Geographic Grouping of Religious Texts Through Historical and Quantitative Analysis

Team: All-Nighters (Francisco Chavez, Alec Panattoni, Tyler Tran, Barry Xue)

Introduction

The practice of religion is an integral part in the lives of a large percentage of the global population and has thrived for centuries, largely thanks to important religious doctrines recorded written to house a religion's ideals and practices. Of particular importance are eight fundamental texts of some of the world's major religions including: the Yoga Sūtra of Patañjali, the Upanishads, the Book of Wisdom (Wisdom of Solomon), the Book of Proverbs, the Book of Ecclesiastes, the Book of Ecclesiasticus (Book of Sirach), the Tao Te Ching, and sacred Buddhist texts which we assume to be related to the Pāli Canon. Contextual background on each of these texts is included below.

The Book of Proverbs, written around 700 BCE, remains significant to two of the world's most prominent religions, serving as a section of the Hebrew bible and a book within the Old Testament. Common themes discussed within include personal values, moral behavior, the meaning of human life, God, and is theologically founded on the idea that "submission to the will of God is the beginning of wisdom". [1]

The Book of Ecclesiastes, is one of the "wisdom books" of the Old Testament, and of Hebrew origin. It also holds a place within the Megillot, a collection of scrolls read at various Jewish religious festivals.[2] Often attributed to Solomon, the book's teaching stresses ideas about wisdom, death, the meaning of life, enjoyment, and God. It is thought to be written between 450 and 200 BCE.

[1] https://en.wikipedia.org/wiki/Book_of_Proverbs
[2] https://www.britannica.com/topic/Ecclesiastes-Old-Testament

Sirach, or the Book of Ecclesiasticus, is another portion of the Old Testament, and of the Hebrew bible. Attributed to Ben Sirach (Ecclesasticus), it was thought to be written around 200 BCE. Notable themes include praise of wisdom, duties to God, friends, parents, others, rules of correct self-conduct, prudence, use of wealth, freedom of choice, and God's gifts. [3]

The Wisdom of Solomon (Book of Wisdom) is of Hebrew origin and a part of the Old Testament. Although it's author is unknown, it is attributed to Solomon and thought to be written around 25 - 75 BCE. Common themes include wisdom, man, God, death, and immortality.[4]

The Pāli Canon, also known as the Tripitaka and as the "Word of Buddha", is the oldest complete canon with Buddhism. Of Indian origin, its writings are thought to have been oral ideas passed down by Gautama Buddha, and was recorded sometime around 29 BCE. It's teachings emphasize respect, peace, purity, life consciousness, meditation, and lists rules for monastics to follow.[5]

The Yoga Sūtra of Patañjali is a text associated with the practice of yoga which has roots in both Hinduism and in Buddhism. Attributed to the Sage Patañjali between 500 BCE and 400 CE, it emphasizes the ideas of sensual impressions (conscious mental states) that are either conductive or detrimental to the goals of Yoga. It also covers the Eight Limbs of Yoga, emphasizing the idea that spiritual practice should be centered on self-reflection.[6]

The Upanishads, thought to have come from the Indus Valley, are a set of late Vedic Sanskrit texts that form the basis of Hinduism. It's central theme is the idea of liberating the soul and returning to the world of Brahman. Through Karma, Samsara, Dharma, one aims to achieve Moksha, or the release from the endless cycle of rebirth.[7]

---

[3] https://bible.usccb.org/bible/sirach/0
[4] https://en.wikipedia.org/wiki/Book_of_Wisdom
[5] https://en.wikipedia.org/wiki/P%C4%81li_Canon
[6] https://chopra.com/articles/yoga-sutras-101-everything-you-need-to-know
[7] https://en.wikipedia.org/wiki/Upanishads

The Tao Te Ching, the sacred texts of Taoism, delivers lessons on how to live in the world with goodness and integrity. Written around 400 BCE, it emphasizes self-reflection and self-awareness.

Data Exploration and Preprocessing

Data exploration began with surveying the given training dataset to get a general sense of how observations and variables were kept. Notably, each row was recorded as a chapter of one of the eight religious texts, with columns housing the words.

	Chapters	foolishness	hath	wholesome	takest	feelings	anger	vaivaswata	matrix	kindled	...	erred	thinkest	modern	reigned	sparingly	visual	tho
0	Buddhism_Ch1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
1	Buddhism_Ch2	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
2	Buddhism_Ch3	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
3	Buddhism_Ch4	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
4	Buddhism_Ch5	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	

As we were more interested in the appearance of words within a specific book, rather than a specific chapter of a book, we molded the data to summarize word appearance in each of the eight texts.

Books	foolishness	hath	wholesome	takest	feelings	anger	vaivaswata	matrix	kindled	convict	...	erred	thinkest	modern	reigned	sparingly	visu
Buddhism	0	0	0	0	19	0	0	0	0	0	...	0	0	0	0	0	
Eccleasiasticus	0	189	3	1	0	14	0	0	3	0	...	0	0	0	1	1	
Ecclesiastes	0	46	0	0	0	5	0	0	0	0	...	0	1	0	0	0	
Proverb	2	65	0	0	0	11	0	0	0	0	...	0	0	0	0	0	
TaoTeChing	0	0	0	0	0	1	0	0	0	0	...	0	0	0	0	0	

