Journal of Asian Studies

Modeling the contested relationship between Analects, Mencius, and Xunzi: Preliminary evidence from a Machine-Learning Approach

--Manuscript Draft--

Manuscript Number:	JAS-D-16-00009R1
Full Title:	Modeling the contested relationship between Analects, Mencius, and Xunzi: Preliminary evidence from a Machine-Learning Approach
Article Type:	Article
Corresponding Author:	Ryan Nichols California State University, Fullerton UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	California State University, Fullerton
Corresponding Author's Secondary Institution:	
First Author:	Ryan Nichols
First Author Secondary Information:	
Order of Authors:	Ryan Nichols
	Edward Slingerland
	Kristoffer Nielbo
	Uffe Bergeton
	Carson Logan
	Scott Kleinman
Order of Authors Secondary Information:	
Abstract:	We present preliminary findings from a multi-year, multi-disciplinary text analysis project using an ancient and medieval Chinese corpus of over 5 million characters in works that date from the earliest received texts to the Song Dynasty. §1 describes 'distant reading' methods in the Humanities and our corpus. §2 introduces topic modeling procedures, answers questions about our data, and discusses complementary relationships between machine-learning and human expertise. §3 explains topics represented in Analects, Mencius and Xunzi that set each of those texts apart from the other two, while §4 explains topics that intersect all three texts. Our results confirm many scholarly opinions derived from close reading methods, suggest a reappraisal of Xunzi's shared semantic content with Analects, and yield several actionable research questions for traditional scholarship. The aim of this exploratory paper is to initiate a new conversation about implications of machine learning for the study of Asian texts.
Opposed Reviewers:	
Response to Reviewers:	Please see attached file for our detailed responses to referee comments. /rn

Modeling the contested relationship between *Analects*, *Mencius*, and *Xunzi*: Preliminary evidence from a Machine-Learning Approach

Traditionally the *Mencius* has been considered as the philosophical heir to the moral philosophy and theory of human nature presented in the *Analects*. However, recent scholars argue that the *Xunzi* is closer to the *Analects* than the *Mencius*. This paper attempts to contribute to the debate by introducing a machine learning approach to supplement traditional modes of inquiry. We make use of a technique known as *topic modeling* to provide a new perspective in ongoing conversations about Confucianism and the relationships between some of the most important source texts in Early Confucian thought.

Topic modeling has already become a complementary source of knowledge and information for scholars across the humanities who are accustomed to using close-reading methods for the extraction of meaning from texts. Topic models identify groups of words (called *topics*) that are statistically likely to co-occur in a text or corpus. Insofar as traditional studies prompt the scholar to bring ideas, themes and assumptions *to* texts, topic modeling reverses this process. In this way topic modeling supplements, confirms or, in some cases, gently challenges conclusions from close-reading traditions. We aspire to combine knowledge of the contents of topics, contents of texts, and expertise in classical Chinese language, culture and thought, and so bring a pioneering navigational tool to the exploration of historically important Chinese documents of deep and wide interest to a readership across Asian Studies, Philosophy, Literature, Religion, and more.

In this paper we explain what topic modeling is, introduce our corpus of ancient and medieval Chinese texts, and discuss preliminary results of our topic modeling research as applied

to questions about the relationship between *Mencius* and *Xunzi* to *Analects*. The *Analects* contains sayings and ideas attributed to Confucius (551–479 BCE) and his followers. Mencius (early 4th cent, BCE-late 4th cent. BCE) and Xunzi (c. 310 – c. 235 BCE / c. 314 – c. 217 BCE) both explicitly stated that they followed the teachings of Confucius. However, since the philosophies of *Mencius* and *Xunzi* differ considerably, the question of who is most in tune with the ideas of Confucius of *Analects* is an open one. As a team composed of experts in pre-Qin Chinese religion and philosophy, Warring States Chinese language and linguistics, and humanities computing, we have and will continue to use traditional close-reading techniques for understanding Chinese thought. We treat the results that follow as the first machine learning steps in a wider interdisciplinary effort to gain deep knowledge of the meaning of some of the world's most influential texts. Our primary goal is to present information capable of starting a new, exciting thread in a millennia-long conversation about the interpretation of a few of the world's most influential texts.

1 Distant Reading & Chinese Texts

Understanding the literary, intellectual, and cultural history of ancient and medieval Chinese literature presents the traditional scholar with imposing challenges. The authenticity, authorship and dates of composition of texts are often either unknown or widely contested (Loewe 1993, Chennault, et al., 2015). Furthermore, since many early Chinese texts are compilations of texts composed by different authors at different times put together by later editors, just what qualifies as a single text is debatable (Boltz 2005). Except for recently excavated manuscripts, most extant early Chinese texts are the products of scribal copying, censorship, redaction, loss of books and

other forms of textual corruption. These documents rarely received study independent of traditional commentary. On top of these concerns the sheer size and complexity of the ancient and medieval Chinese corpus prevents any one individual from mastering all its texts.

Distant reading methods, increasingly popular across the humanities, provide assistance in overcoming some of these challenges and supplementing traditional expertise. Coined by Franco Moretti (2000), this portmanteau refers to the use of computational techniques to identify patterns in corpora. Distant reading differs from innovative studies of texts using human coders because distant reading is primarily computational rather than experimental. Despite this, we see our computational approach as complementary not only to traditional scholarship but to quantitative studies of pre-Qin texts using human coders like Slingerland and Chudek (2011) about the heart-mind, and Clark and Winslett (2011) about supernatural agencies.

The computational methods in distant reading use computer algorithms to identify patterns in literary corpora independent of prior assumptions of experts. The use of computation means the process is largely but not totally free from biases (Goldstone and Underwood 2014, 364). Often these patterns are hidden from the view of scholars not because of scholarly bias but because these patterns only appear at scale, or involve word usage that does not typically catch the human eye. “Computation” refers to the mathematical mapping of variables from one set to variables in another. If we are given a list of whole numbers, we might map the variable *larger than* onto *integers* in order to compute the largest number in the list. An algorithm is a bit-by-bit recipe for implementing a computation. Familiar processes like tying one’s shoes and baking a cake are familiar processes that can be described algorithmically. In these cases, the user of the algorithm supervises its step-by-step iteration and has a specific outcome in mind. Sometimes

algorithms are exploratory and used without this sort of supervision. For example, we might have no prior idea about how many prime numbers there are between 2,576 and 6,509,322. Despite not knowing the outcome in advance, we can still write an algorithm to give us this information.

Topic modeling is one method of unsupervised algorithmic exploration. Topic modeling results have supplemented and invigorated a number of other humanities research areas, including history, philosophy, journalism and prose and poetry studies. In these and other areas, topic modeling results have moved far beyond the digital humanities ghetto to contribute to answers to questions of considerable intellectual value. Literary and historical studies have benefitted the most from topic modeling, as is apparent in the work and influence of M. Jockers and his remarkable study of 19th century novels in the United Kingdom, Ireland and America (Jockers 2013). He explores major themes and often contrast emergent themes in one country's novels, like landlord-tenant relations, with those of another, like race. In history, R. Nelson topic modeled the archives of the *Richmond Daily Dispatch* newspaper from November 1860 to December 1865 during the American Civil War. Nelson tracked changes in relationships between words about the Confederate military draft, fatalities, and patriotism by using the algorithm to compute a mapping of words to words and words to dates. Combining knowledge of dates of movements of the Union army, Nelson found that ads for fugitive slaves spiked on the two occasions when the Union army came closest to Richmond. These results work in harmony with research by historians by providing correlational evidence for a theory: a minority of civil war historians have argued for greater appreciation of the role of the Union army in the destabilization of slavery in the Confederate south, independent of the Emancipation Proclamation (Nelson 2015). Asian Studies has not yet caught up with other humanities areas in

the use of topic modeling, though this may be changing (Chen, Borovsky, Kawano, and Chen 2014; Hou and Frank 2015).

We apply a topic modeling algorithm to a corpus of 5.74 million characters across 96 ancient and medieval Chinese works, including many of the most important texts in the tradition. The texts in this corpus were processed with generous permission of Dr. Donald Sturgeon from the Chinese Text Project (http://ctext.org/). The corpus spans several eras of historical Chinese literature. It includes the pre-Warring States *Book of Poetry* (*Shījīng* 詩經), the Warring States *Dào Dé Jīng* (道德經), the short treatise on philosophy of language *Gōngsūnlóngzi* (公孫龍子) and the lengthy history text *Hàn Shū* (漢書), Han medical texts like *Huángdì Nèijīng* (黃帝內經) and pre-Qin encyclopedic texts like *Lǚ Shì Chūnqiū* (呂氏春秋). (See Appendix 1 for the complete list of texts and Figure 1 for era classifications of the corpus.)

Figure 1 Corpus Composition by Era

Era	Dates	Character Count	Percent of Corpus
Pre-Warring States	Before 480 BCE	30,447	0.53
Warring States	479-222 BCE	1,424,080	24.79
Han	221 BCE-220 CE	3,501,256	60.9
Post-Han to Song	221 CE-1044 CE	786,546	13.6
Totals		5,742,329	1.00

2 Topic Modeling

Comprehensive and friendly introductions to topic modeling for humanists have been written. We do not aim to duplicate those resources. (See Blei 2012b, Weingart 2012, Underwood 2012a and Mohr and Bogdanov 2013.) Yet we see broad benefits in directly providing researchers across subfields of Asian Studies with hands-on knowledge of the topic modeling process, since many scholars of texts of any kind will soon benefit from--and some will need to acquire--the ability to interpret topic models in their research area.

Topic modeling was developed for search and retrieval in large collections of text-heavy data, but topic models efficiently sum, visualize and explore the semantics of any kind of text corpus. Words are assigned to topics based on their tendency to co-occur in texts with other terms found in the topic.[1] For topic modeling we use an algorithm that maps, i.e. computes probabilistic values for, the relations amongst all the terms in the corpus to all the other terms in

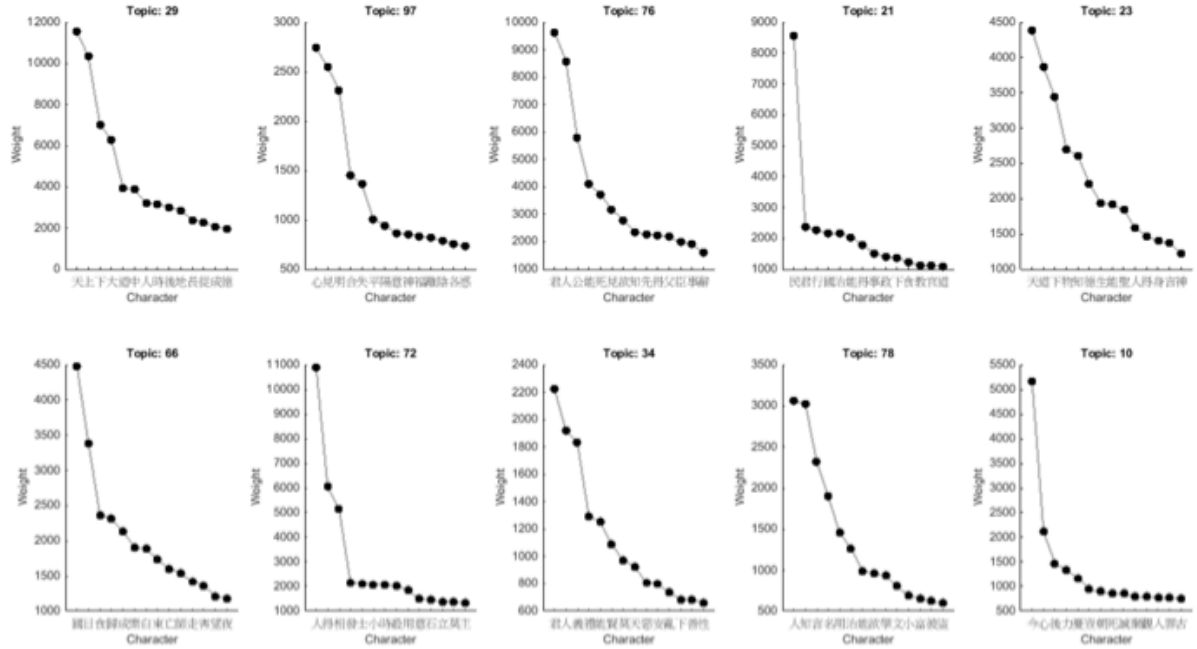
[1] Formally, a topic is a mapping or a probability distribution over terms. 'Terms' here refers to countable linguistic forms such as words or Chinese characters. A number of algorithms and tools can be used to calculate such distributions. We use a sampling-based algorithm for latent Dirichlet allocation known as 'LDA'. LDA is a generative probabilistic model that extracts a set of latent variables (i.e., topics) in large collections of documents. This is implemented in a software environment called MALLET, an acronym for 'MAchine Learning for LanguagE Toolkit.' MALLET has proved effective in modeling humanities data (McCallum 2002). The LDA topic model employed in this study uses a Gibbs sampler, which is a Markov chain Monte Carlo algorithm for obtaining observations from the multivariate Dirichlet distribution. For the underlying mathematics, see Blei 2012a.

the corpus. Through this process, the model extracts topics in large collections of documents. The model is *probabilistic* because the topics consist of words that have a high probability of occurring together in documents (Blei, Ng and Jordan 2003). The model is *generative* because topics are formulated from latent relationships amongst terms in documents. Unlike an algorithm for tying one's shoes, but like an algorithm for discovering a set of unknown prime numbers, our model works without supervision. This means that the algorithm discovers the topics without its being fed prior knowledge about genre or date or any other information about the texts. The upshot is that before seeing the results, we do not know what the topics will be or which topic will have the biggest representation in the corpus.

