A Quantitative Analysis of Chinese Philosophical Texts Using Topic Modelling

Kyle Chen

Chinese 370: Hacking Chinese Studies

May 16, 2023

**Abstract**

The importance of the works of Confucius, Mencius, Laozi, and Zhuangzi, or at least their ideas that were later scribed on to paper by others, are the roots of modern Chinese culture and philosophy. Previously, I had written a paper titled “A Comparison Between the *Mencius* and Contemporary Philosophies” as a qualitative comparison between *The Analects*, *Dao De Jing*, *Mengzi*, and *Zhuangzi*. As such, I had concluded that, in order to best parse the texts in their entirety, having a quantitative comparison via a computer algorithm would aid in generating a more accurate comparison of the works mentioned above, or at least help in supporting the findings in that paper.

The new research that I conducted in this report developed a program that used Python, and specifically the Natural Language Toolkit and Gensim Topic Modelling libraries, as well as the regular expression library to better parse the input texts for more specific output in the form of a graph.

Ultimately, the algorithm, after parsing multiple versions of the corpus, decided on the conclusion that the texts all have some form of individualism as a main topic inherent to them, with less emphasis placed on the societal elements such as community and applied higher weights to the Confucian texts over most of the valid data, with only a few passes placing higher emphasis on the Daoist texts. This paper found that parsing the context of the topics within a specific text leads to much less specific data, or rather much broader information return, from parsing the texts without restricting the input to the topic modelling algorithm. With the selected input, we see much more specific output from the topic modelling algorithm, showing better adherence to previous research done by manually parsing all of the texts. However, both unrestricted (“lax passing”) and restricted (“strict passing”) approaches show that the text topics overall treat individual traits over social traits.

**Introduction**

Of the many philosophical schools of China that developed during the Warring States period prior to the formation of the Qin Dynasty, two of the largest were Confucianism and Daoism, and, of the two, Confucianism was revived as Neo-Confucianism by the Song Dynasty, with one of its core corpora being the *Mencius*, written by Mengzi.

Mengzi’s own philosophical work iterates upon the earlier teachings of Confucius, reiterating the importance of family (“xiao”, filiality), tradition, ritual, and duty. Yet, Mengzi also places importance on the idea that all are born with good intentions and are molded instead by circumstance.

One of the main reasons as for Mengzi’s importance to Confucian thought is the revival of Confucian teachings by Neo-Confucianists in the Song Dynasty, who felt the need to add a spiritual element to the philosophy to make it more appealing to a population that is primarily Buddhist. According to David Hinton, this attempt at redefining Confucianism from governing ideology to a spiritual ideology is why the *Mencius*, alongside *The Great Learning* and *The Doctrine of the Mean* are associated with more contemporary Confucianism.[[1]](#footnote-2)

As the ideas of the *Mencius* has now been ingrained into modern East Asian culture, it perhaps may be better for us to develop some level of understanding of the translated text, even though there may exist cultural cues or expressions that may be lost upon translation. The purpose of this report is to primarily analyze the *Mencius* in its translated form and, secondary to that, compare *Mencius* with the works of Confucius, Laozi, and Zhuangzi in order to provide reasons for why it has taken this new cultural importance developed by Neo-Confucian thinkers.

The Zhuangzi, on the other hand, is a collection of 33 chapters, of which 7 are likely have been written by Zhuangzi himself – called the Inner Chapters – and the other 26 chapters being dubbed the “Outer Chapters” (9-22) or the “Miscellaneous Chapters” (23-33).[[2]](#footnote-3) This, alongside the many different styles of *The Zhuangzi*, to the point that each separate chapter within the Inner Chapters of *The Zhuangzi* can be considered its own short story,[[3]](#footnote-4) seeing as how Zhuangzi himself has a fondness for metaphors over direct wisdom as is the case with Mengzi. Tangentially, *The Zhuangzi* pulls a large amount of influence from both Confucian and Daoist texts, as seen with Zhuangzi’s references to Confucius himself. With all these literary styles and with such a large corpus, if we were to split each text into its own chapters, it is certainly difficult to glean the meaning from each separate passage. As such, using a machine learning approach, such as topic modelling, could allow us to better parse the texts.