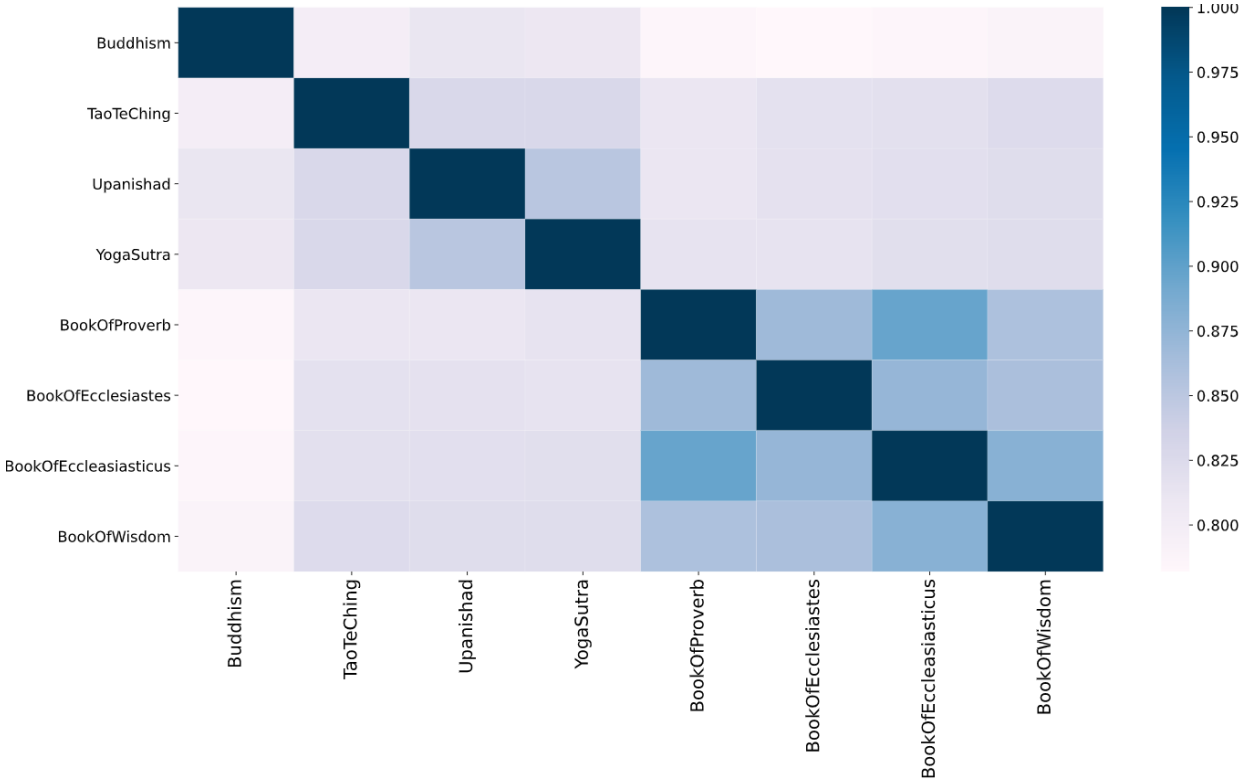
Initial data exploration and preprocessing continued by querying the data to create tables representative of each of the eight books, holding only columns (words) that were actually present in the book. Included here is an excerpt from the dataframe for the Tao Te Ching.[8]

	hath	wholesome	takest	anger	kindled	open	rage	looketh	prosperous	lambs	...	state	stained	aromatical	admireth	taketh	kettle	reigned
521	11	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
522	5	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
523	6	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
524	0	0	0	1	0	0	0	1	0	0	...	0	0	0	0	0	0	0
525	1	0	0	0	0	0	0	1	0	0	...	0	0	0	0	0	0	0