Topic models produce several different types of data, including word weights, corpus weights and text weights. In practice, many digital humanities papers using topic modeling neglect much of these data in preference for focusing on the resulting topics. Since we attempt to exploit the full range of these data to address our research question about the relation between *Analects*, *Mencius* and *Xunzi*, we now briefly explain types of data. The number of times the topic modeling algorithm has assigned a given term to the topic determines its *word weight*. 'Weight' in this context refers to the relative size of the contribution that a word makes in a topic. The centrality of the word (or character or term or glyph) to the topic can be determined by rank-ordering the word weights. Topics are customarily split into short lists of the top-ranked words, the topic's *keywords.* Keywords serve as a metonym for the long list of words in the entire topic. The plot for Topic 29, in Figure 2, shows the character tiān (heaven or God 天) as having the largest weight of any keyword in that topic. In this context, the large word weight of tiān results from its nearly 12,000 occurrences in Topic 29.

Next is *corpus weight*. Figures 3 and 4 illustrate a topic's weight in the corpus, or *corpus weight*. A topic's corpus weight is the ratio of the sum of words in a topic over the total number of words in the corpus. Corpus weights are not standardized (ours sum to 14.9) because they are based on the total occurrence of words within each topic. Instead of representing the weights of individual characters, Figure 3 depicts keywords (in the order of their word weight within the topic) along with the corpus weight of the topic. Figure 3 represents our topic model's findings as to the ten most weighty themes in ancient and medieval Chinese writing. Figure 4 visualizes the corpus weights of all 100 topics in our model.

Figure 2 Keyword loading in Highest Loading Ten Topics in Our Corpus

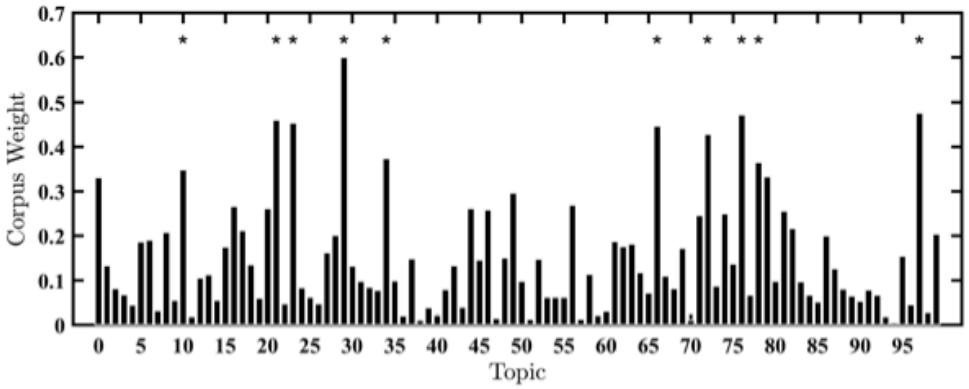


Individual plots in this figure represent the distribution of highest loading keywords within the target topic. Characters along the horizontal axis represent central characters in the target topic, with the most central character nearest the vertical axis. The numbers along the vertical axis represent the number of occurrences of each character. Typically keyword weights approximate a discrete power law distribution, with the weights being inversely proportionate to the keyword's rank for any given topic. This describes topic 21 because people (mín 民) has nearly four-times the word weight as Topic 21's second-ranked character, lord (jūn 君). Contrast Topic 23, which is almost linearly distributed. See Figure 3 below.

Figure 3 Highest Loading 10 Topics in Corpus

Topic #	Corpus Weight	Topic Name	Topic Keywords in Descending Order of Weight
29	0.600	Heaven, Cosmos, & the Way	天 上 下 大 道 中 人 時 後 地 長 從 成 德
97	0.475	Cognition, Perception & Cosmic Fortune	心 見 明 合 失 平 陽 意 神 福 離 陰 各 惑
76	0.471	Roles of Rulership	君 人 公 能 死 見 欲 知 先 得 父 臣 事 辭
21	0.459	Political & Social Order	民 君 行 國 治 能 得 事 政 下 食 教 官 道
23	0.452	Moral-Cosmic Attunement	天 道 下 物 知 德 生 能 聖 人 得 身 言 神
66	0.446	Ritual Sacrifice	國 日 食 歸 成 樂 白 東 亡 師 走 害 望 夜
72	0.428	Political Roles, Political Affairs	人 得 相 發 士 小 時 殺 用 意 石 立 莫 主
34	0.373	Ethical Rulership	君 人 義 禮 能 賢 莫 天 惡 安 亂 下 善 性
78	0.364	Learning & Governance	人 知 言 名 用 治 能 欲 學 文 小 富 彼 盜
10	0.348	Temporal Cognition & Planning	今 心 後 力 憂 豈 朝 死 誠 棄 觀 入 罪 古

Figure 4 Corpus Weights for Topics 0-99



Vertical bars represent corpus weights of all 100 topics in our topic model. Asterisks indicate the ten most central topics in the Chinese corpus.

[2]

The third and final type of data produced by a topic model are *text weights*. This term refers to the proportion of a text's vocabulary that is assigned to a given topic, which represents how saturated a text is by a topic. Text weights are normalized and sum to 1. In each text in the corpus, some of the 100 total topics will have greater representation than others. For example, in

[2] Footnote to Figure 4: Corpus weights are calculated from Dirichlet distributions that serve as hidden or latent variables responsible for the allocation of words to topics. (See Blei 2102b, 79-81; Blei 2012a, footnote 4.) Since MALLET outputs the Dirichlet parameter, which is "roughly proportional to the overall portion of the collection assigned to a given topic" (McCallum 2002), we use this number as a measure of corpus weight.

Xunzi area experts would expect that topics having to do with ritual matters will have bigger representation, and so larger text weights, as compared to *Mencius.*

Let's illustrate text weight, word weight and corpus weight in one example. Consider the topic that loads heaviest into *Xunzi*, Topic 34, which we call "Ethical Rulership." First, Topic 34 has a text weight of 0.256 in *Xunzi*. In contrast, its text weight in *Analects* is only 0.043 and in *Mencius* 0.023. This alone represents a discovery in terms of our research question since the distribution of Topic 34 into Xunzi is 6x greater than its distribution in *Analects* and 11x greater than in *Mencius*. This warrants a practical inference for scholars of ancient Chinese documents, namely, Topic 34 sets *Xunzi* apart from the other two texts.

To understand the significance of this discovery, we turn to look at the characters in Topic 34 and information about them. (See Figure 5 for keywords and word weights of Topic 34.) To avoid misunderstanding topic model results, it is important to understand information about characters that make up topics. Person (rén 人) has 219 occurrences in *Analects*, 611 in *Mencius* and 1,241 in *Xunzi*. Though accurate, for purposes of comparison this way of making the point neglects a couple issues. *Mencius* is 2.3x the size of *Analects*, and *Xunzi* is 5.3x the size of *Analects*, facts which hamper one's ability to interpret semantic importance from character frequencies alone. Zipf's law has the same effect. Zipf's law states that in any natural language given text, a word's frequency is inversely proportional to its rank in the corpus. This means that, in a given text, the most frequent word is typically twice as frequent as the second most frequent word, three times as frequent as the third most frequent word and so forth. A better way of understanding the importance of a character *in a set of texts* is to examine its rank within and across the texts and its rate of occurrence per 1,000 characters. Raw frequencies do not disclose

that rén is the most frequent character in each of *Analects, Mencius* and *Xunzi,* and has a rate of occurrence per 1,000 characters of 28.7, 34.6 and 30.6, respectively. To understand the importance of a character *in a topic* rather than in a text, however, we must consult its word weight. (See Figure 5, column 3.) By doing so, for example, we see that with a word weight of 0.037, nobleman (jūn 君 as in jūnzǐ 君子) is 3x as important to Topic 34 as is peace (ān 安).

Corpus weight is not a helpful statistic unless a topic's corpus weight is put in comparison with others. The corpus weight of Topic 34 is 0.375. Of 100 topics in our model, this is a very large corpus weight, ranking Topic 34's corpus weight eighth of 100 topics. (See Figure 3.) We can infer that, despite the fact that it was not nearly as representative in *Analects* and *Mencius*, *Xunzi'*s particular focus on ethical rulership is well-distributed in the corpus.[3]

[3] We thank an anonymous referee for several comments that led to substantial improvements in our presentation of these types of data throughout the paper

Figure 5 Word Weight & Character-Level Data for Topic 34

Term	English	Word Weight	*Analects* Occurrences	*Analects* Per 1000 characters	*Analects* Term Rank	*Mengzi* Occurrences	*Mengzi* Per 1000 characters	*Mengzi* Term Rank	*Xunzi* Occurrences	*Xunzi* Per 1000 characters	*Xunzi* Term Rank
君	**nobleman**	0.037	160	21.0	2	253	14.3	5	547	13.5	4
人	**person**	0.032	219	28.7	1	611	34.6	1	1241	30.6	1
義	**righteousness**	0.031	24	3.1	63	107	6.1	25	315	7.8	13
禮	**ritual**	0.022	75	9.8	9	68	3.8	40	343	8.5	10
能	**able**	0.021	69	9.0	12	135	7.6	12	519	12.8	5
賢	**virtuous**	0.018	25	3.3	60	74	4.2	37	152	3.7	44
莫	**none, do not**	0.016	18	2.4	89	58	3.3	53	257	6.3	18
天	**day/heaven**	0.016	49	6.4	23	293	16.6	4	598	14.7	3
惡	**evil**	0.014	39	5.1	38	80	4.5	36	190	4.7	30
安	**peace**	0.013	17	2.2	94	23	1.3	167	190	4.7	29

While topic models are a form of unsupervised machine learning, human decisions play some role in what topics are generated. At many junctures we pooled our expertise in programming, in preprocessing, in the classical Chinese language and thought to make decisions that influence the quality of the topics generated by the algorithm. The subset of decisions that are made prior to the application of a topic modeling algorithm to a corpus is referred to as *preprocessing*. Typically, preprocessing begins by stripping punctuation, tokenizing, stemming, segmentation, chunking and application of a stopword list. Due to the nature of our texts in Chinese, we removed punctuation, tokenized and applied a stopword list. For tokenization, we used a procedure that rendered each character separated by spaces before and after from every other character. This allowed us to treat as a unit of semantic meaning each character irrespective of whether it occurred as part of a two- or three-character word.

In a second preprocessing step, we used experts' knowledge to generate a stopword list. A stopword is a high frequency word that tends to be highly ranked in topics but that also tends to make the topics less valuable for interpretation. Stopwords typically consist of common function words. Applying a stopword list means removing those common characters from the corpus prior to analysis. Examples of terms on our stopword list are zhī (之), a grammatical particle which was used as a pronoun and subordination particle, and yě (也), a grammatical particle used to indicate noun predication (among other things). These and other stopwords were removed because during a series of pilot studies those words tended to blur the semantic coherence of topics. We provide a full list of stopwords used in this study in Appendix 4.

In a third preprocessing step, we encountered problems software to implement our topic modeling algorithm because that software was not designed to handle all the Chinese characters in our corpus. We scripted a method of encoding our input and decoding our output that allowed us to work around that problem.[4]

[4]MALLET's default tokenizing rules failed to process some characters in our corpus. To ensure all characters were counted correctly, we converted them to Unicode escape sequences, then to purely alphabetic equivalents, before importing the texts into MALLET. We then converted the MALLET output back to Chinese characters for analysis. A Python-based version of our conversion algorithm is available at <URL withheld>. (N.B. to referees: In accord with standard scientific practice, a Github URL released upon acceptance of submission.)

Moving from preprocessing to processing, the most important decision is the number of topics chosen to model. Too few topics may combine semantically unrelated material into so-called *chimera* topics; too many may cause related material to split into separate topics, redundance between topics and accumulation of irrelevant 'junk' topics (Schmidt 2012). Topic quality is typically determined by semantic coherence of the keywords in the topic. Although significant strides have been taken in algorithmic determination of the ideal number of topics (Marshall 2013), the assessment of topic coherence is typically a product of the scholar's interpretation. There is ongoing discussion in digital humanities scholarship over the interpretive significance of topics--whether they constitute subjects, themes, or discourses--and topic models do not always produce topics that appear semantically coherent to the scholar (Underwood 2012b). To the extent that text corpora are composed of figurative language, such as that found in poetry (our corpus contains some poetry), topic models produce higher rates of apparently incoherent topics (Rhody 2013). After experimenting with a number of models in several pilot studies using different numbers of topics, we settled on 100 topics, which, after the removal of stopwords, seemed to offer the best balance of scope and granularity while yielding the fewest junk or chimera topics.

After the topics are generated, researchers are faced with interpreting them and their relations to texts in the corpus. Some scholars in the field of ancient Chinese thought have argued that contemporary interpretations of ancient Chinese documents, especially philosophical, political, and religious documents, fall victim to debilitating biases and errors, for example, either Orientalizing or Westernizing the texts (Ames 2001). Since the texts were canonized long ago, a commentarial tradition two millennia long continues to structure the (presumed) central

themes of the early Chinese source texts. But this tradition makes assumptions that are open to re-examination. Topic modeling has the potential to reveal the unexpected and even challenge canonical claims about themes and contents of these texts, opening up new avenues for our understanding of ancient and medieval Chinese thought.

