Religious texts study

Ivan Alejandro Diaz Cardenas
23036361
3/12/2023

Introduction

According to a global poll in 2012 by WIN-Gallup International, 59% of the world said that they think of themselves as religious person, 23% think of themselves as not religious whereas 13% think of themselves as convinced atheists.[1] This sums up to almost two thirds of all humans. Most religions are based on scriptures and texts that have been passed generation to generation, creating a huge database of moral and lifestyle rules and stories. It would take a lifetime to study all these religions and their respective textbooks, so the main goal of this project is to analyze them with the several NLP methods were given to us this term, such as Topic Modelling, Sentiment Analisis and Markov Chains and try find similarities and differences between each other. Also, I wanted to create a 'inter-religious' text for my artistic investigation. To do this project, I've chosen 8 religious books or compilation of texts as datasets. The texts I have chosen are: The King James Version Bible (Christianity), The Quran (Islam), The Bhagavad Gita (Hinduism), Tripitaka (Buddhism), Hebrew Bible (Judaism), Bahá'í Writings (Bahá'í Faith), TAO TE CHING (Taoism) and the Kojiki Texts (Shinto).

I would like to think that as humans, independent of our cultural background, we all thrive towards the same goals. In my opinion, even though religions have made many societies thrive and were completely necessary to current society evolution, they have also created and widened gaps. On several occasions, these texts have been misinterpreted and used as an excuse for fulfilling private interests. A main goal of my artistic research is to find the similarities and differences between these texts.

Background

"Sentiment Analysis" is the most common text classification tool that analyses an incoming message and tells whether the underlying sentiment is positive, negative or neutral." (Gupta, 2018) Commonly used in reviews context, this method will be remarkably interesting because it assigns a positive, negative, or neutral rating to the texts, so it will allow to find which religious text has more positive sentiment analysis than others. Also, I will perform Markov Chains. A Markov chain is a "stochastic model created by Andrey Markov that outlines the probability associated with a sequence of events occurring based on the state in the previous event." (Patel, 2022) This has a more artistic objective, to create a new religious document, based on the sum of all of them. I have hope that in a not very distant future, humans can forget their

1. [1] *"Global Index of Religion and Atheism: Press Release" (PDF). Archived from the original (PDF) on 16 October 2012. Retrieved 1 December 2023.*

differences, sometimes highlighted by religions, and understand that we are all the same, independent of our minor genetic differences. Regarding the data domain, most of the religious texts were open source, due to their religious nature. The only consideration would be the translated version. I found in my investigation a paper by Zemlyanker, D. (2022). *Using natural language processing to analyze religious text.* In his paper he also investigates different religious texts and uses NLP to analyze them and find conclusions. The data set he uses is based on is different texts than mine, and he uses specific techniques that I did not use. But it was interesting to understand a different approach to a similar project.

Method

First, I used simple Web Scraping to get the full text from different websites that offered English translations. I based all my NLP techniques on the notebooks given to us throughout this term. Regarding finding the corpus, there were some .txt files that were easy to download directly (such as the K. J. Bible from Project Gutenberg) but other texts such as the Bahá'í Writings had to be scraped from a website. I used this opportunity to practice simple web scraping techniques, including using web-scraper plug-in to get a .json file and together with beautiful soup to only extract the content necessary for the project.

The next step was to check their length. There was one web scraping text file (Tao Te ching) that was 6 times in size in comparison to the King James Bible, therefore I reduced the file size, so it was more similar in length.

After I had my datasets ready it was particularly important to clean the datasets. Religious texts have many chapters and most of the text is not often done in prose, but verse. First I used regex to remove all the numbers and special characters.

```
regex = '[\\!\\"\\#\\$\\%\\&\\\'\\(\\)\\*\\+\\,\\-
\\.\\/\\:\\;\\<\\=\\>\\?\\@\\[\\\\\\]\\^_\\`\\{\\|\\}\\~]' #Eliminate special
characters

    new_text = re.sub(regex, ' ', new_text)  # Eliminate punctuation

    new_text = re.sub("\d+", ' ', new_text)  # Eliminate numbers

    new_text = re.sub("\\s+", ' ', new_text)  # Eliminate spaces
```

My first attempt was using a library from python named clean-text, but it actually cleaned more than what I needed, so I decided to use Regex instead, so I had more control over what I was removing.