[8] https://en.wikipedia.org/wiki/Tao_Te_Ching

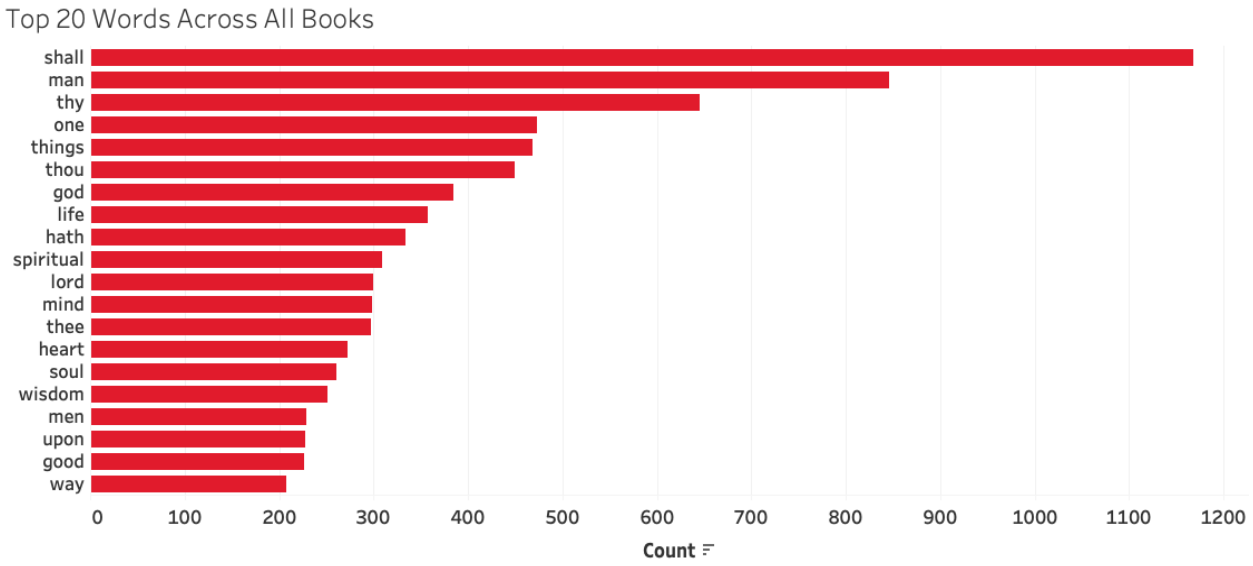
Analysis

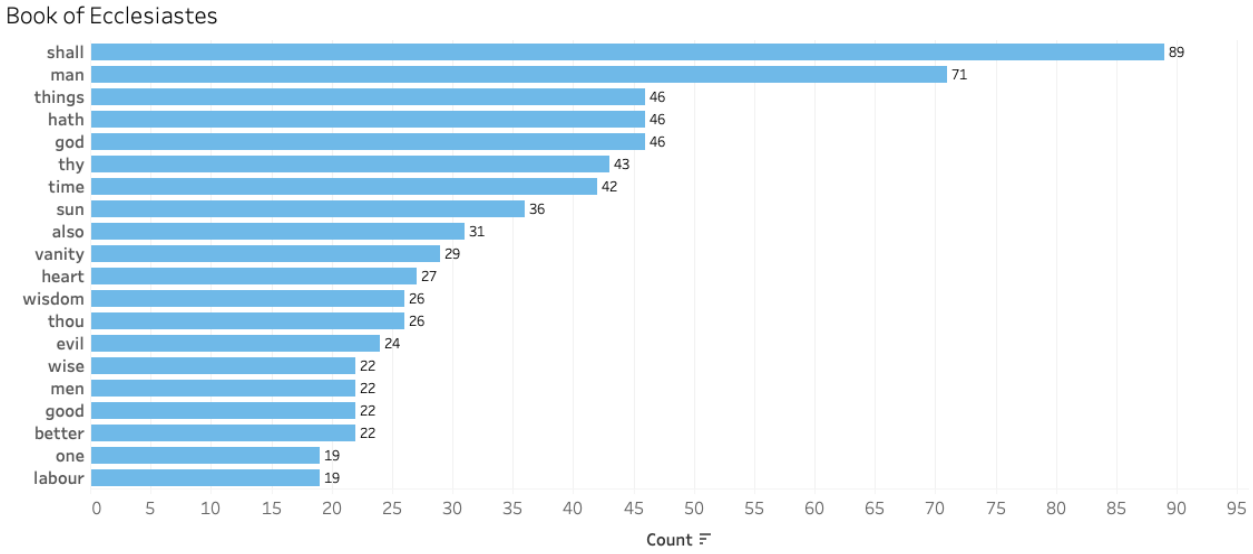
To analyze the composition and themes of each of the books, we sought to determine the most commonly occurring words within each book, and within all eight of the religious texts combined. In this task we elected to consider all words in the dataset.



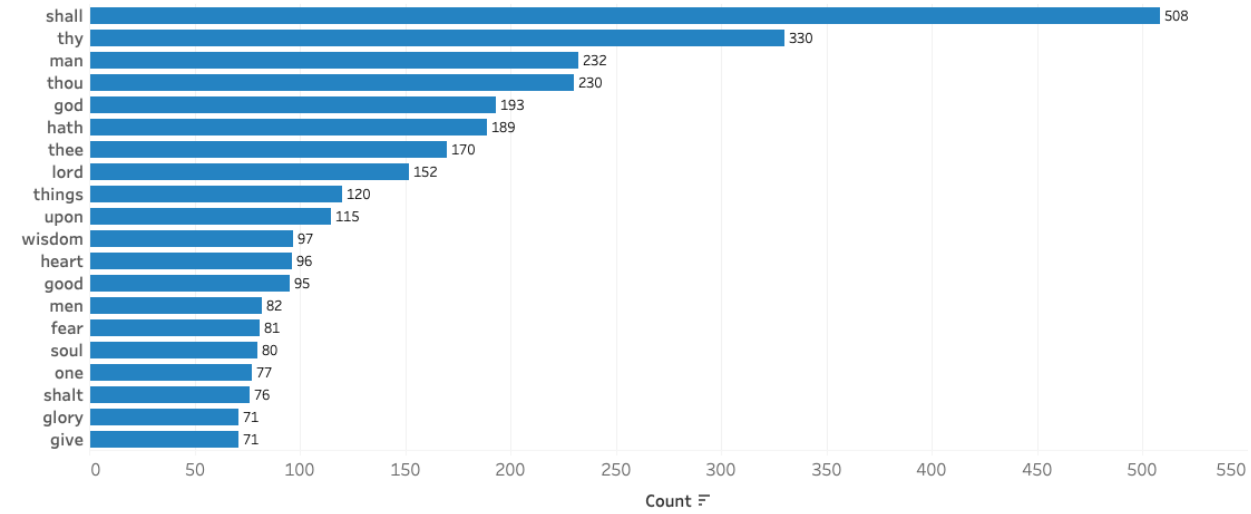
This is a heatmap showing the similarities and differences between all the books. It provides a nice overview of the general relationship between all the books. The four books: The Book of Proverbs, and the three books from The Old Testament share a great deal of vocabulary compared to others. Without looking into the meaning of the vocabulary used, there is already a noticeable trend across the books which implies there might exist a larger framework that can generally classify the books. To better the understanding of the text, it is good to look into the meaning and context of each book's vocabulary.