At the same time that our results may challenge leading interpretations of certain texts, we are well aware that our interpretations of the topics may be subject to biases of which we ourselves are unaware. Since three of the six of us publish actively in early Chinese thought, we aimed to minimize scholarly biases of our own that, unbeknownst to us, might influence our interpretations of our topics. For this reason, we decided that interpretation of our topics should be informed by independent expert knowledge in historical Chinese thought and language. So we enlisted the help of over 60 experts in the field to independently code topics. We refer to these results frequently in what follows to demonstrate a partial validation of our interpretations. This process worked as follows. These expert coders were presented with word clouds showing a target topic's keywords. First they were given an open-ended question reading, "Suppose you had to guess what is the theme of this word cloud. What are one to three English words you would use to describe this theme?" Second, experts were asked how confident they were about their judgment in the open-ended question. Third, experts received a forced-choice question with answers enabling us to probe their opinions about the contents of these topics. They could tell us, there, that the topic is about military, politics, philosophy, mind, etc. Due to the likelihood of chimera or junk topics, and limitations among our experts, we included an option of 'uncategorizable' as well. Fourth and finally, if an expert coder responded to a top-level multiple-choice saying that, in her opinion, the topic was about *military*, they would receive a

supplemental forced-choice question inquiring whether the topic represented issues including *weaponry, peace, the state, war, violence, order*, and/or *government*. (Note that experts were always able to select multiple answers.) These three levels of answers allowed us to use the expertise of generous volunteers knowledgeable about ancient and medieval Chinese thought to partially confirm or contest our broad, and our granular, interpretations of specific topics. (See Appendix 2 for the survey text and for an example of a word cloud.[5])

To take an example from our own corpus, consider Topic 27 in Figure 6. Traditional scholars skeptical of our methods may think that a topic as incoherent as 27 is evidence that our method is of little assistance in answering research questions about early Chinese thought. However, to experts of Warring States philosophical discourse on logic and language associated with Later Mohists and the School of Names, this topic makes perfect sense. These logicians focused on problems of reference (zhǐ 指) and how words are related to "things" (wù 物). They wanted to know whether a "white horse" (báimǎ 白馬) is a "horse," a famous example, and how attributes such as "hard and white" (jiān bái 堅白) relate to substances. Such is our initial interpretation, but to minimize our own bias and error, we take additional steps. We partially confirm this interpretation of Topic 27 by reviewing its text weights in specific texts to determine in which documents the weight of Topic 27 is heaviest. Since it loads heaviest in the School of Names text *Gōngsūnlóngzi* (公孫龍子), this provides further vindication of our interpretation. We then examined responses from our independent expert coders to determine whether their interpretations were supportive of our 'Logic and Language' interpretation. One of

[5] N.B. to referees: In accord with standard scientific practice, survey data will be released upon publication in the form of an online appendix.

three experts assigned Topic 27, presumably not as knowledgeable about philosophical materials in our corpus as about other materials, did not understand this topic. This was revealed in his or her answer to the top-level forced-choice question, uncategorizable. The other two coders agreed that it was a coherent topic. Furthermore, these two knew precisely what this topic was about. In open-ended questions, they reported that Topic 27 concerned "logicians, philosophy," "disputation," and "appearance, language."

Figure 6 Topic 27 Keywords & Weights

Chinese	Pinyin	English	Word Weight
馬	mǎ	horse	0.049
白	bái	white	0.04
物	wù	thing	0.035
生	shēng	birth, life	0.033
汝	rǔ	you	0.031
無	wú	without, nothingness	0.028
見	jiàn	see	0.022
指	zhǐ	finger, point	0.022
色	sè	color	0.019
列	liè	column	0.019

3 What topics make *Analects*, *Mencius* and *Xunzi* each unique?

Longstanding debate surrounds the relationship between Confucius of *Analects* and his two declared successors, Mencius and Xunzi (Lau 2000 [1953]; Van Norden 1992). The notion that it is *Mencius*, rather than *Xunzi*, which is the true inheritor of the teachings of Confucius contained in *Analects* has deep roots in the traditional Chinese commentarial tradition.

Tang Dynasty scholar Han Yu (768–824) first asserted authentic transmission of the teachings of Confucius ended with Mencius. The point was reiterated by Song dynasty (960-1279) Neo-Confucians and canonized by Zhu Xi's (1130-1200) inclusion of *Mencius* in the collection of Confucian texts referred to as *Four Books* (along with *Analects, Great Learning* (Dà Xué 大學) and *Doctrine of the Mean* (Zhōng Yōng 中庸). The Four Books were a central part of the core curriculum memorized by students and examination candidates from the early 1300s to the abolition of the examination system in 1905. Xú Fùguān 徐復觀 (1904-1982) reinforced the traditional idea that when Confucius speaks of 'human nature' he is expressing the same idea that Mencius later formulated, viz. that human nature is good (Xú 1969, 89; see also Zhāng 2012, 197). Following in Xú's footsteps, influential contemporary scholar Fù Pèiróng 傅佩榮 (1950-) argues Mencius' theory of the potential for goodness latent in human nature "is an excellent expression of Confucius's thought" (Fù 2011, 872).

Others argue this traditional interpretation is problematic and that analysis of the language of self-cultivation, including craft metaphors, indicates closer affinities between *Analects* and *Xunzi*. If human nature is like a raw piece of jade, values have to be carved into it by an outside force (*Analects* 1.15, tr. adapted from Slingerland (2003, 6-7); *Xunzi* 27.514-523, tr. Hutton 2014, 309). Knowledge of normative values is not innately present in human nature; it has to come from an external source. In contrast, *Mencius* contains an internalist theory that assumes normative values to be innate. Human nature is a seed with potential to grow into a fully developed plant (Ivanhoe 2000; Slingerland 2003).

Which topics do we select for analysis in order to initiate a new conversation about this canonical issue? In this section we examine those topics with text weights in one *and only one* of *Analects*, *Mencius* and *Xunzi* such that this placed the respective topics in a given text's top ten weightiest topics. (See Figure 7.) In other words, in this section we discuss only topics that render each of these texts unique and *different from* one another. In this section we discuss Topics 5 and 82 in *Analects*, 10, 99 and 86 in *Mencius*, and 23, 71, and 17 in *Xunzi*. In the following section we discuss those topics that our texts share *in common with* one another.

Figure 7 Weightiest 10 topics in each of *Analects*, *Mencius* & *Xunzi*

Topic	Keywords	Text Weight in *Analects*
61	孔 問 仁 言 人 禮 行 聞 道 貢	0.307
76	君 人 公 能 死 見 欲 知 先 得	0.130
63	禮 君 人 喪 士 父 樂 母 侯 廟	0.069
78	人 知 言 名 用 治 能 欲 學 文	0.074
21	民 君 行 國 治 能 得 事 政 下	0.040
34	君 人 義 禮 能 賢 莫 天 惡 安	0.043

5	大 祭 食 門 婦 先 入 既 服 出	0.038
33	人 大 天 知 王 得 世 — 心 已	0.026
29	天 上 下 大 道 中 人 時 後 地	0.034
82	公 王 德 成 事 民 告 用 聞 既	0.029
Topic	**Keywords**	**Text Weight in *Mencius***
21	民 君 行 國 治 能 得 事 政 下	0.102
61	孔 問 仁 言 人 禮 行 聞 道 貢	0.121
33	人 大 天 知 王 得 世 — 心 已	0.122
99	王 人 下 孟 取 相 或 士 他 好	0.114
76	君 人 公 能 死 見 欲 知 先 得	0.089
29	天 上 下 大 道 中 人 時 後 地	0.041
18	下 王 詩 天 亡 士 得 侯 善 臣	0.043
86	文 利 學 用 大 古 賢 義 能 今	0.036
63	禮 君 人 喪 士 父 樂 母 侯 廟	0.027
10	今 心 後 力 憂 豈 朝 死 誠 棄	0.027
Topic	**Keywords**	**Text Weight in *Xunzi***
34	君 人 義 禮 能 賢 莫 天 惡 安	0.256
78	人 知 言 名 用 治 能 欲 學 文	0.084
29	天 上 下 大 道 中 人 時 後 地	0.057
71	人 治 事 法 世 行 功 明 主 亂	0.058
21	民 君 行 國 治 能 得 事 政 下	0.053
23	天 道 下 物 知 德 生 能 聖 人	0.052
76	君 人 公 能 死 見 欲 知 先 得	0.038
18	下 王 詩 天 亡 士 得 侯 善 臣	0.039
17	國 法 民 兵 賞 力 利 刑 重 上	0.035
63	禮 君 人 喪 士 父 樂 母 侯 廟	0.025

3.1 *Analects*

We focus first on *Analects*, which is a collection of sayings attributed to Confucius (551-479 BCE) and his followers and contains material likely dating predominantly to the early Warring

States.[6] Two topics in the top ten differentiate *Analects* from other texts, including other texts within Confucianism. These are Topic 5, with a text weight in *Analects* of 0.038, which we label "Sacrifice, Ritual, Mourning," and 82, with a text weight of 0.029, which we call "Virtuous Rule & Moral Suasion." Since the text weight of Topic 5 in *Analects* is 0.038, this means that 3.8% of *Analects* is composed of the clustered terms representing Topic 5. (See Figure 8.)

[6] The bulk of the textual material in *Analects* was composed over the span of at least several centuries from the early Warring States period to the 3rd century BCE. See Cheng (1993, 313-23); Makeham (1996); Qu Wanli (1983, 382-9); Brooks and Brooks (1998) and Slingerland (2000). As indicated by Hunter (2014), its compilation likely occurred in the Han dynasty.

Figure 8 Topics Differentiating *Analects, Mencius* and *Xunzi* from one another

Document	Text Weight	Topic	Corpus Weight	Name	Topic Keywords in Descending Order of Weight
Analects	0.029	82	0.22	Virtuous Rule & Moral Suasion	公 王 德 成 事 民 告 用 聞 既 實 能 先 政
Analects	0.038	5	0.19	Sacrifice, Ritual, Mourning	大祭 食 門 婦 先 入 既 服 出 飲 相 小 哭
Mencius	0.027	10	0.35	Past & Future Cognition	今 心 後 力 憂 豈 朝 死 誠 棄 觀 入 罪 古
Mencius	0.114	99	0.2	Language of *Mencius*	王 人 下 孟 取 相 或 士 他 好 長 舍 章 羊
Mencius	0.036	86	0.2	Benefit & Moral Excellence	文 利 學 用 大 古 賢 義 能 今 國 商 良 富
Xunzi	0.052	23	0.45	Moral-Cosmic Education	天 道 下 物 知 德 生 能 聖 人 得 身 言 神
Xunzi	0.058	71	0.25	Legal Order	人 治 事 法 世 行 功 明 主 亂 亡 得 相 用
Xunzi	0.035	17	0.21	Institutional Rulership	國 法 民 兵 賞 力 利 刑 重 上 勝 官 戰 爵

Keyword characters in Topic 5 include great (dà 大), sacrifice (jì 祭), feed or eat (shí 食), gate or school (mén 門), woman, wife (fù 婦), first or before (xiān 先), enter (rù 入), submit or ritual garb (fú 服), exit or go out (chū 出), drink (yǐn 飲), assist or assist someone (xiàng/xiāng 相), and weep or cry (kū 哭). These terms describe a semantic space revolving around important

rituals and sacrifices, particularly those involved in ancestor worship and mourning. Independent coders reported that this topic concerned ritual and religion. Topic 5 has heavy text weight in only a handful of the texts in the corpus including in *The Classic of Rites* (Lǐjì 禮記, 0.175) and *Yili* (儀禮, 0.170) and *The Rites of Zhou* (Zhōulǐ 周禮, 0.052), which contains the core of the Book of Changes or *Yìjìng* (易經). (See Figure 9.) Together these three texts form a unit known in the Chinese commentarial tradition as the *Three Ritual Texts* (Sānlǐ 三禮). The bulk of each of these works consists of long lists of ritual prescriptions specifying the correct way of executing various rites and sacrifices, for example, specifying which clothes to wear and which color of accouterments to use. In form and content, parts of the *Analects,* especially Chapter 10, are strikingly similar to the *Three Ritual Texts*. This is a distinctive feature of *Analects* in comparison to *Mencius* and *Xunzi*. The fact that unsupervised topic modeling is able to precisely track this scholarly insight powerfully demonstrates the potential of this new research tool.

Results of Topic 5 appear to have important implications in adjudication of an ongoing debate about the role of sacrifices and spirits in *Analects*. Consider the opinion of a key voice in Chinese intellectual history about Confucius, spirits and sacrifices. Feng Youlan (馮友蘭) writes that Confucius ‘‘displayed a rationalist attitude [toward spirits], making it probable that there were other superstitions of his time in which he did not believe’’ (Feng 1952-3, I, 58). Thomas Wilson recently used a close-reading method to oppose this interpretation and emphasize the *Analects*’ advocacy of ritual, sacrificial rituals to ancestors and deities in particular. Wilson reasons that “Contrary to modern accounts, imperial-era commentaries on the *Analects* 論語 disclose the figure of Confucius as committed to pious sacrifices to gods and spirits” (2014, 185).