**Preliminary Findings**

In my previous research on the subject, the *Mencius* shares much with its philosophical predecessor, *The Analects*, where both discuss in depth the importance of tradition, Ritual, and Duty. Mencius himself not only references the *Book of Songs* during conversation, but he also references *The Analects* in order to provide greater context for his own reasoning. One of the key differences between the ideas of Mencius and Confucius, however, is that Mencius places much more emphasis on the will of the people.

Laozi and Zhuangzi, on the other hand, talk about the pure flexibility of human nature, and how one should strive to go back towards a natural or original state, like that of water.[[4]](#footnote-5)

One of the key differences between the ideas of Mencius and Confucius is that Mencius places much more emphasis on the will of the people. Confucius, on the other hand, primarily discusses how people should act within society.

Thus, we have our 6 topics: Tradition, Ritual, Duty, Individuality, Flexibility, Society. Similarly, after reading the papers, or at least the introduction where the translator has a list of important themes with their Pinyin spelling and English translation[[5]](#footnote-6),[[6]](#footnote-7),[[7]](#footnote-8),[[8]](#footnote-9), and using internet resources such as ChatGPT, we can see key words within each of the texts. Thus, in order to commence a stricter parsing of the data, we can record these important words as follows:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| "dao": "way" | "ziran": "naturalness", "spontaneity" | "ren": "benevolence" | "tian": "heaven", | "xing": "nature", "disposition" | "reality" |
| "de": "power", "virtue" | "pu": "simplicity" | "xin": "heart" | "zhong": "faithfulness", "loyalty" | "qi": "life" | "zhi xing", "zhixing": "conduct" |
| "wuwei", "wu wei": “non-striving” | "wu": "emptiness" | "junzi": "gentleman", "noble" | "jun": "ruler", "king" | "youyou": "wandering" | "zhongyong": "doctrine of the mean", which includes "harmony", "balance", "moderation" |
| "yin", "yang": “harmony” | "sage" | "li": "duty" | "xiao": "filial" | "dream" |  |

Note that this is a non-exhaustive list of important words, and more can be added later.

**Methods**

In order to use the topic modelling algorithm, we must first obtain our corpus for use in the algorithm. Thus, we obtained four PDFs of the texts previously mentioned, and subsequently the UTF-8 formatted text files of a differing translation, for use. We also have four Chinese versions of the texts for later interpretation. Forking the topic modelling Python3 script “genism\_1.py” for use in this project, we now have a base program to develop our topic modelling script off of.

We can then save the files in a directory called “texts” for interpretation through Python; either directly or with the PyPDF PDF reader, as Python cannot natively read PDF files. After cleaning the texts, since reading PDFs with PyPDF leads to the string saving formatting elements, with a regular expression, we can then perform topic modelling over each set of texts, since we don’t want disparities between versions to influence output.

As such, we have 3 sets of texts: a translated set of the texts in PDF form, a translated set of the texts in text file form, and a set of the texts in Chinese in text file form. In order to best use the topic modelling algorithm, we can split the translated text files by chapter, giving us 148 separate documents. We can also develop a stricter parsing of the texts by creating a list of most important words and their translations and using a regular expression to obtain a listing of all instances of aforementioned important words and apply the topic modelling to this newly parsed version as well. Thus, we can apply the topic modelling method to 7 different sets of texts: 3 of which are the unedited texts, 1 being the chapter-divided version of the texts, and the remaining 3 are the parsed versions of the English texts.