Common words, such as “i” would be expected to be prevalent in religious texts serving as instructions to individuals seeking to learn religious tenets. Furthermore, we anticipated seeing words associated with prominent religious figures and general teachings of these religions, such as “wisdom” and “God”. Included below are the top twenty words of each particular book, and of all books together.

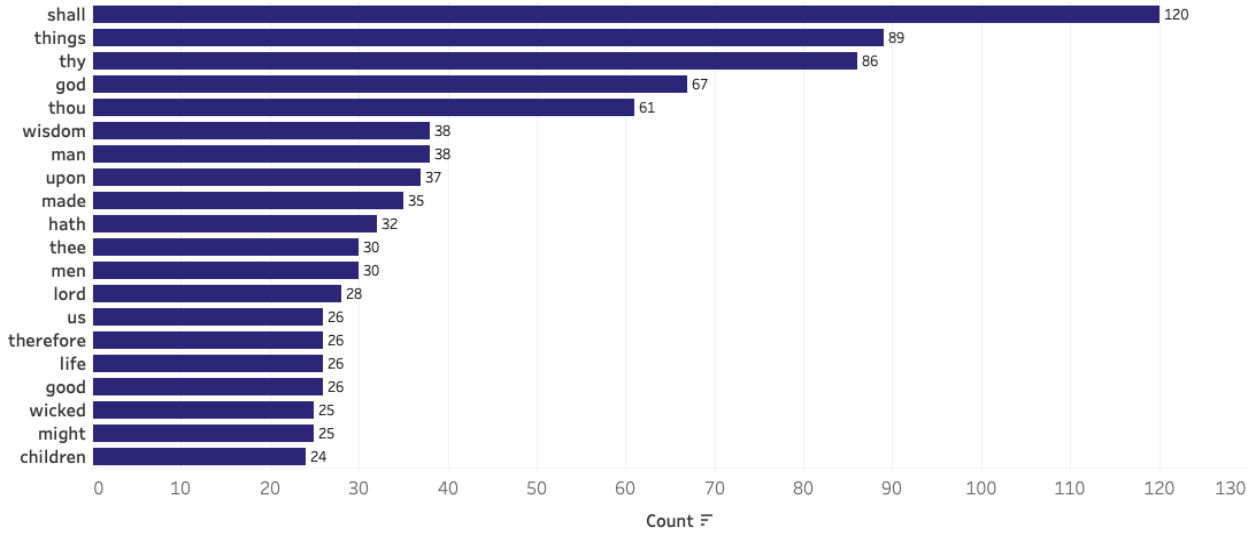