Unlike Xunzi, who explicitly reports his intentions to endorse the use of sacrifice for social-functional reasons (Ch. 19, "Discourse on Ritual"; see Campany 1992), the text of the *Analects* leaves readers uncertain what are Confucius' intentions about sacrifice. For this reason, debate about Confucius' relation to sacrifice will not be easily settled by topic modeling or by close reading. Feng Yu-lan cites *Analects* 7.20 and Wilson cites *Analects* 3.6; Feng Yu-lan cites 6.22 and Wilson cites 3.12; and so on. The fact that numbers of scholars argue that belief in gods in early China has prudential, not rational, origins is one indication that Wilson is likely correct. Prudential concerns arose through divination and knowing the future (Overmyer et al., 1995), ancestor reverence and seeking ancestors' blessings (Eno, 1990a, 1990b) and avoiding curses through shamanism (Ching 1997). Interpretation of Topic 5 offers some evidence in favor of the unique importance of practices associated with these sources of religion, especially religious ritual and sacrifice (jì 祭), for the compilers of the received *Analects*. This, in turn, could be seen as bolstering Wilson's position vis-a-vis that of Feng Yu-lan. If the compilers of the *Analects* were as rationalist as, say, Xunzi, we would not expect to see Topic 5 so prominently, and uniquely, featured in this text.

Topic 82, "Virtuous Rule & Moral Suasion," is a reflection of the fact that numerous passages in the *Analects* discuss the importance of the charismatic virtue (dé 德) of rulers, dukes (gōng 公) and kings (wáng 王) as they govern (zhèng 政) the people (mín 民). Rulers are advised to employ officials with virtue (dé 德) and ability (néng 能) to serve (shì 事) them by bringing affairs (shì 事) to completion (chéng 成). Independent coders reported in open-ended questions that this topic concerns "history, statecraft, philosophy" and "civil-affairs, reports, officialdom."

Turning to the word ranks of its keywords across the three target texts, we see governance or order (zhèng 政) is much more important in *Analects* (43 occurrences, 31st in rank, 5.6/1000 characters) than to authors of *Mencius* (54 occurrences, 57th in rank, 3.0/1000) and *Xunzi* (95 occurrences, 85th in rank, 2.3/1000). Topic 82 is the heaviest weighted topic in *Guoyu* (國語), which is a collection of historiographical/fictional anecdotes set in the pre-Qin period. Many of these anecdotes feature professional persuaders or diplomats who use their command of language to persuade rulers to "do the right thing."

Figure 9 Topic 5's Text Weights Across Texts in the Corpus

Text	**Text Weight of Topic 5**
Yílǐ 儀禮	0.175
Lǐjì 禮記	0.17
Zhōulǐ 周禮	0.052
Dàdàilǐjì 大戴禮記	0.04
Analects (Lúnyǔ) 論語	0.038
Báihǔtōngdélùn 白虎通德論	0.034
Mùtiānzǐzhuàn 穆天子傳	0.033
Kǒngzǐjiāyǔ 孔子家語	0.032
Shìmíng 釋名	0.029
Ěryǎ 爾雅	0.027

Topics 5 and 82 represent themes that are unique to *Analects* and not shared with *Mencius* or *Xunzi.* The prevalence of Topic 5, "Sacrifice, Ritual, Mourning," and Topic 82, "Virtuous Rule & Moral Suasion," confirms the scholarly consensus that Confucius (as he is portrayed in *Analects*) was a moral philosopher who emphasized the value of ritual and sacrifice

as elements in individual self-cultivation practice. The normative values enshrined in this tradition could be transmitted to disciples and followers through teaching. In many reported conversations with rulers, these same values were transmitted through advice on how to govern a state through virtue rather than laws and military force. These topics focus on institutionalized ritual practice and do not reveal much concern with cognition or emotion or other internal states of mind.

3.2 *Mencius*

Mencius was a self-proclaimed follower of the teachings of Confucius and flourished as an adviser to rulers in the mid- to late 4th century BCE. *Mencius* is generally agreed to have been composed in the Warring States period. Three topics set *Mencius* apart from *Analects* and *Xunzi,* Topic 10, "Temporal Cognition," Topic 86, "Benefit and Moral Excellence," and Topic 99, "Stylistic Features of *Mencius*". Both 10 (0.027) and 86 (0.036) have rather light weights in *Mencius* in comparison to Topic 99 (0.114).

Topic 10 is difficult to characterize, a point reflected in our coding results. In open-ended questions, expert coders described this topic as concerned with "loyalty, official service," "emotion, masculinity, mind," "psychology, self-cultivation, morality," and "mind, emotion, leave." Topic 10 is dominated by two sets of words. The first concerns temporality and includes present (jīn 今), later (hòu 後), and ancient (gǔ 古). We take this to be indicative of *Mencius'* frequent comparison between a golden age of the past and the fallen present. The second set concerns cognition or thought. This includes heart-mind (xīn 心), worry (yōu 憂), sincere (chéng

誠),regard or gaze upon (guān 觀), and guilt or crime (zuì 罪). Heart-mind represents the seat of cognition and emotion and is a common term in *Analects*, *Xunzi* and *Mencius*. Cognition and emotion terms cluster in this topic in part because *Mencius* focuses on internal reflection and mental regulation of emotion. Topic 10 has heavy text weights in two texts, *Yandanzi* (0.085) and *Jian Zhu Ke Shu* (0.084).

Topic 86 appears to represent "Moral Excellence and [criticism of] Profit." Independent coders reported that this topic is concerned with "culture and profit," "learning, cultivation of culture," "ethics," and "study, ancient, benefit." The top term, pattern or culture (wén 文), is used in names (King Wen 文王, Duke Wen, etc.) in all but 4 of the 51 occurrences in *Mencius*. These 4 refer to 'decorative pattern'; 'rhetoric'; '(rhetorical) style'; and 'to refine,' respectively. While wén does mean 'high culture, civility, or civilization' in other texts, it is not used in this meaning in *Mencius,* as observed by Bergeton (2013). Topic 86 is, therefore, a case where coders may be misled by the polysemy of the word wén. Benefit or profit (lì 利) is next. *Mencius* often criticizes the pursuit of profit (lì 利) as inferior to what is right (yì 義). As indicated by the inclusion of both present (jīn 今) and ancient (gǔ 古) in Topic 10, *Mencius* often contrasts amoral behavior of the present with morally superior behavior of the ancient or Golden Age (gǔ 古). Study (xué 學) is a step on the path of self-cultivation. Study elicits innate potential to become virtuous (xián 賢) and good (liáng 良), traits needed to serve one's state (guó 國). Virtue terms such as these bind Topic 86 together. Topic 86 has a large text weight in only *Discourses on Salt and Iron*, a debate about taxation (0.159).

The contents of Topic 86 provide evidential support for an interpretation of *Mencius*' as advocating internalist belief in the innate potential for goodness in human nature. This engages *Mencius'* discussion at 3B9 in which he attacks the doctrines of Yáng Zhū (楊朱) and Mò Dí (墨翟), who advocate egoism and altruism, respectively. Mencian Confucianism repudiates these act-based ethics in favor of the cultivation of character (Csikszentmihalyi 2002). This is uncontroversial but it leads to an ongoing interpretive problem about self-cultivation. Consider Mencius' four 'sprouts' of virtues (sì duān 四端) in 2A6, where he writes that "if one is without the heart (xīn 心) of compassion, one is not human. … The feeling of compassion is the sprout of benevolence" (Van Norden 2008, 46; see also the archer analogy at 2A7). On one interpretation of these passages, the cultivation of feelings appears to be the source of moral virtue in *Mencius*, making *Mencius* representative of an 'internalist' streak. This allegedly contrasts with moral motivation and cultivation as found in *Analects* and *Xunzi*. These two texts are thought advocate a greater number of, and greater roles for, externalist sources of morality like ritual (lǐ 禮), patterned civility (wen) and rectification of names (zhèngmíng 正名). We submit that the fact that Topic 86 is central to *Mencius* and not to *Analects* and *Xunzi* provides key evidence in support of the interpretation that, within the Confucian tradition, *Mencius* is a uniquely and distinctively "internalist" (Slingerland 2003) or "inside-out" (Klein 2000; see also Wong 1991 & Ihara 1991) thinker.

Topic 99 appears to represent features of dialogic text in *Mencius*. Independent coders reported that this topic is concerned with "Mencius" and "sage, teaches, king." This bland description from coders provides some confirmation that Topic 99 stands apart from other topics

that have richer semantic content and coherence. The fourth most frequently occurring word in this topic is Mencius' own name (mèng 孟) which obviously occurs hundreds of times in the text, and the third is *xia* 下 , usually meaning "below" or "under." In this topic, it must be picking out the frequent use of *xia* 下 in *Mencius* chapter titles, which are classified into A (*shang* 上) and B (*xia* 下). The most frequent word is king (wáng 王). This probably reflects the fact that many of the dialogical exchanges in *Mencius* are between kings and Mencius himself. Topic 99 not only sets Mencius apart from Analects and Xunzi, it also sets Mencius apart from all other texts. Its weight in Mencius is 0.114, which is twice its weight in any other text. This aptly confirms our designation of this topic as reflecting "Stylistic Features of *Mencius*."

Although it lacks thematic coherence, Topic 99 is a good example of what we have come to call "stylistic" topics, which load almost exclusively in one text in the corpus and seem to represent textually-distinct clusters of terms. These tend to be dominated by meaningless function words not removed in our stopword list, stylistic tics or commonly used proper names, but they also sometimes point to distinctive conceptual themes. Tied for eighth most frequent word in the topic, for instance, is *hào* 好, "to be fond of, like, desire," which picks up Mencius's concern with internally-driven preferences and desires. Although they are perhaps less interesting philosophically or conceptually, these stylistic topics have potential use in tracing textual lineages, dating texts, classifying newly-discovered texts or picking up surprising continuities in themes between texts. The next heaviest text weight for Topic 99 after Mencius is the obscure *Yùzi* 鬻子 fragments (0.059), usually classified as "Daoist." This may tell us

something about stylistic or conceptual influences, or convergent thematic concerns, between these otherwise disparate texts.

3.3 *Xunzi*

Xunzi is a compilation of various texts, including philosophical essays, attributed to Xunzi, and exchanges between Xunzi and others. Like Mencius, Xunzi was a self-proclaimed follower of the teachings of Confucius. He flourished as a teacher and adviser to rulers in the 3rd century BCE. Although compiled in its present form in the Han dynasty, the bulk of the material in *Xunzi* was composed in the late Warring States period. Philosophically, the Xunzi is a 3rd century BCE development of core ideas in the *Analects* that incorporates ideas from other pre-Qin philosophies.

From a modeling perspective, what is most intriguing is the semantic scope of *Xunzi's* heavyweight topics. Independent coders reported that Topic 71,"Ordered Government," concerns "politics," "law, humanity, worldly," and "humanity, the world and dealing with affairs." At 0.058 this is the fourth heaviest topic in *Xunzi*. The following passage nicely illustrates how the 12 most prominent keywords in Topic 71 cluster in the *Xunzi*:

> There are men (rén 人) who create order (zhì 治); there are no rules (fǎ 法) creating order (法) of themselves. The rules (fǎ 法) of Archer Yi have not perished (亡), but not every age (shì 世) has an Archer Yi. . . . Thus, rules (fǎ 法) cannot stand alone, and categories

> cannot implement (xíng 行) themselves. . . . One who tries to correct the arrangements of the rules (fǎ 法) without understanding their meaning, even if he is broadly learned, is sure to create chaos (luàn 亂) when engaged in affairs (shì 事). And so, the enlightened (míng 明) ruler (zhǔ 主) hastens to obtain (dé 得) the right person (rén 人). . . . If one hastens to obtain (dé 得) the right person (rén 人), . . . [then] one's accomplishments (gōng 功) will be grand . . . . (*Xunzi* 12.1-20, tr. adapted from Hutton (2014: 117))

As shown here, 明 míng (bright, clear; perspicacious, enlightened) often refers to the far-sightedness associated with sages and desired in rulers (zhǔ 主) in early China (Brown and Bergeton 2008). By obtaining (dé 得) and the right person (rén 人) to assist him, the ruler can create order (zhì 治) and avoid chaos (luàn 亂), thereby achieving great accomplishments (gōng 功). The ruler's discernment is therefore more important than blind enforcement of rules or promulgated models (fǎ 法). This sounds like a classically Confucian claim, albeit with much more emphasis on institutional structures than we see in *Analects* or *Mencius*. This thematic distinctiveness is in turn reflected in the fact that Topic 71 is completely absent from *Analects* and loads at 0.010 in *Mencius*.

Another topic unique to *Xunzi* among the classical Confucian works is Topic 17, "Institutionalized Rulership," which loads at 0.035 in the *Xunzi* but is absent from both *Analects* and *Mencius*. This topic similarly involves ordering the state (guó 國), but through what sounds like much more Legalistic means. The people (mín 民) and officials (guān 官) can be managed with rewards (shǎng 賞), noble rank (jué 爵), and profit (lì 利). Punishments (xíng 刑) figure into

Legalist governance, particularly heavy (zhòng 重) ones. This topic also includes terms related to violence and force: troops, weapons (bīng 兵), victory (shèng 勝) and war (zhàn 戰). These terms for warfare are arguably more frequent in *Xunzi* than in *Mencius* and *Analects* due to *Xunzi'*s Chapter 15 "Debate on Military Affairs" (Yìbīng 議兵), and tend to be associated with military strategists, such as Sunzi, author of *The Art of War* (Sūnzi Bīngfǎ 孫子兵法), or Legalist thinkers such as Han Feizi, Xunzi's student. Since *The Art of War* and *Hanfeizi* are in our corpus, we can look to Topic 17's weight in them to confirm or disconfirm our reasoning. We find, indeed, that both texts appear among the heaviest texts into which Topic 17 loads, as do several other military and Legalist texts. Their presence here supports some scholars' view that *Xunzi* is more focused on institutional means of social control than is Confucius or Mencius (one citation here). Independent coders reported that this topic concerned "governance" and "legalism."[7]

[7] The following passage elegantly illustrates how the highest ranked words in Topic 17 are used in the *Hanfeizi*:

> If laws (fǎ 法) are equable, there will be no corruption among the officials. . . . [People] should only be able to obtain rank (jué 爵) through their own effort (lì 力). . . The state (guó 國) bestows office (guān 官) and rank (jué 爵) based on accomplishments; this is called devising plans with utmost wisdom and conducting war (zhàn 戰) with awe-inspiring courage. . . . If punishments (xíng 刑) are heavy (zhòng 重) and rewards (shǎng 賞) sparse, this means that the superior (shàng 上) cares for the people (mín 民), and that the people (mín 民) will die for rewards (shǎng 賞). . . . If profit (lì 利) issues from one

Fu Peirong (2011, 872) takes the fact that Xunzi was the teacher of prominent legalists, such as Hanfei and Li Si (the Prime Minister of the First Emperor of the Qin dynasty), to indicate that *Xunzi*'s theory of human nature as "bad" stands out against the theory of a human nature with innate potential for goodness he sees in *Analects* and *Mencius*. However, this may not be the best explanation for the prevalence of Topics 71 and 17 in the *Xunzi.* Large text weights of Topics 17 and 71 probably derive from *Xunzi*'s greater interest in discussing details of government institutions like 'laws,' 'punishments,' 'officials.' Unlike Confucius and Mencius, Xunzi had more personal experience serving as an official. Hence it is only to be expected that his practical and less theoretical discussion would lead him to write about "Institutionalized Rulership" (Topic 17) and "Ordered Government" (Topic 71) more than *Analects* and *Mencius*.