After cleaning with regex, I had to find which specific stop words to use, so the final dataset made more sense.

```
religious_stop_words = ["hath", "ye", "thy", "lo", "thou", "thus", "name"]
```

After this I performed lemmatization and tokenization with this algorithm (provided in class):

```
# Function originally from:
https://www.programcreek.com/python/?CodeExample=get%20wordnet%20pos

def get_wordnet_pos(word):

    tag = nltk.pos_tag([word])[0][1][0].upper()

    tag_dict = {"J": wordnet.ADJ,

                "N": wordnet.NOUN,

                "V": wordnet.VERB,

                "R": wordnet.ADV}


    return tag_dict.get(tag, wordnet.NOUN)
```

My first approach was to use the dataset as a whole, but I realized that, since I was not familiar to all the religions, it was better to apply the algorithms to each individual text first.

My investigation continued using Counter from Collections library to perform a word count for each of the religious texts and then find the first 500 most repeated words for each text. Which gave remarkably interesting insights. I was surprised how efficient these tools are to get a rough but accurate idea on massive texts. For example the main common word in all of them is a variation of ‘god’ (the name each text books gives to their own god), but in the King James Bible the most repeated word was ‘shall’(Just an interesting fact)

To map it in a more efficient manner I decided to use WordCloud library, that makes into a graphic the most repeated and important words in size to get a more interesting visual representation for each text. After this I performed sentiment analysis on each of the texts, to find out which were more "positive” than others. (the results are available in the git repository).

Next step was performing a TF-IDF(Term Frequency - Inverse Document Frequency), but instead of doing it individually I performed it to the sum of all texts. I used TruncatedSVD (https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html) algorithm from sci-kit and then I tried experimenting with the n-gram values, finding that the most interesting results were with 2, 3 combination, even though it takes longer to process. Lastly, I used the notebook from class on Markov Chains to create a new generative text based on all the religious texts.

Results

My results are separated into three categories:

- **The first category would be the results from web scraping. Since some of the text weren't readily available as a complete downloadable text file I used beautiful soup and was able to efficiently extract data from certain web sites that had consecutive HTMLs to find the continuous text I was searching for. It proved to be more complex when the content wasn't consecutive and I had to use the web-scraper chrome plug-in to create a json file. In any case it was very interesting to learn the basics, and how it can be applied to other investigation projects.**

- **The second category, is performing word count and word clouds for each of the religious texts. This gave me insight and a very shallow but informative summary into the content of each of the religious texts. I also performed sentiment analysis on the separated texts, to get information regarding how "positive" each one of them was. Here are some of the word counts. The word clouds are saved in the project git repository if interested:**

```
King James Bible: ({'shall': 9838, 'unto': 8997, 'lord': 7964, 'god': 4472, 'said':
3999, 'thee': 3827, 'upon': 2748, 'man': 2735, 'israel': 2575, 'king': 2540,
'son': 2392, 'people': 2143, 'came': 2093, 'house': 2024, 'come': 1971 }]
```

```
Quran: ({'allah': 2918, 'unto': 1868, 'lord': 1009, 'said': 762, 'say': 706,
'thee': 635, 'day': 524, 'us': 456, 'believe': 434, 'may': 405, 'earth': 395,
'shall': 364, 'verily': 363,}]
```

```
Bhagavad_Gita: ({'one': 66, 'thee': 61, 'soul': 56, 'life': 48, 'shall': 48,
'heart': 41, 'fn': 40, 'lord': 37, 'unto': 35, 'man': 33, 'know': 31, 'yet': 30,
'work': 29, 'things': 28, 'prince': 27, 'mind': 27, 'worship': 27,}]
```

```
Tripitaka: ({'one': 2696, 'bhikkhus': 2594, 'mind': 2533, 'bhikkhu': 2201,
'bhagava': 1954, 'dhamma': 1480, 'sutta': 1436, 'four': 1252, 'said': 1229,
'venerable': 1229, 'life': 1202, 'also': 1173, 'body': 1126, 'world': 1121,
'sir': 1115, 'king': 1090}]
```

