so that we can find something akin to Aldous Huxley's Perennial Philosophy. The Perennial Philosophy that we extract from more texts could potentially tie together the underlying foundations of all transcendental theories. We see tremendous scope in this technology, since it can help us uncover elemental narratives in texts.

V. Conclusion

To summarize, we uncovered what seems to be profound similarities across the different religious texts. For instance, the expression of sentiment in the three books from the Old Testament changed over time. But perhaps the most striking of all was the way in which the Tao Te Ching and the Book of Buddhism shared such far-reaching similarities when we grouped them by concepts rather than words.