The text weights of Topics 17, 71 and 23 show they are all highly representative of *Xunzi*. Yet Topic 23 (0.052 in *Xunzi*), with a corpus weight of 0.453, the fifth weightiest topic in the entire corpus, has much greater overall importance to ancient and medieval Chinese literature. Topic 23 has very heavy text weights across texts in the Daoist school including *Dao de Jing* (0.434) and *Heshanggong laozi* (0.358). It constitutes only 0.007 of the *Mencius*, however, and is completely absent from the *Analects*. We dub Topic 23 "Moral-Cosmic Attunement." Its heaviest terms are heaven or sky (tiān 天), way (dào 道), and under (xià 下), as in the world (tiānxià 天下), literally 'under heaven'. These terms all tend to refer to the structure of the

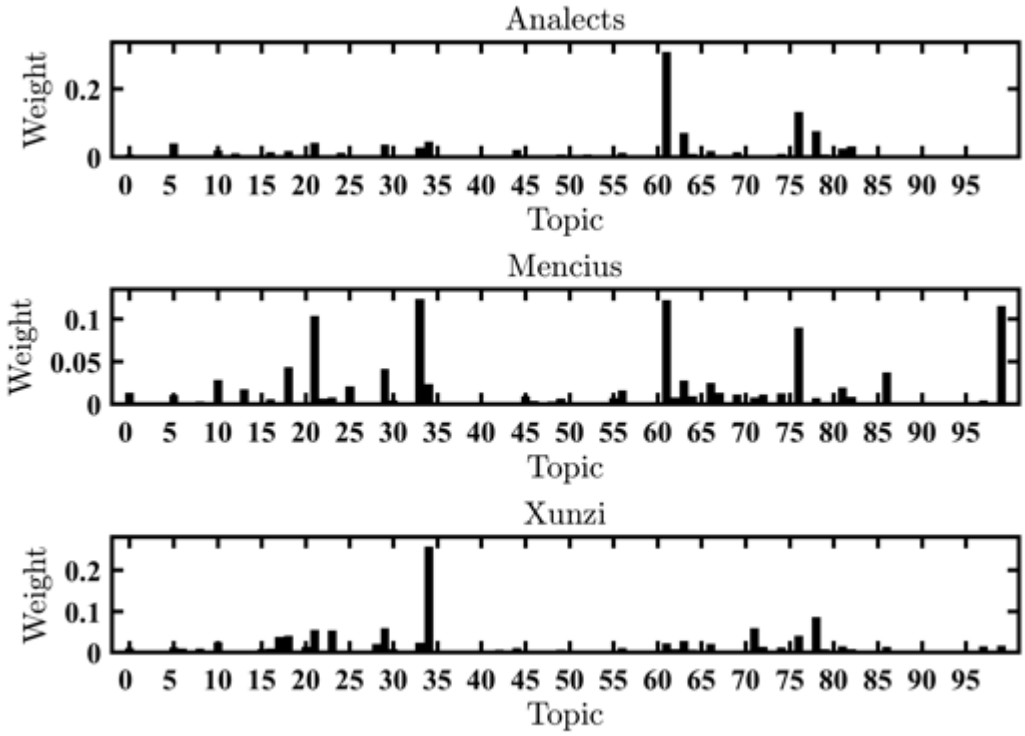
---

source only, the state (guó 國) will have no enemies; if profit (lì 利) issues from two source, its soldiers (bīng 兵) will be half useful. (*Hanfeizi* 53, Wáng Xiānshèn 王先慎 (2006: 471-3)

universe. In the *Dao De Jing* and the *Xunzi* the 'sky' or 'heaven' is an impersonal force, not a moral agent as in *Mencius* and *Analects*. Prominent moral terms include way (dào 道), virtue (dé 德), and sage (shèng 聖). Philosophical terms include way, creature or thing (wù 物), knowledge (zhī 知), and saying or teaching (yán 言).

To conclude this section, Topic 5 suggests that *Analects* emphasizes and discusses specific ritual prescriptions more frequently than *Mencius* or *Xunzi*. Topic 82 suggests that *Analects* proposes a toolkit for moral suasion and social change differing from those of the author's close intellectual ancestors. Topic 17 reveals *Xunzi* is more interested in detail-oriented discussions of government institutions than *Mencius* or *Analects* and 71 that *Xunzi* has a more elaborate theory of the use of legal structures and military force than *Analects* and *Mencius*. Topic 23 indicates that Xunzi, despite his critique of Daoist thought, shares with Daoists an interest in moral-cosmic attunement that is absent from the *Analects* and *Mencius*. Topics 86 and 10 are not as thematically well defined as Topics 5, 17, 71, and 23 and possess large text weights in a variety of texts across different genres. The fact that 86 and 10 have heavy text weights in *Mencius* is less helpful in setting this text apart from *Analects* and *Xunzi*. In spite of this, the overall match between machine-generated topics and scholarly studies of the defining characteristics of these three texts remains impressive. This should enhance our confidence in the general method. Furthermore, the details of our findings represent original contributions to the scholarly debate by either weighing in on one side or suggesting novel lines of attention or inquiry.

Figure 10 Text Weights in *Analects*, *Mencius* & *Xunzi* Across the Corpus



4 Intersecting topics in *Analects, Mencius*, and *Xunzi*

In the previous section was intended to provide an understanding of what makes each text different from the other two. Those discussions combine with the discussion of this section to preliminarily address our guiding research question about whether *Mencius* or *Xunzi* better represents the content of *Analects.* Using the same list of the ten most weighty topics in each of our three texts, we calculated the topic intersections between documents as a shared set of

topics.[8] Now we focus on topics that load into pairwise unions of these three texts in an effort to explore not their differences but their similarities. Before discussing those topics, however, we briefly report on the four topics that lie at the union of all three texts. (See Figures 11 & 12.)

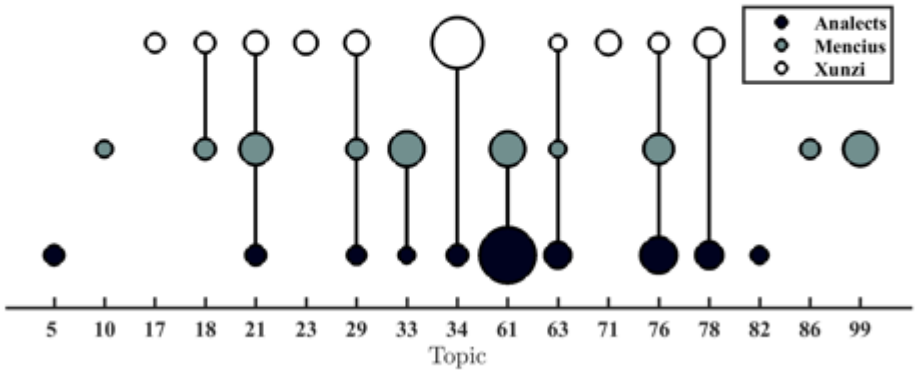
Figure 11 Formal Interpretation Matrix of Intersections of *Analects, Mencius* and *Xunzi* with Topic Keywords (∩=intersection of sets)

Document	Topic's Weight in Text (Text Weight)	Topic	Topic's Weight in Corpus (Corpus Weight)	Name	Topic Keywords in Descending Order of Weight
(Mencius ∩ *Xunzi* ∩ *Analects)*	0.04/0.06/0.03	29	0.6	Heaven & the Way	天 上 下 大 道 中 人 時 後 地 長 從 成 德
(Mencius ∩ *Xunzi* ∩ *Analects)*	0.1/0.05/0.04	21	0.46	Political & Social Order	民 君 行 國 治 能 得 事 政 下 食 教 官 道
(Mencius ∩ *Xunzi* ∩ *Analects)*	0.03/0.03/0.07	63	0.18	Ritual, Family & Governance	禮 君 人 喪 士 父 樂 母 侯 廟 親 主 命 事
Xunzi ∩ *Analects*	0.25/0.04	34	0.37	Ethical Rulership	君 人 義 禮 能 賢 莫 天 惡 安 亂 下 善 性
Xunzi ∩ *Analects*	0.08/0.07	78	0.37	Learning & Governance	人 知 言 名 用 治 能

[8] To put this point semi-formally, A ∩ B is the intersection of A with B, that is, the set of all the elements in A that are also contained in B and not contained in any other elements. We applied this definition to a 10*3 matrix, representing ten topics (for each document) in rows and three documents in columns. See Figure 12.

					欲 學 文 小 富 彼 盜
Mencius ∩ *Analects*	0.12/0.3	61	0.19	Language of *Analects*	孔 問 仁 言 人 禮 行 聞 道 貢 仲 學 知 路
Mencius ∩ *Analects*	0.12/0.03	33	0.08	Human, Heaven and Political Order	人 大 天 知 王 得 世 一 心 已 義 且 今 見

Figure 12 Topic Intersections in *Analects, Mencius* and *Xunzi*



Caption: Topic intersections of ten most central topics for each document. Circles represent presence of topic within the ten most central topics, circle size is proportional to the relative within document centrality (topic weight) and links (vertical lines) indicate an intersection.

At the intersection of *Analects, Mencius*, and *Xunzi* are Topics 29 "Heaven, Cosmos & The Way," 76 "Roles of Rulership," 21 "Political & Social Order," and 63 "Ritual, Family, & Governance." These topics possess some of the largest corpus weights of all 100 topics in our

model. 29 is ranked #1, 76 #3, 21 #4, and 63 #29. Our expert coders report that Topic 29 is concerned with "cosmology, virtue" and "cosmology, time, philosophy." They report Topic 76 is concerned with "Lordly leadership" and "politics and leadership." Topic 21 is concerned with "governance, kingdom, people" and "people, masses" and Topic 63 with "ritual, "masters of ritual (rú)," "ritual, rites, ceremonies."

This information supports two major inferences about shared semantic content in these texts. First, these heavy corpus weights provide strong evidence that topics at the heart of early Confucianism are exceptionally well seeded throughout ancient and medieval Chinese literature. When we compare topics weighty in these texts with topics weighty in what are traditionally referred to as "Daoist," "Legalist" and "Mohist" texts, we find that those loading into Confucian texts are much, much more likely to be represented with heavy corpus weights.[9] Second,

[9] The CTP genre classification emerges from traditional Chinese content-based and form-based library taxonomies as well as recent categories. The five genre categories "Confucianism," "Mohism," "Daoism," "Legalism" and "School of Names" are English translations of a classification system that can be traced back to the Western Han scholar and historian Sima Tan (c. 165-100 BCE; see Csikszentmihalyi 2002; see Goldin 2011 on "legalism"). The "School of the Military" and the "Miscellaneous Schools" categories can be traced back to Ban Gu's (32-92) classification of books in the *Han Shu*. Knowledge of the representation of genre in our corpus is helpful for the sake of interpreting topics. For example, the biggest genre is History (53%) followed by Confucianism (16%) and Ancient Classics (6%). However, due to several shortcomings of the CTP classification of genre, we do not use genre for analyses. Consider

consider that Mencius and Xunzi both self-identified as masters of ritual (*rú* 儒), who followed in Confucius' *rú* footsteps. The fact that *Analects*, *Mencius* and *Xunzi* have shared interest in Topics 29 ("Heaven & the Way"), 76 ("Roles of Rulership"), 21 ("Political & Social Order"), and 63 ("Ritual, Family & Governance") indicates that the pre-Qin concept of *rú* referred to a set of philosophies characterized by a high degree of internal coherence. Earlier we noted that Topic 86 separated *Mencius* from other texts by virtue of its internalist perspective about moral motivation and normativity. We contrasted internal moral motivation from external motivation, which we associated with ritual and law. Here, however, we see that *Mencius* (0.027) and *Xunzi* (0.025) share in Topic 63, "Ritual, Family & Governance," to the same degree. This topic is fronted by ritual or rites (lǐ 禮, word weight=0.042), which is why it is strong in *Analects* (0.069). This apparent conflict in our interpretation can be resolved by keeping in mind that Mencius's internalist stance naturally makes him less interested in discussing the external tradition embodied in the rituals and rites, but this does not mean that he dismisses them altogether.