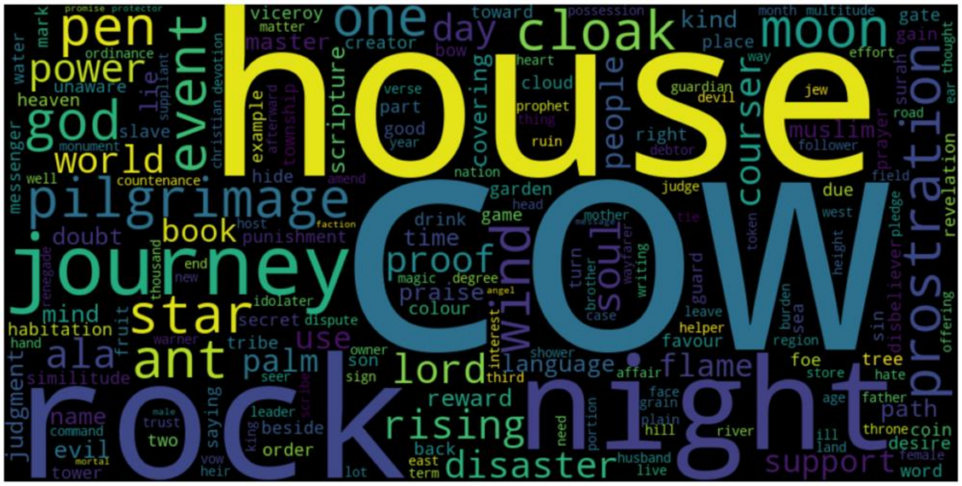
```
Tanakh: ({'adonai': 6652, 'god': 4033, 'people': 3519, 'said': 3432, 'one': 2716,
'isra’el': 2520, 'king': 2329, 'son': 2171, 'land': 1869, 'like': 1606, 'house':
1479, 'us': 1382, 'day': 1374}]
```

```
Writings of Bhallua: ({'god': 5880, 'thee': 3403, 'unto': 2924, 'one': 2151,
'lord': 2021, 'upon': 1836, 'may': 1471, 'people': 1250, 'things': 1246, 'every':
1231, 'world': 1222, 'earth': 1218, 'truth': 1142, 'day': 1114}]
```

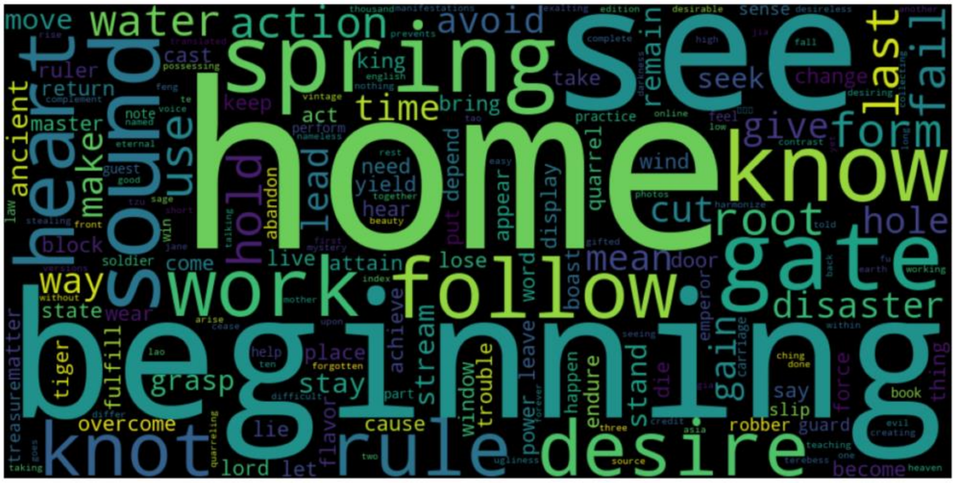
```
Tao Te Ching: {'tao': 1118, 'one': 864, 'things': 800, 'therefore': 736, 'good':
592, 'man': 592, 'without': 576, 'great': 576, 'virtue': 512, 'people': 504,
'heaven': 488, 'like': 480, 'sage': 448}]
```

```
Kojiki: ({'name': 1232, 'deity': 806, 'august': 718, 'motowori': 689, 'sect':
638, 'next': 627, 'augustness': 613, 'note': 606, 'great': 510, 'see': 492,
'heavenly': 462]
```