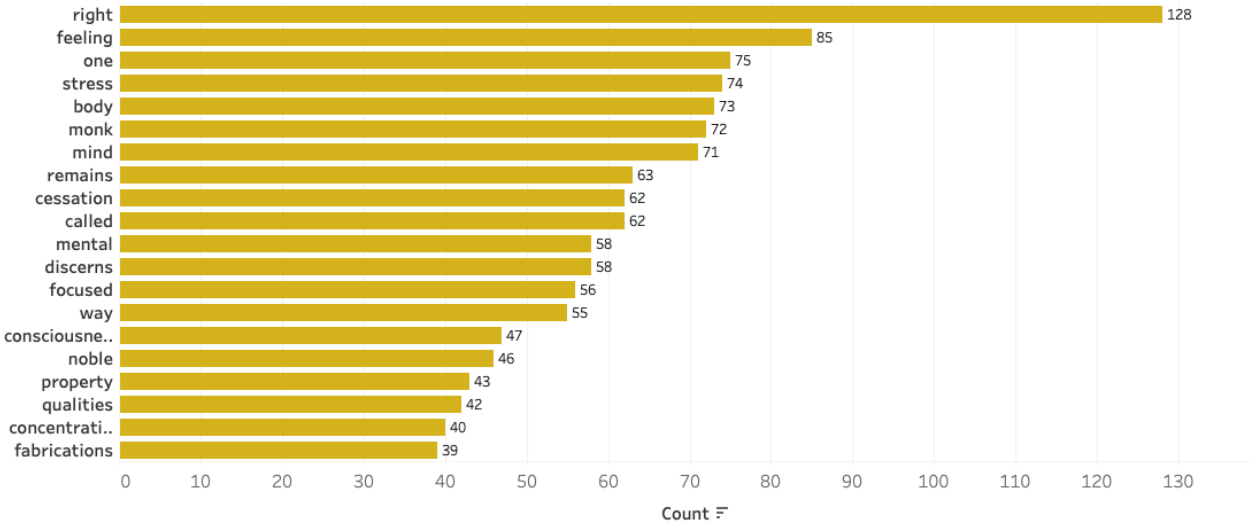
Book of Ecclesiasticus



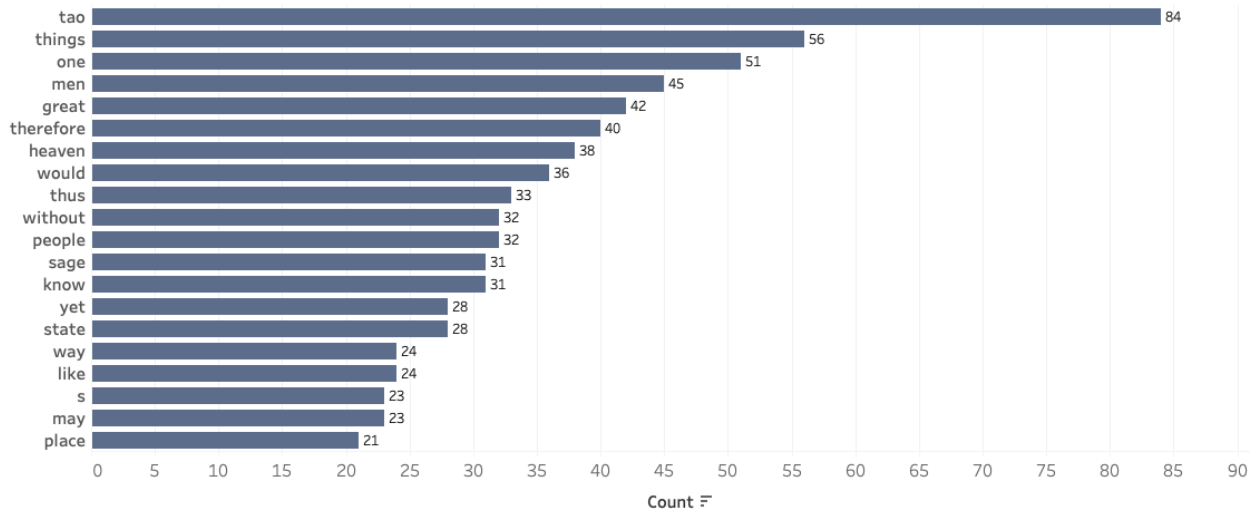
Book of Wisdom



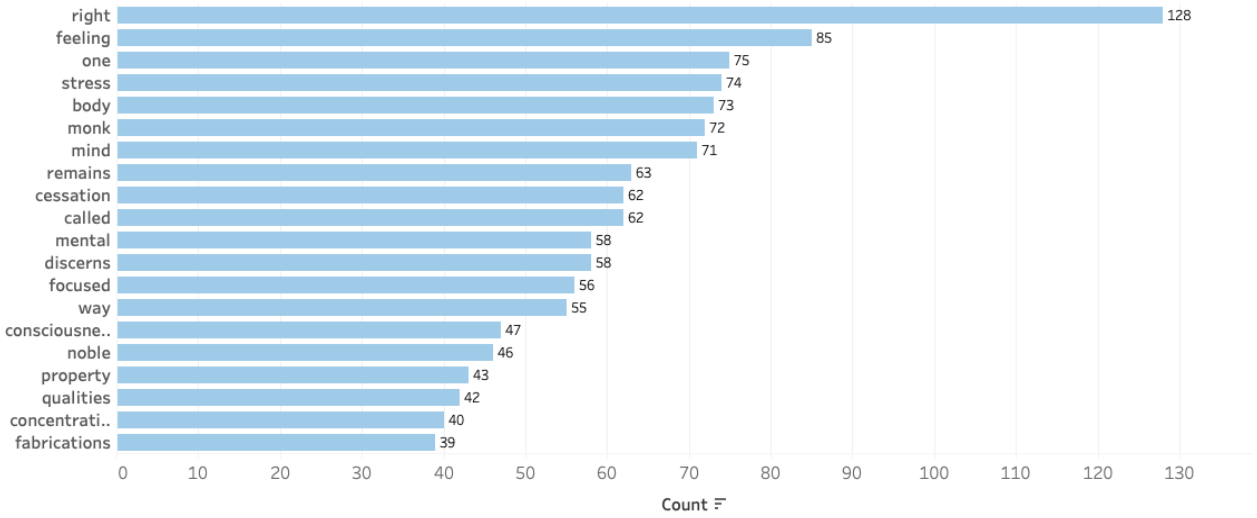
Unnamed Buddhist Text



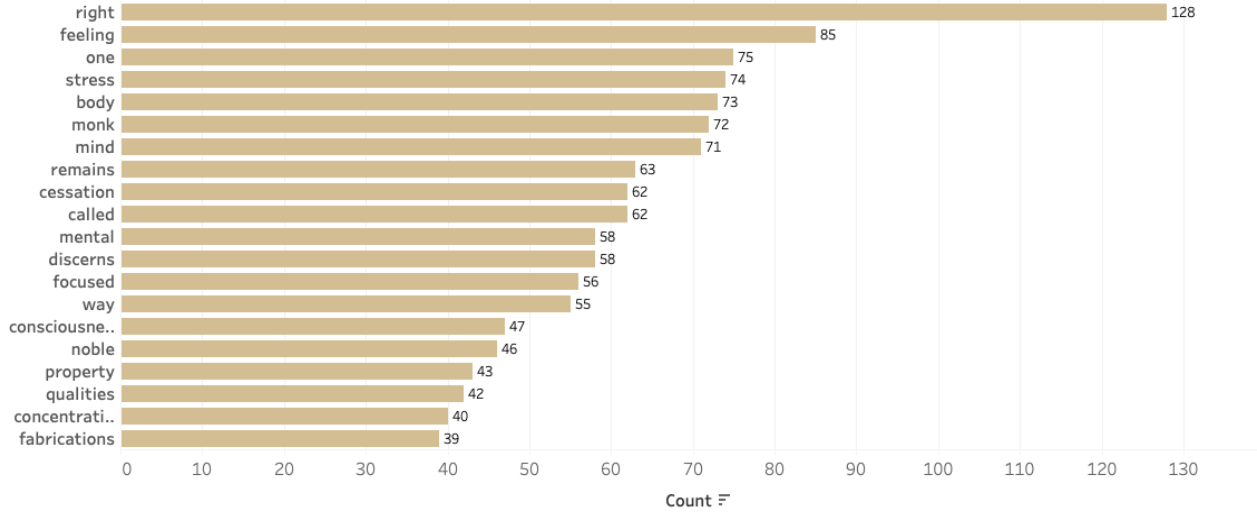
Tao Te Ching



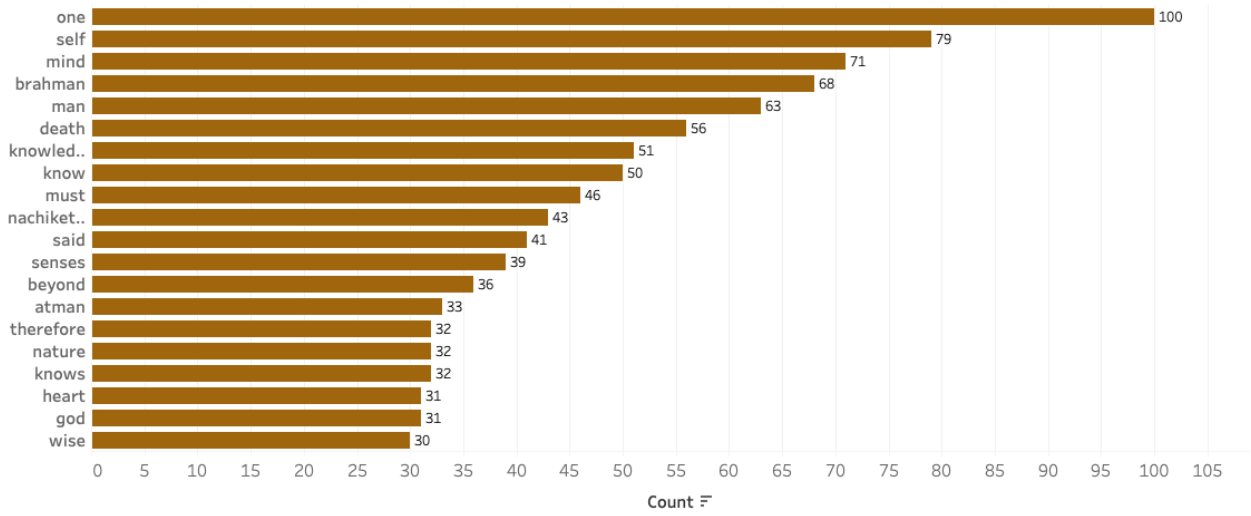
Book of Proverbs

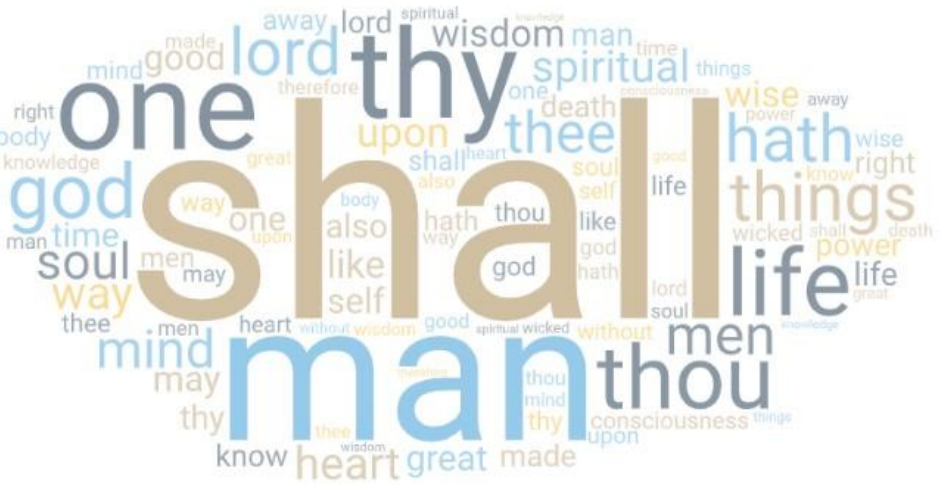


Yoga Sutra

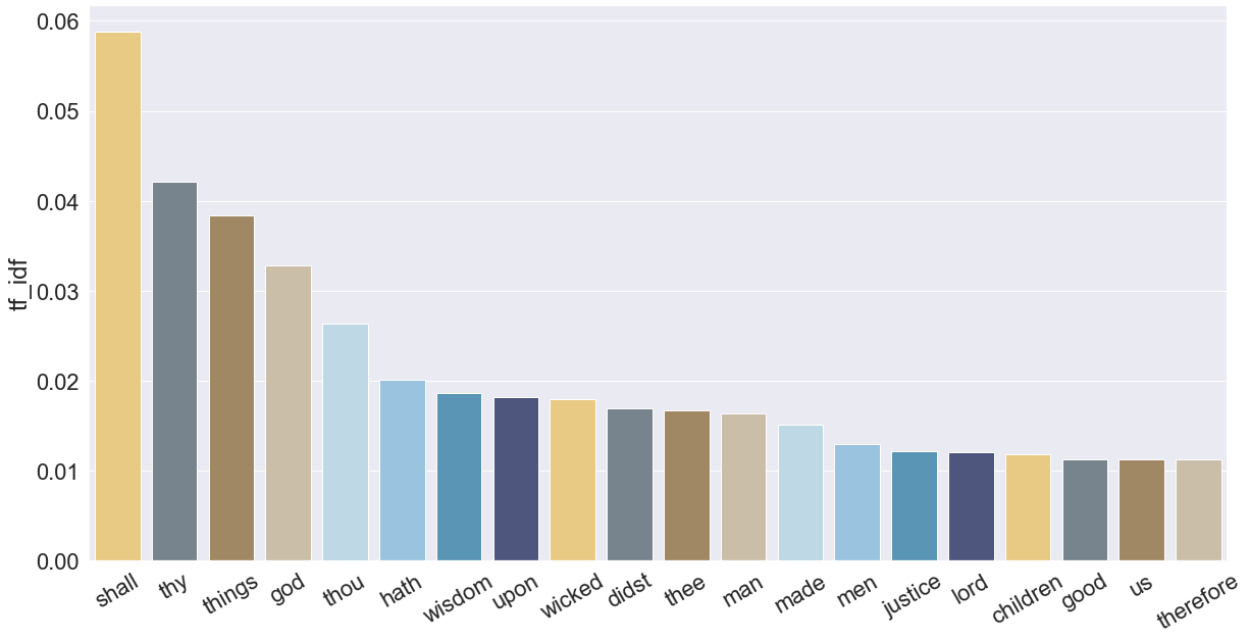


The Upanishads