**Initial Setbacks and Issues**

Outside of the aforementioned PDF parsing issue, the machine learning algorithm would occasionally output the highest weighted words with similarly large weights, henceforth called “turboboosting” due to the algorithm seemingly boosting the weights of words without properly correcting the overall set of weights such that the sum of the weights equal one, with the likely reason being that there wasn’t enough in the corpus for the machine learning algorithm to iterate over. As such, development time was shifted to see if a regular expression could parse all the chapters of the documents at runtime, therefore allowing any and all document formatting to pass through. However, the issue was that the UTF-8 text files were of differing formats, and thus the regular expression would fail due to issues regarding expression size and flexibility. Thus, the fix was ultimately to split the works along chapters manually, leading to an overall corpus of 148 documents.

Another issue overall is an issue of nondeterminism with the output data; that is, if we were to run the Python script twice, we would get two different sets of outputs. As such, increasing the size of the parsed corpus is the best way to get a valid set of outputs from the model algorithm.

**Results**

Based on the above research methods, we have obtained the following results:

1. From several of our runs of the script, we can see that both the lax and strict parsing methods on the standard corpuses reveal important overlap of concepts in the four documents we are trying to interpret. For example, in Figure 1, we can see that, of the valid information in the figure from the PDF documents, “wuwei” (“non-striving”), “ziran” (nature), and “honor” are highly weighted, implying that each of the texts cover these concepts in depth. Similarly, in the text file corpus, we see that “dao/tao” (way), “wu” (emptiness), and “yang” – although somehow missing “yin” – appear enough to be weighted higher in the algorithm. However, after parsing and applying the algorithm to the split versions of the corpus elements, our output shows very broad, unimportant information, as can be seen in Figure 10, with higher weights applied to words such as “they”, “their”, and “when”.
2. The Chinese text parsing through the algorithm gives us poor results due, as can be seen in Figure 7 where it places emphasis on the names of philosophers (“孔子日” – “Confucius/the master said”), which may imply that the model is placing emphasis more on the origin of the quote rather than the content.
3. The strictly parsed text as can be seen in Figures 13-21, shows us the importance of different concepts based on the differing formats and translations of the corpus, if we consider the weight of the words. In Figure 13-15, we see emphasis being placed on “junzi” (gentleman), “disposition”, and “filial”, with *The Analects* having the highest weight. In Figures 16-18, the algorithm placed emphasis on “heart”, “filial”, and “noble” and placed a high weight to the other Confucian text in the corpus, the *Mencius*. The final set of strictly parsed text, as seen in Figures 19-21, had a large weight placed on “li” (duty), and smaller weights placed on “way” and “ren” (benevolence), in which the highest weights are on chapters in *The Analects* and, oddly enough, the *Zhuangzi*.

**Conclusion**

Initially, I had issues with translation of those same phrases into English, where more inconsequential and simple words such as “one” or “his” are boosted far above more core elements such as “faith”, “duty”, or “dao”. As such, a separate parsing of the texts, perhaps manually, for each text’s core elements (i.e. if “dao” appears in a chapter of a text 5 times, then we may decide to reduce the elements of the text down to just 5 “daos”) may just provide superior, more usable results. With the results above, certainly the script outputs usable, if not more interesting, data for interpretation.

Despite my efforts in producing better data with the topic modelling algorithm, certain issues were unable to be resolved, such as “turboboosting” and poor information return, instead requiring multiple runs of the script with sometimes different filter values. Of the usable data, obtained by editing the string filter within the script and running the script multiple times, I was able to obtain useful enough data from the texts.

Of this useful data, we can conclude that “Dao/Tao/the Way” is important to all of the texts within the corpus, but with different definitions, as there exists a large divide between the Confucian and Daoist definitions of “dao”. With our lax and strict passes on the corpus, we see that the lax passes emphasizes more spiritual words than the strict passes, with the strict passes in turn emphasizing more tangible words such as “duty”. Yet, both passes still place weights on the personal topics within the texts, as “duty”, “filial”, “honor”, “wu”, and “ziran” are personal elements to be worked upon in the texts, matching up at least in part with my previous research, where the work of Mengzi, Laozi, and Zhuangzi all place emphasis on the individual.