This It was useful for example in the Buddhist Tripitaka the word ‘Bhikkhus’ repeated a lot, I researched the meaning and it means monks, so it helped to quickly get a glimpse of the text. The word clouds in the repository are also very interesting. These are some:

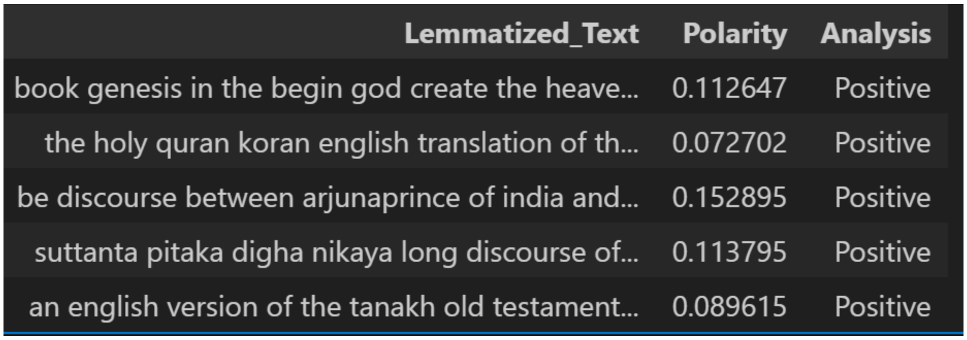


Quran

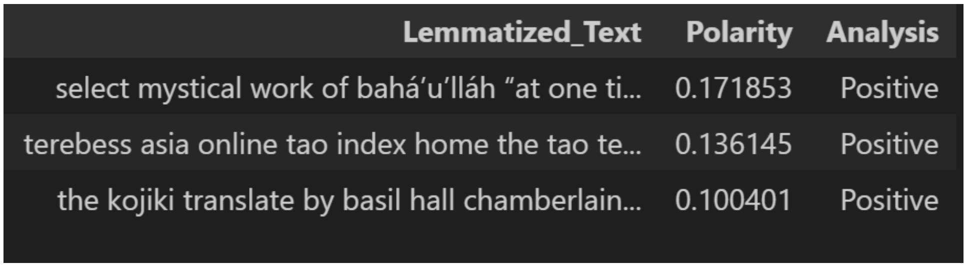


Tao Te Ching

This is the sentiment analysis for each of the religious texts.



	Lemmatized_Text	Polarity	Analysis
	book genesis in the begin god create the heave...	0.112647	Positive
	the holy quran koran english translation of th...	0.072702	Positive
	be discourse between arjunaprince of india and...	0.152895	Positive
	suttanta pitaka digha nikaya long discourse of...	0.113795	Positive
	an english version of the tanakh old testament...	0.089615	Positive



	Lemmatized_Text	Polarity	Analysis
	select mystical work of bahá'u'lláh "at one ti...	0.171853	Positive
	terebess asia online tao index home the tao te...	0.136145	Positive
	the kojiki translate by basil hall chamberlain...	0.100401	Positive

Fortunately, they are all positive.