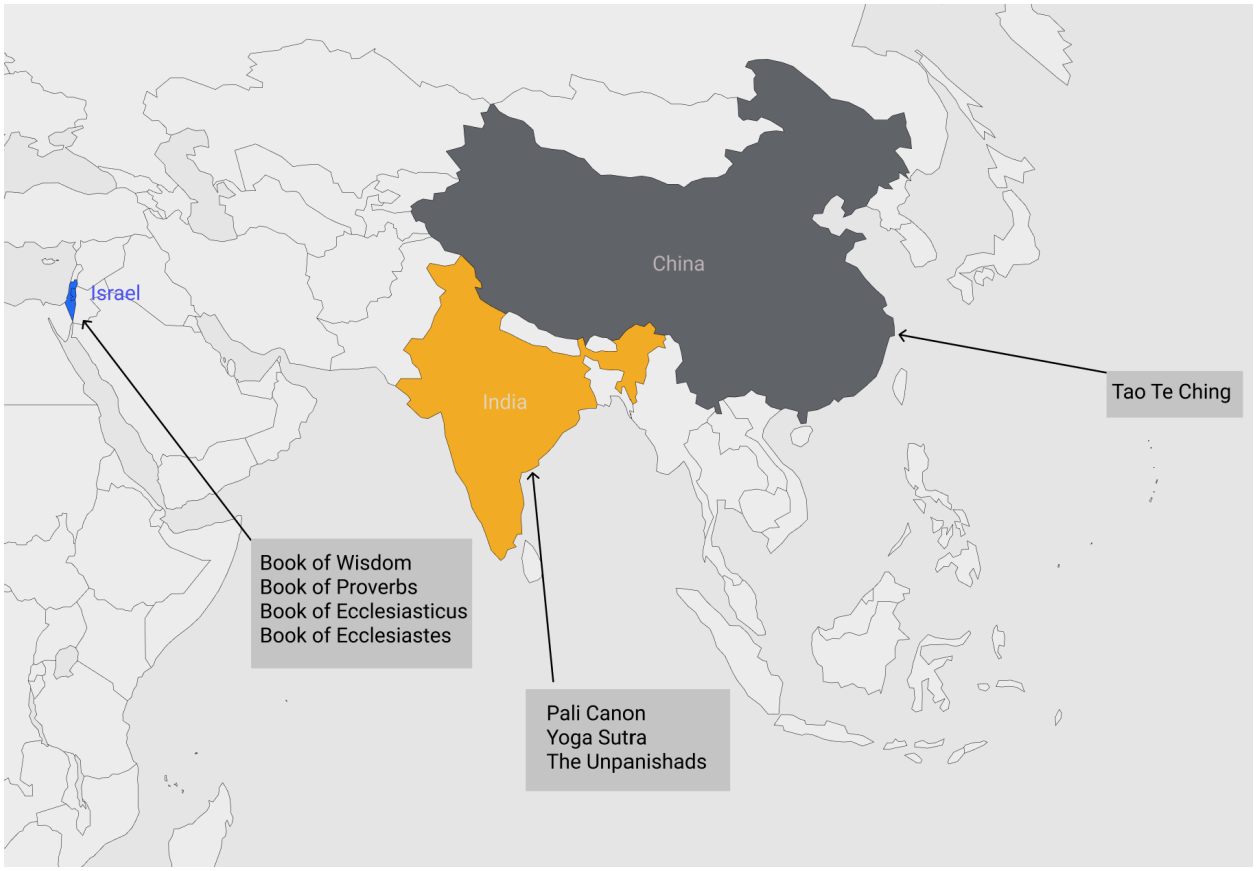


Of equal interest to the analysis was the relative importance of a word to a religious text. This was measured using TF-IDF (term frequency-inverse document frequency). TF-IDF scores were measured to determine the most relevant words to each text. Specifically term frequency was calculated as the occurrence of the word within the target book divided by the total number of words within the target book. Similarly, inverse document frequency was computed by finding the log of the number of books divided by how many of the eight books contained the specific word. Basic normalization was performed to prevent bias in term frequency based on the relative total lengths of each of the books. Included below is a subset of the findings, showcasing the twenty most important words to the Book of Wisdom as measured by TF-IDF.