To appreciate how results at the intersection of all three texts might inform current debate, let's continue examination of Topic 63 in light of some secondary literature. Topic 63 does not prominently feature moral terms. The difference between Topic 34, sitting at the intersection of

---

CTP's "Excavated texts" category. The fact that it includes documents that CTP calls "*Mawangdui*" and "*Guodian*" mistakenly leads users to infer that these documents represent the Mawangdui and Guodian manuscripts. In fact these CTP documents only represent different versions of the *Dao De Jing*. Further, the "Ancient Classics" genre includes Song Dynasty material. Thus we exercise extreme caution in using some of the CTP genres.

only *Analects* and *Xunzi* but not *Mencius*, and Topic 63 allows us to grasp a subtle but important difference in the social scope of the shared moral ideal across the three books. Topic 34 sees the virtue of righteousness (yì 義, word weight=0.031) traveling together with terms connoting high social status like lord or nobleman (jūn 君,=0.037) and rituals (lǐ 禮=0.022). This informs our understanding of what E. Brindley has called the "sociology of the junzi." Specifically, considerations about how the distribution of Topics 63 and 34 differs between the three books can add considerable subtlety to this scholarly debate. In contrast to *Mencius'* appeal to internal states (see discussion of Topic 86 above), *Analects* and *Xunzi* have in common a linkage between rites that accord high social status and certain moral virtues. Righteousness, with its connotations with animal sacrifices to gods, not benevolence, has a particularly heavy word weight in Topic 34.

This suggests an interpretive hypothesis worth exploration through traditional scholarship. Some scholars, including P.J. Ivanhoe, argue that the Confucian "ethical ideal" is "something anyone can achieve and a way of being human that can be manifested in a wide range of social roles" (2008, 5). Others concur (Wills 2012, 25; Hsu 1965, 162). However, Brindley (2009), echoing Hall and Ames (1987, 188), argues with some force that achievement of the status of gentleman or nobleman (jūnzi 君子) is restricted to high status males, or males who are entitled to perform certain rites.

While we concur with Brindley on this matter, our topic modeling results suggest value in further research on two pleasingly concrete questions: First, is the *jūnzi* ideal preferentially associated with ritual and the virtue of righteousness, rather than other virtues benevolence (rén

仁)? This is suggested by an analysis of Topic 63 and the word weights of its keywords. Second, consider that Brindley states in the title of her paper that her thesis is restricted to *Analects*. When that restriction is accompanied by claims about "Confucian" morality it can sow confusion. Might there be significant cross-textual variety in Early Confucian texts' association of the *jūnzi* ideal with high social status? Topic 63's text weights in *Analects, Mencius* and *Xunzi* place it in each text's top ten, but distribution into *Analects* is twice that of its distribution into *Mencius* and *Xunzi* (see Figure 7). This raises the probability that Brindley and Ivanhoe are talking past one another. Could it be that Brindley emphasizes what is true but only of *Analects,* while Ivanhoe emphasizes what is true of *Mencius* and *Xunzi* but not *Analects*? We do not know, and do not presume to know, the answers to these questions. Yet on the strength of our results, we surmise this is likely to be true. More important, our interpretation of the results suggests value in pursuing a concrete research question with close-reading methods to narrow in on an answer.

Since our interest lies in renewing a conversation about the influence of *Analects* on subsequent Ruist and Confucian thought, we move from our brief review of topics at the intersection of all three texts to discussion of those topics that intersect only in pairwise fashion. The topics at the intersection of *Analects* and *Xunzi* include Topics 34, "Moral Leadership," and 78, "Learning & Governance." The corpus weights of Topics 34 and 78 rank #8 and #9 of 100 total topics, which suggests they have wide influence. Independent coders report that Topic 34 is concerned with "subject-ruler relations" and "virtue, politics," and that Topic 78 is concerned with "governance, learning, talent" and "human, knowledge, culture."

Topic 34 has a heavy text weight in Xunzi (0.256) and moderate weight in Analects (0.043) but rather low weight in Mencius (0.023). Why might Topic 34, on "Ethical Rulership,"

not load as heavy into *Mencius*? Since moral leadership is a common theme in that text, this finding is intriguing. Investigating heavyweight characters in Topic 34, *Analects* and *Xunzi* represent rites (lǐ 禮) at very similar rates (9.8/1000 and 8.5/1000 respectively). But the rate of 禮 in *Mencius* falls far below this (3.8/1000). Since in the *Analects* and the *Xunzi* human nature (xìng 性) is initially ignorant of normative values, these texts recommend use of the rituals and rites (lǐ) to build morally refined gentlemen (jūnzǐ 君子) who possess traits such as righteousness (yì 義), worthiness (xián 賢), and goodness (shàn 善). If an individual does not assimilate traditions embodied in rites then he will be nothing but a savage or a beast. With rites and rituals, the Confucianism of *Analects* and *Xunzi* says that the nobleman orders himself and leads a state that is neither chaotic (luàn 亂) nor characterized by widespread badness (è 惡), but rather at peace (ān 安). While not unimportant, rites play a much less significant role in the philosophy of *Mencius*, since the Mencian strategy of self-cultivation is directed at motivating normative behavior through appeal to and training of the innate sprouts of virtue (Ivanhoe 1993), what Slingerland (2003) has termed the "internalist" Confucian strategy. Topic 34's emphasis on rites in *Analects* and *Xunzi* is just what we expect to observe given the 'internalist' morality represented in Topic 86, which had a heavy topic weight in *Mencius.*

Topic 78, "Learning and Governance," contains a few concepts, including 文 wén (pattern; patterned civility, high culture), emphasized in *Analects* and *Xunzi* but not *Mencius*. Learning (xué 學) and knowledge (zhī 知) have higher saturation in *Analects* and *Xunzi.* Their presence in the context of Topic 78 suggests that keys to rulership involve cognitive preparation of the mind for rule or management (zhì 治). This contrasts with the model of rulership discussed

in *Mencius* Books 1 and 2 in which kings are challenged to deeper levels of empathy and emotion. Contents of 78 raise the probability that *Analects* and *Xunzi* are semantically linked by virtue of their advocacy of a set of normative values deriving from learning (xué) and an external, refined (wén) tradition. The heavy text weights of Topic 78 in *Analects* (0.074) and *Xunzi* (0.084) also provide counter-evidence to claim that differences between *Mencius* and *Xunzi* on the contents of human nature are merely a matter of emphasis rather than the result of different views of the moral resources located within the individual (for a representation of this view, see Lau 1970: xix-xxii).

In terms of differentiating the influence of *Analects* on *Mencius* and on *Xunzi,* evidence weighs in favor of greater discursive overlap between *Analects* and *Xunzi.* This appears to *reduce* the probability that the traditional theory about *Mencius'* closer relation to *Analects* is correct. But one might suspect that consideration of topics at the intersection of *Analects* and *Mencius* will *increase* the justification of a closer relation between *Analects* and *Mencius*. At this intersection we have Topics 61, "Language of *Analects*," and 33, "Humanistic ethics." Both 61 and 33 occur in the top ten largest loading topics in *Analects* (0.307 and 0.026) and *Mencius* (0.121 and 0.122). Yet note they also both fall within the top 13 topics of *Xunzi* (0.020 and 0.021). As a result, we infer that no heavy loading topics falling at the intersection of *Analects* and *Mencius* successfully differentiate their semantic contents from *Xunzi'*s contents. This alone increases the probability that *Xunzi* tracks the semantic contents of *Analects* more closely than does *Mencius*, all things considered.

Ranked 65th, with a low corpus weight of 0.077, Topic 33 is not commonly represented in the corpus. We call this topic "Human, Heaven and Political Order." Independent coders

report that Topic 33 is concerned with "heaven, knowledge," "ruling," and "kingship." Three of three independent coders in forced-choice questions report that Topic 33 concerns leadership, though kingship and statecraft were also regarded as important. The dominant keywords in this topic are people (rén 人), big or great (dà 大), heaven (tiān 天), know (zhī 知), rulership (wáng 王), thinking (xīn 心), and righteousness (yì 義). Examining character frequencies, we find that terms from this topic often appear at similar rates in *Xunzi* and *Mencius,* and dissimilar rates in *Analects.* For examples, with respect to 心, it is the 14th most frequent term in *Mencius* (126 occurrences, 7.1/1000 characters) and 32nd in *Xunzi* (168 occurrences, 4.1/1000) but only the 255th in *Analects* (6 occurrences, 0.78/1000). This is so despite the fact that Topic 33 sits at the intersection of *Mencius* and *Analects* but not all three texts. Righteousness is the 13th most frequent term in *Xunzi* (315 total, 7.8/1000 characters) and the 25th in *Mencius* but 63rd (107 occurrences, 6.0/1000), in *Analects* (24 occurrences, 3.1/1000). The effect of these character level data is to raise doubts about Topic 33's ability to pull *Analects* and *Mencius* together and away from *Xunzi.*

Topic 61 ranks 29th in the corpus with a corpus weight of 0.188. Topic 61, "Language of *Analects*," contains the character the name of Confucius (kǒng 孔), as well as three other characters used in names of followers of Confucius (zhòng 仲, lù 路 and gòng 貢). We infer that Topic 61 represents linguistic features of *Analects'* and *Mencius'* literary style, particularly dialogic prose, explaining why it is neither prominent in *Xunzi* nor in the corpus as a whole. While the *Analects* and the *Mencius* consist mostly of reported dialogues, the *Xunzi* also contains lengthy philosophical essays. In sum, although topics 33 and 61 feature among the top ten topics

in the *Analects* and the *Mencius*, (61 is #13 and 33 is #12 in *Xunzi*), the *Xunzi* also contains high word frequencies of keywords found in 33 and 61. Examination of topics at the intersection of our three texts, and at the pairwise intersection of two of three of our texts, appears to serve as evidence to shift the burden of proof somewhat onto those traditionalists who argue that Mencius is the inheritor of Confucius' mantle.

5 Conclusion

Topic modeling is an extremely powerful tool for the study of the intellectual tradition embodied in the extant corpus of early Chinese texts. As illustrated here, topic modeling algorithms produced without being fed prior knowledge of ancient and medieval Chinese thought and literature produce a set of topics that accurately reflects insights by scholars working with traditional methods of analysis. The ability to replicate such scholarly consensus is quite remarkable and underlines the robustness of topic modeling data. More importantly, this "unsupervised" technique can uncover new or unexpected connections—or, sometimes more importantly, a lack of connections—invisible to the individual scholar reading through the early enormous Chinese corpus on his or her own.

We believe that textual scholars can benefit greatly from new methodologies such as topic modeling, as well as other automated, machine-learning means for the "distant reading" of texts. The widespread availability of textual corpora in digital form has yet to substantially alter the manner in which we approach our material—to date, they have been used primarily as glorified concordances. Techniques such as topic modeling represent entirely new ways to analyze and explore texts that can generate novel insights and allow us to grapple with

prodigious amounts of textual material. In the end, however, the true usefulness of topic modeling lies in how it can be brought to bear on controversial questions that divide scholarly opinions. In this paper, we have attempted to show how topic modeling can provide a fresh source of input that may help resolve age-old scholarly debates, such as questions concerning the intellectual relationship of *Analects*, *Mencius*, and *Xunzi*.

To be sure, our topic modeling approach has a number of limitations. First, in inexpert hands, far too many topics might be dismissed as uninterpretable 'junk topics'. Second, extensive polysemy in classical Chinese presents interpretive challenges for us and those who follow, e.g., 文 in Topic 86 would be mistaken as 'culture' without expert knowledge of its use in names and in other meanings (decoration, etc.). This why topic modeling and other automated techniques require teamwork between scholars with deep familiarity with the corpus in question and those able to employ digital humanities and statistical tools. This will no doubt require traditional scholars to challenge themselves to overcome aversions to the use of machine-learning. One of the goals of the research projects that have funded the current research is to not only raise awareness among humanities scholars of the existence of such techniques, but also to de-mystify them and make them, or their results, more easily accessible to the scholarly community. Most importantly, it should always be emphasized that "distant reading" techniques can never be a substitute for qualitative, close reading. Besides the obvious ways in which close reading is necessary for any genuine understanding of a text, the actual significance of automated results can never be assessed without reference to such understanding.

To conclude, our study of the thematic relationships between *Analects*, *Mencius* and *Xunzi* has been presented in the spirit of advancing new threads in old conversations. We are not

especially optimistic that the fruits of our narrow research here should lead to significant developments in the research programs of well-established, orthodox scholars. Yet we are confident that the coming wave of like-minded machine-learning research, to be conducted by a new generation of researchers in Philosophy, Religion, and Asian Studies, will lead to groundbreaking changes to our knowledge about Early Confucianism. We see this potential for a couple reasons. First, while machine-learning efforts like ours are subject to several forms of bias, such biases are no greater than those associated with traditional close reading methods, where scholars implicitly loyal to their professors or to a privileged theory (or to Confucius) sometimes fall into unproductive patterns of textual commentary reduplicated across generations. This is demonstrably true in the context of Early Confucian moral philosophy. There large groups of scholars continue to believe Early Confucian moral thought is best represented by each of a dozen Western normative ethical theories. This is so despite the fact that these theories are mutually inconsistent, a state of affairs described recently as a crisis in the field (Nichols 2014). Machine-learning efforts are likely to be especially good at breaking up longstanding interpretive stalemates in secondary literatures. Second, our small contribution confirms a number of scholarly opinions on several shared themes across these three documents. This is important since it suggests that our methods are sound. For example, our findings from Topic 34 support a theory that *Analects* and *Xunzi* share an 'externalist' theory of human nature and moral self-cultivation while findings from Topic 86 support attribution to *Mencius* of an 'internalist' moral philosophy, confirming widely disseminated interpretations in the secondary literature.