- **The third category would be Topic Analysis, which was performed on all of the texts together, and the use of Markov chains to generate (somehow not very consequent) but beautiful text.**

These are the first 12 words for the main topics with n-grams 2-3:

Topic 0 ['lord host' 'shall eat' 'god lord' 'house lord' 'haply may' 'cry unto' 'burnt offering' 'lord world' 'one another' 'beside allah' 'believe good work' 'messenger allah']

Topic 1 ['equal heart' 'set freefrom' 'fn sanskrit' 'devilish womb' 'strive thereto' 'inward breath' 'adhiyajna lord' 'longarmed lord' 'life within' 'self friend' 'thousand yugas' 'darken soul']

Topic 2 ['shin great' 'ancient matter' 'heaven shin' 'heavenly deity' 'motowori commentary' 'great august deity' 'august mausoleum' 'dwelt palace' 'deity born' 'male deity' 'reply say' 'next deity']

Topic 3 ['king maha' 'naked ascetic' 'vi im' 'future buddha' 'enlighten one' 'sense base' 'devas men' 'tzva ot' 'adonai elohim' 'sensual pleasure' 'achieves remains' 'universal monarch']

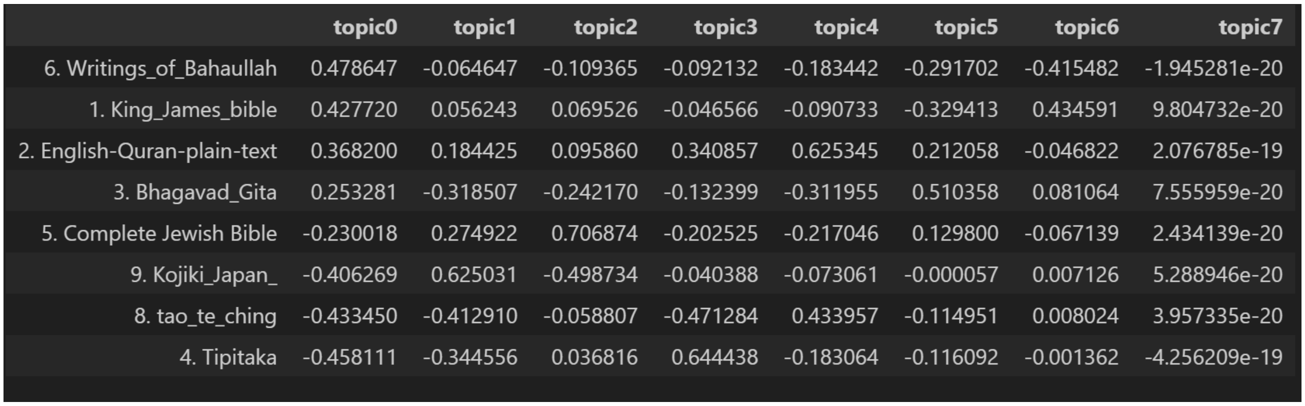
Topic 4 ['offering adonai' 'lord yeshua' 'open land' 'sin offering' 'word adonai come' 'adonai give' 'el son' 'god give' 'olive oil' 'good news' 'adonai make' 'adonai god isra']

Topic 5 ['bear witness' 'thee art' 'best beloved' 'shone forth' 'book god' 'draw nigh unto' 'love thee' 'implore thee' 'people earth' 'concourse high' 'word god' 'amongst men']

Topic 6 ['lord god shall' 'child ammon' 'take away' 'unto child' 'hearken unto' 'king assyria' 'answer say unto' 'shalt make' 'unto god' 'word lord come' 'shall go' 'israel say']