Subsequently, to best proceed with the findings of this paper, it would be best to use SpaCy to interpret Chinese multicharacter words or certain phrases in Chinese in order to obtain better data with regards to intent of the authors and the frequency of their use in other works. Thus, the bag-of-words approach used in the basic Topic Modelling algorithm lacks enough complexity to parse the much more metaphorical language of the texts. Alternatively, having a supporting model to parse the more metaphorical meanings within the texts is necessary before applying the topic modelling algorithm. Of the data obtained and their graphs, certain filters have to be applied to them, else they will return much broader results which include less important information, such as “I”, “saw”, “year”, et cetera (Figure 22). As such, in order to obtain the best possible information from the algorithm, we must edit the filter on each word for each individual set of corpuses.

**Acknowledgements**

I would like to thank Professor Vierthaler for providing the code base for which the script used in this paper was built upon.

**Appendix**

A picture containing screenshot, rectangle, electric blue, text

Description automatically generated

Figure : Highest Weighted Words, PDF corpus (Robert Eno translation), obtained with filter: no\_below=int(len(refined\_texts) \* .25) + 1, no\_above=.75.

A picture containing screenshot, text, line, rectangle

Description automatically generated

Figure : Topic Weights, PDF corpus.

A picture containing rectangle, screenshot, square, line

Description automatically generated

Figure : Corpus Element Weights, PDF Corpus, weighted more towards the Dao De Jing (Topic 2) and The Analects (Topic 1). Note that Topic 1: The Analects, Topic 2: Dao De Jing, Topic 3: The Mencius, and Topic 4: The Zhuangzi.

A picture containing screenshot, text, plot, line

Description automatically generated

Figure : Highest Weighted Words, Text file corpus (James Legge translation), obtained with filter no\_below=int(len(refined\_texts) \* .05) + 2, no\_above=.75.

A picture containing screenshot, text, rectangle, line

Description automatically generated

Figure : Topic Weights, Text file corpus.

A picture containing screenshot, rectangle, text, line

Description automatically generated

Figure : Corpus Weights, Text file corpus.

A picture containing text, screenshot, rectangle, font

Description automatically generated

Figure : Highest Weighted Words, Chinese language corpus.

A picture containing text, screenshot, rectangle, line

Description automatically generated

Figure : Topic Weights, Chinese corpus.

A picture containing rectangle, screenshot, square, line

Description automatically generated

Figure : Corpus Weights, Chinese corpus.

A picture containing screenshot, electric blue, plot, text

Description automatically generated

Figure : Highest Weighted Words, English Chapter-Split corpus (James Legge translation), filter no\_below=int(len(refined\_texts) \* .05) + 2, no\_above=.75.

A picture containing screenshot, rectangle, text, square

Description automatically generated

Figure : Topic Weights, English Chapter-Split corpus.

A picture containing screenshot, line, text, plot

Description automatically generated

Figure : Corpus Weights, English Chapter-Split corpus.

A picture containing screenshot, diagram, electric blue, text

Description automatically generated

Figure : Highest Weighted Words, PDF corpus (Robert Eno translation, strict parse), filter no\_below=int(len(refined\_texts) \* .25) + 1, no\_above=.75.

A picture containing screenshot, text, rectangle, line

Description automatically generated

Figure : Topic Weights, PDF corpus (strict parse).

A blue squares on a white background

Description automatically generated with low confidence

Figure : Corpus Weights, PDF corpus (strict parse)

A picture containing screenshot, text, diagram, line

Description automatically generated

Figure : Highest Weighted Words, Text file corpus (James Legge translation, strict parse), filter no\_below=int(len(refined\_texts) \* .25) + 1, no\_above=.75.