Comparative analysis of texts was also conducted based on cursory background research. As many of these texts stem from the same religion, we sought to explore if religiously related texts would have a similar set of most common words. As expected, the underlying religion a book belonged to was a relatively good measure of a book's similarity based on common words. For example, the Book of Proverbs and the Book of Ecclesiastes-- both of which are included in the Old Testament-- shared eleven of their top twenty most common words and nine words based on TF-IDF comparison.

Intriguingly, the unnamed Buddhist text bore little similarity to the Tao Te Ching. Instead, it was more closely related to both the Yoga Sūtra and the Upanishads. Thus, we assumed it was likely that the unnamed text could bear resemblance to the Pāli Canon, one of the most prominent Buddhist texts, but that could not be verified. Moreover, it was hypothesized that the discovered similarity could be attributed to the common origin region of these religious texts rather than underlying religion.



Proposal

Therefore, we hypothesize that we would see strong similarities between religious texts coming from the same geographical region, where each region's texts would show large differences from the texts of other regions, as measured by the popular words of the region's texts and their important words, as measured by TF-IDF.

When we began grouping these books by religion, we found that there were five groups that emerged based on the books coming from the same religion. Specifically, the Book of Wisdom , the Book of Proverbs, the Book of Ecclesiastes, and the Book of Ecclesiasticus all come from Judeo-Christian faith. Buddhism, Hinduism, and Taoism, all warranted their own categories as religions, with the Yoga Sutras also getting their own categorization for comparison against books of the other groups. When we began by comparing the books Proverbs and Ecclesiastes, we found that they shared 9 similar TF-IDF keywords, and 11 similar words in their respective top word counts. When we compared the Book of Wisdom with the Ecclesiastes, we found that they shared 10 similar TF-IDF keywords, while also sharing the same 10 keywords in their top word counts. When we compared the Book of Wisdom with the Ecclesiasticus, they shared 13 similar TF-IDF keywords, while also sharing the same 13 identical words in their top word counts.

When we approached grouping these books by geographical region, we found that there were less groups than when we grouped by religion. This likely happened because of historical and thematic similarities indicative of their origin region. For example, the books Yoga Sutra and the Buddhism text share two similar TF-IDF keywords, while also sharing top counts of the keywords “one,” “body,” “mind,” and “consciousness.” Then, when we compared Upanishad with Yoga Sutra, they shared 5 keywords in terms of TF-IDF, while also sharing top counts of the keywords “one,” “self,” “mind,” and “man.” When we compared Upanishad and Buddhism, they shared 2 TF-IDF keywords named “one” and “mind,” while also sharing these same keywords for top word counts. This has led us to believe that these three books, the Upanishad, Yoga Sutra, and the Buddhism text, all share similarities regarding self-ref keywords. It is also important to note that the Yoga Sutra shares the Hindu system of Dualism, and the religion Buddhism agrees with many Hindu fundamentals such as karma, dharma, moksha, and reincarnation.

It is clear that historical context plays a massive part in the comparisons that can be made amongst these religions, even if the TF-IDF algorithm says otherwise. This is especially true with the comparison of the Buddhist and Taoist texts. The algorithm indicates that there is little

to no relation between the texts. However, based on the prevalence of the word “monk” in the popular words for the unnamed Buddhist text and the potential similarities to the famous Buddhist text, the Pali Canon, it is likely that the Buddhist text predates Chinese Buddhism. This would be in line with our hypothesis that religious text originating from the same region would be similar while those from other regions would strongly differ. Modern associations between Chinese Buddhism and Taoism are potentially further proof of our hypothesis, as similarities between those sects of Buddhism and Taoism may have increased as Buddhism’s popularity in China expanded. Furthermore, the Tao Te Ching was written in Classical Chinese, which is incredibly difficult to translate. As such, the chosen words for translation may be simply a poor comparison to the other religious texts present. Therefore, with the historical pretext of potential issues with the translation of the Taoist text and the uncertainty of the popularity of the Buddhist text in China at the time, the result might not be representative of the reality. Thus, it can't be used as evidence to conclude that there are no associations between the two texts.

Conclusion

In exploratory analysis, it was discovered that some religious texts thought to be similar in theme, such as the unnamed Buddhist text and the Tao Te Ching, were not quite as predictive of similarity by TF-IDF or by most common words. Instead, geographical origin of the text was shown to be a better indicator of similarity for these measurements. This was confirmed by grouping the texts based on known geographical region and comparing TF-IDF and count values against grouping by religious origin. The Judeo-Christian texts, for example, all come from the same region. After grouping the text based on geographical origin, the groupings were compared against each other, confirming that the most prominent and popular words of these religious texts are a product of their location. Therefore, it is possible to say that religious texts coming from the same geographical region likely have similar themes given their similarity in common words and in important words as measured by TF-IDF.

Furthermore, using the Random Forest classifier, we built a prediction algorithm that predicts the origin of a given religious text. After training it with the training data, the prediction result from

the testing dataset shows that there is a lack of accuracy and it is heavily biased towards the three Judeo-Christian Texts. One reason for this phenomenon is that there are many similarities between the three texts due to their geographic proximity. Although proving to be less effective in practice, the result from the classification problem confirms our previous findings on the correlation between geolocation and the vocabulary usage in each text.

Finally, although the unnamed Buddhist text and the Tao Te Ching bore little resemblance to each other in both most common words and in most important words, the lack of similarities can be qualitatively explained by the history of the Tao Te Ching, and a likely candidate for the Buddhist text represented. Moreover, these differences do not necessarily extend into the practices of these religions today, as they have evidently influenced one another as Buddhism historically expanded beyond its origins in India.