Topic 7 ['subject nature' 'material benefit honour' 'deva world' 'first jhana' 'ariya truth' 'mind object' 'exalt one' 'mind bhikkhu' 'lead cessation' 'four ariya' 'king maha' 'sense base']

And the corresponding chart:



	topic0	topic1	topic2	topic3	topic4	topic5	topic6	topic7
6. Writings_of_Bahaullah	0.478647	-0.064647	-0.109365	-0.092132	-0.183442	-0.291702	-0.415482	-1.945281e-20
1. King_James_bible	0.427720	0.056243	0.069526	-0.046566	-0.090733	-0.329413	0.434591	9.804732e-20
2. English-Quran-plain-text	0.368200	0.184425	0.095860	0.340857	0.625345	0.212058	-0.046822	2.076785e-19
3. Bhagavad_Gita	0.253281	-0.318507	-0.242170	-0.132399	-0.311955	0.510358	0.081064	7.555959e-20
5. Complete Jewish Bible	-0.230018	0.274922	0.706874	-0.202525	-0.217046	0.129800	-0.067139	2.434139e-20
9. Kojiki_Japan_	-0.406269	0.625031	-0.498734	-0.040388	-0.073061	-0.000057	0.007126	5.288946e-20
8. tao_te_ching	-0.433450	-0.412910	-0.058807	-0.471284	0.433957	-0.114951	0.008024	3.957335e-20
4. Tipitaka	-0.458111	-0.344556	0.036816	0.644438	-0.183064	-0.116092	-0.001362	-4.256209e-19

This is an example of a segment of the text using Markov chains:

"Will not take thy store with our money;

he called his sister,

she was for meat offering,

mingled with you over Jordan,

this house and the borders of Egypt to Gilgal.

Beside the bows of cloud,

covered with seven days,

and with a heap of the eighth day.

Honour upon them and they came near,

and my brethren shalt thou,

shalt not dog into a heaven offering,

made haste for the sacrifice of the children"

(I added the punctuation:)

Discussion

Web scraping was very interesting, but when the web pages became more complex it was harder to scrape them. With more time I would’ve tuned more the web scraper and learn how to scrape more complex websites with more detailed results.

First, I tried to use my dataset as a whole, and then I realized that to get more interesting conclusions it was first necessary to analyze each text individually to find relevant conclusions about the similarity of religions. I also discovered (after many attempts) that it was easier to save lemmatization results as a csv file. This way I could have the texts lemmatized and tokenized and I would not have to wait around 40 minutes each time for my program to process the data.

Among these analysis were the word count and the word clouds, that were very interesting because they did give an insight into each religions main topics. Also the sentiment analysis gave interesting information about the ‘nature' of each text.

Then doing the LSA topic weighting was also very interesting, but I think due to the way these religious texts are written it was harder to find the moral topics I was expecting to appear. To improve the results in further investigation, I would change the settings and for example remove all nouns.

A way to improve this would be performing a different type of lemmatization, using only verbs and maybe removing the nouns with the POS(Part of speech function). Because there are lot of specific main characters in these books which make it harder to find the “message” I was searching for. Also I would change the n-grams to provide better insights, once I cleaned the texts a bit further.

I didn't have enough time to apply Word RNN with pyTorch, that would be the next step. The Markov chains word generator was an interesting experiment, but due to its algorithmic nature the output is mostly sentences that doesn’t make much sense. I had hopes of creating a new religious text that was the sum of all the previous ones, but I'll have to do it in further research. Also it was among my ideas to create a chatbot that could reply a religious answer to any question, but that will also have to wait.

Conclusion

In conclusion it was a very interesting insight of NLP and the various tools it provides for analyzing data. I tried my best to utilize most of the notebooks given to us throughout the term in order to understand the entire process better. Even though I found interesting results, I would need more time to fine-tune all the tools.

I have now a glimpse of an immensely powerful investigation tool, that provides a lot of information about a certain subject, and that can be easily used in connection with other artistic projects. Even though some of the results weren’t what I expected, I am very happy with the process and the new horizons these techniques bring to my research.