A picture containing screenshot, text, rectangle, square

Description automatically generated

Figure : Topic Weights, Text file corpus (strict parse).

A picture containing rectangle, screenshot, square, line

Description automatically generated

Figure : Corpus Weights, Text file corpus (strict parse).

A picture containing screenshot, text, rectangle, font

Description automatically generated

Figure : Highest Weighted Words, English Chapter-Split corpus (James Legge translation, strict parse), filter no\_below=int(len(refined\_texts) \* .25) + 1, no\_above=.75.

A picture containing screenshot, rectangle, text, square

Description automatically generated

Figure : Topic Weights, English Chapter-Split corpus (strict parse).

A picture containing text, screenshot, line, font

Description automatically generated

Figure : Corpus Weights, English Chapter-Split corpus (strict parse).

A picture containing screenshot, electric blue, plot, line

Description automatically generated

Figure : Highest Weighted Words, PDF corpus (Robert Eno translation), obtained with filter: no\_below=int(len(refined\_texts) \* .05), no\_above=.75.

Bibliography

Confucius. *The Analects*. Translated by James Legge. Scanned by the Chinese Text Project. Accessed May 3, 2023.

Confucius. *The Analects*. Translated by Robert Eno. Indiana: University of Indiana, 2015. Accessed April 29, 2023.

Hinton, David. Analects. Berkeley: Counterpoint, 2014.

Hinton, David. The Four Chinese Classics : Tao Te Ching, Analects, Chuang Tzu, Mencius. Berkeley: Counterpoint, 2013. Accessed October 27, 2020. ProQuest Ebook Central.

Gassmann, Robert H.. Menzius : Eine Kritische Rekonstruktion Mit Kommentierter Neuübersetzung. Berlin/Boston: De Gruyter, Inc., 2016. Accessed October 30, 2020. ProQuest Ebook Central.

Laozi. *Dao De Jing*. Translated by James Legge. Project Gutenberg, 1995.

Laozi. *Dao De Jing*. Translated by Robert Eno. Indiana: University of Indiana, 2010. Accessed April 29, 2023.

Mengzi. *Mencius*. Translated by James Legge. Scanned by the Chinese Text Project. Accessed April 29, 2023.

Mengzi. *Mencius*. Translated by Robert Eno. Indiana: University of Indiana, 2019. Accessed April 29, 2023.

Zhuangzi, et al. *Zhuangzi*. Translated by James Legge. Scanned by the Chinese Text Project. Accessed May 3, 2023.

Zhuangzi, et al. *Zhuangzi*. Translated by Robert Eno. Indiana: University of Indiana, 2010. Accessed April 29, 2023.

1. David Hinton, *The Four Chinese Classics : Tao Te Ching, Analects, Chuang Tzu, Mencius* (Berkeley: Counterpoint, 2013), 394. [↑](#footnote-ref-2)
2. Zhuangzi. Zhuangzi. Translated by Robert Eno. Indiana: University of Indiana, 2010, iii. [↑](#footnote-ref-3)
3. Zhuangzi. Zhuangzi. Translated by Robert Eno. Indiana: University of Indiana, 2010, 1. [↑](#footnote-ref-4)
4. Laozi, *Dao De Jing*, trans. Bruce R. Linnell (Project Gutenberg, 2015), 16. [↑](#footnote-ref-5)
5. Conficius, *The Analects*, trans. Robert Eno. Indiana: University of Indiana, 2015, vi-vii. [↑](#footnote-ref-6)
6. Laozi, *Dao De Jing*, trans. Robert Eno. Indiana: University of Indiana, 2010, 6-7. [↑](#footnote-ref-7)
7. Zhuangzi, *Zhuangzi*, trans. Robert Eno. Indiana: University of Indiana, 2010, 2-5. [↑](#footnote-ref-8)
8. Mengzi, *Mencius*, trans. Robert Eno. Indiana: University of Indiana, 2019, 10-11. [↑](#footnote-ref-9)