Many of the world's literary traditions are available in digital, fully searchable form, the result of enormous effort. This new format affords exciting possibilities for supplementing,

confirming or challenging our traditional qualitative techniques with entirely new quantitative methods capable of perceiving patterns invisible to human minds. Our results call for attention to a handful of explicit issues in ancient and medieval Chinese textual studies. More broadly, we hope that our preliminary distant reading of *Analects*, *Mencius* and *Xunzi* here gives a sense of the power, scope and possibility of these new tools—not as replacements for our traditional modes of analyzing texts, but as sources of potential new discoveries, interventions in ongoing interpretive cruxes, and catalysts for new conversations.

Bibliography

Ames, R. (2001). New Confucianism: A Native Response to Western Philosophy. In S. Hua, editor. *Chinese Political Culture: 1989-2000*. Armonk, NY: M.E. Sharpe.

Bergeton, U. (2013). *From Pattern to 'Culture'?: Emergence and Transformations of Metacultural Wén*. Ph.D. Dissertation, University of Michigan.

Blei, D. M. (2012a). Probabilistic topic models. *Communications of the ACM* 55: 77. http://doi.org/10.1145/2133806.2133826

Blei, D. M. (2012b). Topic Modeling and Digital Humanities. *Journal of Digital Humanities*, *2*(1). Retrieved from http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/

Blei, D., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning* 3: 993–1022.

Boltz, W. (2007). The Composite Nature of Early Chinese Texts. In M. Kern (Ed.), *Text and ritual in early China* (50–78). Seattle, Wash.: Univ. of Washington Press.

Brooks, E. Bruce, and A. Taeko Brooks. (1998). *The Original Analects: Sayings of Confucius and His Successors*. New York: Columbia University Press.

Brown, M and Bergeton, U. (2008). Seeing Like a Sage: Three Takes on Identity and Perception in Early China. *Journal of Chinese Philosophy* 35: 641-62.

Brindley, E. (2009). "Why Use an Ox-Cleaver to Carve a Chicken?" The Sociology of the Junzi Ideal in the Lunyu." *Philosophy East and West*, 59(1), 47–70. http://doi.org/10.1353/pew.0.0033

Chen, Jack W., Zoe Borovsky, Yoh Kawano, and Ryan Chen. (2014). The *Shishuo Xinyu* as Data Visualization. *Early Medieval China* 20: 23–59. doi:10.1179/1529910414Z.00000000013.

Cheng, A. (1993). Lun yǔ 論語. In M. Loewe (Ed.), *Early Chinese Texts: A Bibliographical Guide* (313–323). Berkeley, Calif.: Institute of East Asian Studies.

Cheng, A. (2009). Lun-yü. In J. Lagerway & M. Kalinowski, editors. *Early Chinese religion: Part one: Shang through Han (1250 BC–22 AD)* (313–323). Leiden: Brill.

Csikszentmihalyi, M. (2002). Traditional taxonomies and revealed texts in the Han. In L. Kohn & H. D. Roth, editors. *Daoist Identity: History, Lineage, and Ritual* (81-101). Honolulu: University of Hawaii Press.

Clark, K. J., and J. T. Winslett. (2011). The Evolutionary Psychology of Chinese Religion: Pre-Qin High Gods as Punishers and Rewarders. *Journal of the American Academy of Religion* 79: 928–60. doi:10.1093/jaarel/lfr018.

Fung, Yu-lan (Feng Youlan). 1952. *A History of Chinese Philosophy*. Princeton: Princeton University Press, 1952–1953.

Fù, Pèiróng 傅佩榮. (2011). *論語三百講（下篇)*. (聯經出版事業股份有限公司).

Goldin, Paul Ratika. 2011. Persistent Misconceptions about Chinese 'Legalism.' *Journal of Chinese Philosophy* 38: 88-104.

Goldstone, A and T. Underwood. (2014). The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us. *New Literary History* 45: 359-384. doi: 10.1353/nlh.2014.0025.

Hall, D. L., & Ames, R. T. (1987). *Thinking through Confucius*. Albany: State University of New York Press.

Hou Y. and A. Frank. (2015). Analyzing Sentiment in Classical Chinese Poetry. *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. Beijing: Association for Computational Linguistics and The Asian Federation of Natural Language Processing* (pp. 15–24). Stable URL: http://www.aclweb.org/anthology/W15-3703. Accessed 7-15-2016.

Hsu [Xu], Z. (1977). *Ancient China in transition: an analysis of social mobility, 722-222 B.C.* Stanford, Calif.: Stanford University Press.

Hunter, M. (2014). Did Mencius know the Analects? *T'oung Pao* 100: 33–79. http://doi.org/10.1163/15685322-10013p02

Hutton, E. L., trans. (2014). *Xunzi: the complete text*. Princeton: Princeton University Press.

Ihara, C. K. (1991). David Wong on Emotions in Mencius. *Philosophy East and West* 41: 45.

Ivanhoe, P. J. (2000). *Confucian Moral Self Cultivation* (2nd ed.). Indianapolis/Cambridge MA: Hackett Publishing Company.

Ivanhoe, P. J. (2008). ''The Shade of Confucius: Social Roles, Ethical Theory, and the Self,'' (34-49). R. L. Littlejohn and M. Chandler, eds. In *Polishing the Chinese Mirror: Essays in Honor of Henry Rosemont, Jr*. New York: Global Scholarly Publications.

Jockers, M. L. (2013). *Macroanalysis: digital methods and literary history*. Urbana: University of Illinois Press.

Kern, M. (2015). "Speaking of Poetry: Pattern and Argument in the "Kongzi Shilun"." In Gentz, J. and D. Meyer (eds.), *Literary Forms of Argument in Early China* (175-200). Leiden: Brill.

Kline, T.C. (2000). Moral Agency and Motivation in the Xunzi. In T.C. Kline & P.J. Ivanhoe, eds., *Virtue, Nature and Moral Agency in the Xunzi* (155-175). Cambridge, MA: Hackett Publishing Company.

Lau, D.C. (1970). *Mencius*. New York: Penguin.

Li, Xueqin. (2010-). *Qīnghuá Dàxué cáng Zhànguó zhújiǎn* 清華大學藏戰國竹簡 (Vol. 1-5). Shanghai: Zhongxi Shuju 中西書局.

Liu, Xiaogan. (2003). *Classifying the Zhuangzi Chapters.* Ann Arbor, MI: University of Michigan Center for Chinese Studies.

Loewe, M. (ed.). (1993). *Early Chinese texts: a bibliographical guide*. Berkeley, Calif.: Society for the Study of Early China : Institute of East Asian Studies, University of California, Berkeley.

Makeham, J. (1996). The Formation of Lunyu as a Book. *Monumenta Serica* 44: 1–24.

Marshall, E. A. (2013). Defining population problems: Using topic models for cross-national comparison of disciplinary development. *Poetics* 41: 701–724. http://doi.org/10.1016/j.poetic.2013.08.001

McCallum, A. K. (2002). MAchine Learning for LanguagE Toolkit (MALLET). Stable URL: http://mallet.cs.umass.edu/. Accessed 7-15-2016.

Meeks, E., & Weingart, S. B. (2012). The Digital Humanities Contribution to Topic Modeling. *Journal of Digital Humanities* 2:1. Retrieved from http://journalofdigitalhumanities.org/2-1/dh-contribution-to-topic-modeling/. Accessed 7-15-2016.

Mohr, J. W., & Bogdanov, P. (2013). Introduction—Topic models: What they are and why they matter. *Poetics*, *41*(6), 545–569. http://doi.org/10.1016/j.poetic.2013.10.001

Moretti, F. (2000). Conjectures On World Literature. *New Left Review* 1: 54-68

Nelson, R. (2015, May 28). Mining the Dispatch. Retrieved from http://dsl.richmond.edu/dispatch/pages/home. Accessed 7-15-2016.

Nichols, R. (2015). Early Confucianism is a System for Social-Functional Influence and Probably Does Not Represent a Normative Ethical Theory. *Dao* 14: 499–520. http://doi.org/10.1007/s11712-015-9464-8

Nichols, Ryan. (2011). "A Genealogy of Early Confucian Moral Psychology." *Philosophy East and West* 61: 609–29. doi:10.1353/pew.2011.0057.

Nylan, M. (2001) *The Five "Confucian" Classics*. New Haven: Yale University press.

Qu, Wanli. (1983). *Shijing quanshi* (詩經詮釋). Tapbei: Guojia tushuguan chuban she.

Rhody, L. (2013). Topic Modeling and Figurative Language. *Journal of Digital Humanities* 2(1). Retrieved from http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/. Accessed 7-15-2016.

Rubin, T. N., Chambers, A., Smyth, P., & Steyvers, M. (2012). Statistical topic models for multi-label document classification. *Machine Learning* 88: 157–208.

Schmidt, B. (2012). Words Alone: Dismantling Topic Models in the Humanities. *Journal of Digital Humanities* 2(1). Retrived from http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/. Accessed 7-15-2016.

Slingerland, Edward. (2000). Review: Why Philosophy Is Not "Extra" in Understanding the Analects. [*The Original Analects* by Brooks, E. Bruce; Brooks, A. Taeko]. *Philosophy East and West* 50(1):137-141.

Slingerland, Edward. (2003). *Effortless Action: Wu-wei as Conceptual Metaphor and Spiritual Ideal in Early China*. New York: Oxford University Press.

Slingerland, E., & Chudek, M. (2011). The Prevalence of Mind-Body Dualism in Early China. *Cognitive Science* 35(5): 997–1007. doi:10.1111/j.1551-6709.2011.01186.x

Twitchett, D. & M. Lowe. (1986). *The Cambridge History of China Volume 1: The Ch'in and Han Empires, 221 BC–AD 220.* Cambridge: Cambridge University Press.

Underwood, T. (2012a, April 7). Topic Modeling Made Just Simple Enough. <http://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/>. Accessed 7-15-2016.

Underwood, T. (2012b, April 1). What kinds of 'topics' does topic modeling actually produce? <http://tedunderwood.com/2012/04/01/what-kinds-of-topics-does-topic-modeling-actually-produce/>. Accessed 7-15-2016.

Van Norden, B.W. (2008). *Mengzi: with selections from traditional commentaries*. Indianapolis: Hackett.

Van Norden, B. W.(1992). Mengzi and Xunzi: Two Views of Human Agency. *International Philosophical Quarterly*, *32*(2), 161–184. http://doi.org/10.5840/ipq199232212

Wáng, Xiānshèn 王先慎, ed. 2006. Hán fēi zǐ jí jiě 韓非子集解. Beijing: Zhonghua shuju.

Weingart, S. (2012). *Topic Modeling for Humanists: A Guided Tour*. Retrieved from http://www.scottbot.net/HIAL/?p=19113. Accessed 7-15-2016.

Wills, J. E. (2012). *Mountain of fame: portraits in Chinese history*. Princeton, N.J: Princeton University Press.

Wilson, T. (2014). Spirits and the Soul in Confucian Ritual Discourse. *Journal of Chinese Religions*, *42*(2), 185–212. http://doi.org/10.1179/0737769X14Z.00000000013

Wong, D. B. (1991). Is There a Distinction between Reason and Emotion in Mencius? *Philosophy East and West* 41(1): 31-44. http://doi.org/10.2307/1399716

Xú, Fuguan 徐復觀 (1969) *Zhongguo renxing lun shi: Xian Qin pian [History of Chinese Views on Human Nature: The Pre-Qin Period]*. Taipei: The Commercial Press.

Zhang, L. (2012). *The Concept of Humanity in an Age of Globalization*. Göttingen: V&R Unipress.

N.B. Per *Journal of Asian Studies* policies, should the paper be accepted, these appendices will not appear in print but will be placed online instead.

Appendix 1 Texts, Genres & Dates

Text	**Genre**	**Era**
Analects (論語)	Confucianism (儒家)	WS
Mengzi (孟子)	Confucianism (儒家)	WS
Liji (禮記)	Confucianism (儒家)	WS
Xunzi (荀子)	Confucianism (儒家)	WS
Xiao Jing (孝經)	Confucianism (儒家)	WS
Shuo Yuan (說苑)	Confucianism (儒家)	Han
Chun Qiu Fan Lu (春秋繁露)	Confucianism (儒家)	Han
Han Shi Wai Zhuan (韓詩外傳)	Confucianism (儒家)	Han
Da Dai Li Ji (大戴禮記)	Confucianism (儒家)	Han
Baihutong (白虎通)	Confucianism (儒家)	Han
Xin Shu (新書)	Confucianism (儒家)	Han
Xin Xu (新序)	Confucianism (儒家)	Han
Yangzi Fayan (揚子法言)	Confucianism (儒家)	Han
Zhong Lun (中論)	Confucianism (儒家)	Han
Kongzi Jiayu (孔子家語)	Confucianism (儒家)	Han
Qian Fu Lun (潛夫論)	Confucianism (儒家)	Han
Lunheng (論衡)	Confucianism (儒家)	Han
Tai Xuan Jing (太玄經)	Confucianism (儒家)	Han
Fengsu Tongyi (風俗通義)	Confucianism (儒家)	Han

Kongcongzi (孔叢子)[1]	Confucianism (儒家)	Han
Shen Jian (申鑒)	Confucianism (儒家)	Han
Zhuangzi (莊子)	Daoism (道家)	WS
Dao De Jing (道德經)	Daoism (道家)	WS
Liezi (列子)	Daoism (道家)	Post-Han
He Guan Zi (鶡冠子)	Daoism (道家)	Han
Wenzi (文子)	Daoism (道家)	Han
Wen Shi Zhen Jing (文始真經)	Daoism (道家)	Post-Han
Lie Xian Zhuan (列仙傳)	Daoism (道家)	Post-Han
Yuzi (鬻子)	Daoism (道家)	WS
Heshanggong (河上公)	Daoism (道家)	Han
Hanfeizi (韓非子)	Legalism (法家)	WS
Shang Jun Shu (商君書)	Legalism (法家)	WS
Shen Bu Hai (申不害)	Legalism (法家)	WS
Shenzi (慎子)	Legalism (法家)	WS
Jian Zhu Ke Shu (諫逐客書)	Legalism (法家)	WS
Guanzi (管子)	Legalism (法家)	WS
Mozi (墨子)	Mohism (墨家)	WS
Mo Bian Zhu Xu (墨辯注敘)	Mohism (墨家)	Post-Han
Gongsunlongzi (公孫龍子)	School of Names (名家)	Post-Han

[1] As observed by Kern (2015: 189) "traditionally dated to the late third century BCE but most likely composed only in Eastern Han times, even if including earlier material."