Ethical considerations

The intended use of my project is to create consciousness about how all religions teach similar moral principles, and is us humans who deviate from the origin, and use them instead as a tool for division. I do not intend to put one religion over the other or underestimate their importance to our species. My dataset is taken mostly from public sources, as the nature of these religious texts is meant for the public. It was important to understand the use of the file ROBOTS.txt in html websites, which are a part of conscious scraping and respecting intellectual property.

LLM disclaimer

I have used LLMs in my project to assist me, but it also provided interesting results during the process. As the book “You look like a thing and I love you” by Janelle Shane says: ‘The danger of AI is not that it's too smart, but that it's not smart enough; AI has the approximate brainpower of a worm’ (Shane, 2020). At the beginning of the project, I was using AI more, for example, to clean the text, but I later discovered that the REGEX formula it provided was achieving unwanted results and cleaning too much, so I had to go back and find information from other sources (technical at CCI helped me a lot). Another example I used it for was to take all the contents of a lemmatization and tokenization and save them as a csv file, so I didn’t have to start again every time I ran the notebook. I think for these simple tasks is useful, but it can easily provide wrong answers that will cost you time in the future. This course gave me an insight into how LLM’s actually work and were coded, the moral issues they have, promoting old bias and prejudist ideas that might be in the dataset they're trained on.

Bibliography

- Corpus

The King James Version of the Bible. (1989). [online] *Project Gutenberg*. Available at: https://www.gutenberg.org/ebooks/10 [Accessed 24 Nov. 2023].

Brown, W.B. ed., (n.d.). *THE HOLY QURAN (KORAN)*. [online] Translated by M. Marmaduke Pickthal. *Internet Archive*. Available at: https://archive.org/details/EnglishQuran [Accessed 6 Dec. 2023]. Opensource.

www.gutenberg.org. (2000). *The Project Gutenberg E-text of The Bhagavad-Gita*. [online] Available at: https://www.gutenberg.org/files/2388/2388-h/2388-h.htm Produced by J. C. Byers. HTML version by Al Haines. Translator: Sir Edwin Arnold .

Stern, D.H. (2001). *Complete Jewish Bible*. Messianic Jewish Publisher.

www.bahai.org. (n.d.). *Writings of Bahá’u’lláh | Bahá’í Reference Library*. [online] Available at: https://www.bahai.org/library/authoritative-texts/bahaullah/.

Myanmar Government (n.d.). *Tipitaka English Translation from Myanmar Government (DPPS)*. [online] *Internet Archive*. Available at: https://archive.org/details/TipitakaEnglishTranslationFromMyanmar/ [Accessed 3 Dec. 2023].

terebess.hu. (n.d.). *Tao Te Ching, English by Gia-fu Feng and Jane English, Terebess Asia Online (TAO)*. [online] Available at: https://terebess.hu/english/tao/gia.html.

Sacred-texts.com. (2019). *The Kojiki Index*. [online] Available at: https://sacred-texts.com/shi/kj/index.htm translated by Basil Hall Chamberlain.

Other Bibliography

Gupta, S. (2018). *Sentiment Analysis: Concept, Analysis and Applications*. [online] Towards Data Science. Available at: https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17.

McDonald, D. (2014). A Text Mining Analysis of Religious Texts. *The Journal of Business Inquiry*, 13(1), pp.27–47.

Patel, V. (2022). *Markov Chain Explained | Built In*. [online] builtin.com. Available at: https://builtin.com/machine-learning/markov-chain.

Shane, J. (2020). *You Look Like A Thing And I Love You.* S.L.: Wildfire.

Zemlyanker, D. (2022). *Using natural language processing to analyze religious text – NU Sci Magazine*. [online] nuscimagazine.com. Available at: https://nuscimagazine.com/using-natural-language-processing-to-analyze-religious-texts/.