The Art of War (孫子兵法)	School of the Military (兵家)	WS
Wu Zi (吳子)	School of the Military (兵家)	WS
Liu Tao (六韜)	School of the Military (兵家)	WS
Si Ma Fa (司馬法)	School of the Military (兵家)	WS
Wei Liao Zi (尉繚子)	School of the Military (兵家)	Han
Three Strategies (三略)	School of the Military (兵家)	Han
Hai Dao Suan Jing (海島算經)	Mathematics (算書)	Han
The Nine Chapters (九章算術)	Mathematics (算書)	Han
Sunzi Suan Jing (孫子算經)	Mathematics (算書)	Post-Han
Zhou Bi Suan Jing (周髀算經)	Mathematics (算書)	Han
Huainanzi (淮南子)	Miscellaneous Schools (雜家)	Han
Lü Shi Chun Qiu (呂氏春秋)	Miscellaneous Schools (雜家)	WS
Gui Gu Zi (鬼谷子)	Miscellaneous Schools (雜家)	Han
Yin Wen Zi (尹文子)	Miscellaneous Schools (雜家)	WS
Deng Xi Zi (鄧析子)	Miscellaneous Schools (雜家)	WS
Shiji (史記)	Histories (史書)	Han
Chun Qiu Zuo Zhuan (春秋左傳)	Histories (史書)	WS
Lost Book of Zhou (逸周書)	Histories (史書)	WS
Guo Yu (國語)	Histories (史書)	WS
Yanzi Chun Qiu (晏子春秋)	Histories (史書)	WS
Wu Yue Chun Qiu (吳越春秋)	Histories (史書)	Han
Yue Jue Shu (越絕書)	Histories (史書)	Han

Zhan Guo Ce (戰國策)	Histories (史書)	WS
Yan Tie Lun (鹽鐵論)	Histories (史書)	Han
Lie Nü Zhuan (列女傳)	Histories (史書)	Han
Guliang Zhuan (轂梁傳)	Histories (史書)	Han
Gongyang Zhuan (公羊傳)	Histories (史書)	Han
Han Shu (漢書)	Histories (史書)	Han
[Qian] Han Ji ([前]漢紀)	Histories (史書)	Han
Dong Guan Han Ji (東觀漢記)	Histories (史書)	Han
Hou Han Shu (後漢書)	Histories (史書)	Post-Han
Zhushu Jinian (竹書紀年)	Histories (史書)	Han
Mutianzi Zhuan (穆天子傳)	Histories (史書)	WS/Han[2]
Gu San Fen (古三墳)	Histories (史書)	Post-Han
Yandanzi (燕丹子)	Histories (史書)	Post-Han
Xijing Zaji (西京雜記)	Histories (史書)	Post-Han
Book of Poetry (詩經)	Ancient Classics (經典文獻)	Pre-WS
Shang Shu (尚書)	Ancient Classics (經典文獻)	Han
Book of Changes (周易)	Ancient Classics (經典文獻)	WS
The Rites of Zhou (周禮)	Ancient Classics (經典文獻)	WS
Chu Ci (楚辭)	Ancient Classics (經典文獻)	WS
Yili (儀禮)	Ancient Classics (經典文獻)	WS

[2] Text in 6 parts. Parts 1-4 are authentic 350 BCE texts. Part 5 is a post-Han addition. Part 6 is a compilation of WS texts.

Shan Hai Jing (山海經)	Ancient Classics (經典文獻)	Han
Jiaoshi Yilin (焦氏易林)	Ancient Classics (經典文獻)	Han
Jingshi Yizhuan (京氏易傳)	Ancient Classics (經典文獻)	Song (forgery)[3]
Shi Shuo (詩說)	Ancient Classics (經典文獻)	Post-Han
Shuo Wen Jie Zi (說文解字)	Etymology (字書)	Han
Er Ya (爾雅)	Etymology (字書)	WS
Shi Ming (釋名)	Etymology (字書)	Han
Fang Yan (方言)	Etymology (字書)	Han
Ji Jiu Pian (急救篇)	Etymology (字書)	Han
Huangdi Neijing (黃帝內經)	Chinese Medicine (醫學)	Han
Nan Jing (難經)	Chinese Medicine (醫學)	Han
Shang Han Lun (傷寒論)	Chinese Medicine (醫學)	Han
Jinkui Yaolue (金匱要略)	Chinese Medicine (醫學)	Han
Guodian(郭店)	Excavated texts (出土文獻)	WS
Mawangdui (馬王堆)	Excavated texts (出土文獻)	Han

Appendix 2 Example Word Cloud and Survey Given Independent Coders

[3] Probably a Song forgery. Thomas F. Aylward (2007:171) cites Twitchett, et al. (1986:692) *Cambridge History of China, Volume I* as stating that the *Jingshi yizhuan* is not authentic, but was written during the Song dynasty.

Word Cloud for Topic 27

Survey Text

1 Suppose you had to guess what is the theme of this word cloud. What are one to three English words you would use to describe this theme?

2 Please indicate how confident you are about your answer to the previous question by using the slider bar below.

0=Completely Uncertain 7=Completely Certain

0 1 2 3 4 5 6 7

3 Consider the categories below. Please select ALL categories into which you believe the content of this word cloud belongs.

Virtue or Morality Philosophy Religion Military Uncategorizable

Knowledge Fortune or Luck Mysticism Mind Leadership

Politics Body Cosmos

4 Since you selected "Mind" among the previous answers, please select ALL concepts below that represent the content of the word cloud.

Cognition Emotion Belief Rationality Feelings Perception

Judgement Skill Soul Memory

5 Since you selected "Military" among the previous answers, please select ALL concepts below that represent the content of the word cloud.

Victory Government Weaponry State Peace Violence Order War

6 Since you selected "Politics" among the previous answers, please select ALL concepts below that represent the content of the word cloud.

Lord Emperor Statecraft Minister Sage King Law Official

7 Since you selected "Fortune" among the previous answers, please select ALL concepts below that represent the content of the word cloud.

Weather Dates Fate Calendar Law

8 Since you selected "Cosmos" among the previous answers, please select ALL concepts below that represent the content of the word cloud.

Sagehood Ability Seasons Human Benefit World

9 Since you selected "Knowledge" among the previous answers, please select ALL concepts below that represent the content of the word cloud.

Earth Reflection Dao World Human Culture

10 Since you selected "Virtue" among the previous answers, please select ALL concepts below that represent the content of the word cloud.

Speak Order Life Desire Wisdom Goodness Worthy Peace Respect

11 Since you selected "Leadership" among the previous answers, please select ALL concepts below that represent the content of the word cloud.

Royalty Statecraft King Education Family

12 Since you selected "Philosophy" among the previous answers, please select ALL concepts below that represent the content of the word cloud.

Language Emperor Ruism Sage King Qi Mencius Confucius Logic

13 Since you selected "Body" among the previous answers, please select ALL concepts below that represent the content of the word cloud.

Medicine Health Bodily organs Yin Medicine Qi Biology Yang

14 Since you selected "Religion" among the previous answers, please select ALL concepts below that represent the content of the word cloud.

Gods Spirit Sacrifice Religion Heaven Deities

15 Since you selected "Mysticism" among the previous answers, please select ALL concepts below that represent the content of the word cloud.

Divination Ritual Sacrifice Spirit Deities Mourning

Qi Gods

Appendix 3 Traditional and Topic-Classifier Dates for 尚書 *Shang Shu* Chapters[4]

Chapter	**Estimated date of composition**	**Era**	**Topic-classifier era (if applicable)**
堯典 - Canon of Yao	275-175 BCE	Late Warring States/ Early Han	
舜典 - Canon of Shun	275-175 BCE	Late Warring States/ Early Han	
大禹謨 - Counsels of the Great Yu	200 CE	Han*	
皋陶謨 - Counsels of Gao-yao	275-175 BCE	Late Warring States/ Early Han	
益稷 - Yi and Ji	200 CE	Han*	
禹貢 - Tribute of Yu	275-175 BCE	Late Warring States/ Early Han	
甘誓 - Speech at Gan	275-175 BCE	Late Warring States/ Early Han	**Warring States**
五子之歌 - Songs of the Five Sons	200 CE	Han*	
胤征 - Punitive Expedition of Yin	200 CE	Han*	

[4] The estimated dates of composition are based in large part on Nylan's "Textual Authority in the pre-Han and Han." As observed by Nylan, the 25 so-called Archaic Script chapters are "known to be late compilations, dating possible as late as the early fourth century AD, though they contain much earlier material that presumably circulated in the Warring States, Han, and Wei periods in form of oral traditions" (Nylan 2001: 134). Since the latest date of composition or compilation is less than century after the end of the Han and since these chapters are contain a significant amount of earlier material, we have tentatively dated them all to 200 CE. To make it easier to identify these chapters, they have all been marked as "Han*."

湯誓 - Speech of Tang	275-175 BCE	Late Warring States/ Early Han	
仲虺之誥 - Announcement of Zhong-hui	200 CE	Han*	
湯誥 - Announcement of Tang	275-175 BCE	Late Warring States/ Early Han	**Warring States**
伊訓 - Instructions of Yi	200 CE	Han*	
太甲上 - Tai Jia I	200 CE	Han*	
太甲中 - Tai Jia II	200 CE	Han*	
太甲下 - Tai Jia III	200 CE	Han*	
咸有一德 - Common Possession of Pure Virtue	200 CE	Han*	
盤庚上 - Pan Geng I	350-200 BCE	Warring States	
盤庚中 - Pan Geng II	350-200 BCE	Warring States	
盤庚下 - Pan Geng III	350-200 BCE	Warring States	
說命上 - Charge to Yue I	200 CE	Han*	
說命中 - Charge to Yue II	200 CE	Han*	
說命下 - Charge to Yue III	200 CE	Han*	
高宗肜日 - Day of the Supplementary Sacrifice to Gao Zong	1000-700 BCE	Pre-Warring States	
西伯戡黎 - Chief of the west's Conquest of Li	450-250 BCE	Warring States	
微子 - Count of Wei	450-250 BCE	Warring States	**Han**
泰誓上 - Great Declaration I	200 CE	Han*	
泰誓中 - Great Declaration II	200 CE	Han*	

泰誓下 - Great Declaration III	200 CE	Han*	
牧誓 - Speech at Mu	450-250 BCE	Warring States	
武成 - Successful Completion of the War	200 CE	Han*	
洪範 - Great Plan	450-250 BCE	Warring States	**Han**
旅獒 - Hounds of Lu	200 CE	Han*	
金縢 - Metal-bound Coffer	450-250 BCE	Warring States	
大誥 - Great Announcement	1000-700 BCE	Pre-Warring States	
微子之命 - Charge to the Count of Wei	200 CE	Han*	
康誥 - Announcement to the Prince of Kang	1000-700 BCE	Pre-Warring States	
酒誥 - Announcement about Drunkenness	1000-700 BCE	Pre-Warring States	
梓材 - Timber of the Rottlera	800-500 BCE	Pre-Warring States	
召誥 - Announcement of the Duke of Shao	1000-700 BCE	Pre-Warring States	
洛誥 - Announcement concerning Luo	1000-700 BCE	Pre-Warring States	
多士 - Numerous Officers	800-500 BCE	Pre-Warring States	
無逸 - Against Luxurious Ease	800-500 BCE	Pre-Warring States	
君奭 - Prince Shi	450-250 BCE	Warring States	

蔡仲之命 - Charge to Zhong of Cai	200 CE	Han*	
多方 - Numerous Regions	800-500 BCE	Pre-Warring States	
立政 - Establishment of Government	450-250 BCE	Warring States	
周官 - Officers of Zhou	200 CE	Han*	
君陳 - Jun-chen	200 CE	Han*	
顧命 - Testamentary Charge	450-250 BCE	Warring States	
康王之誥 - Announcement of King Kang	450-250 BCE	Warring States	
畢命 - Charge to the Duke of Bi	200 CE	Han*	
君牙 - Jun-ya	200 CE	Han*	
冏命 - Charge to Jiong	200 CE	Han*	
呂刑 - Marquis of Lu on Punishments	450-250 BCE	Warring States	
文侯之命 - Charge to the Marquis Wen	450-250 BCE	Warring States	
費誓 - Speech at Bi	450-250 BCE	Warring States	
秦誓 - Speech of the Marquis of Qin	450-250 BCE	Warring States	

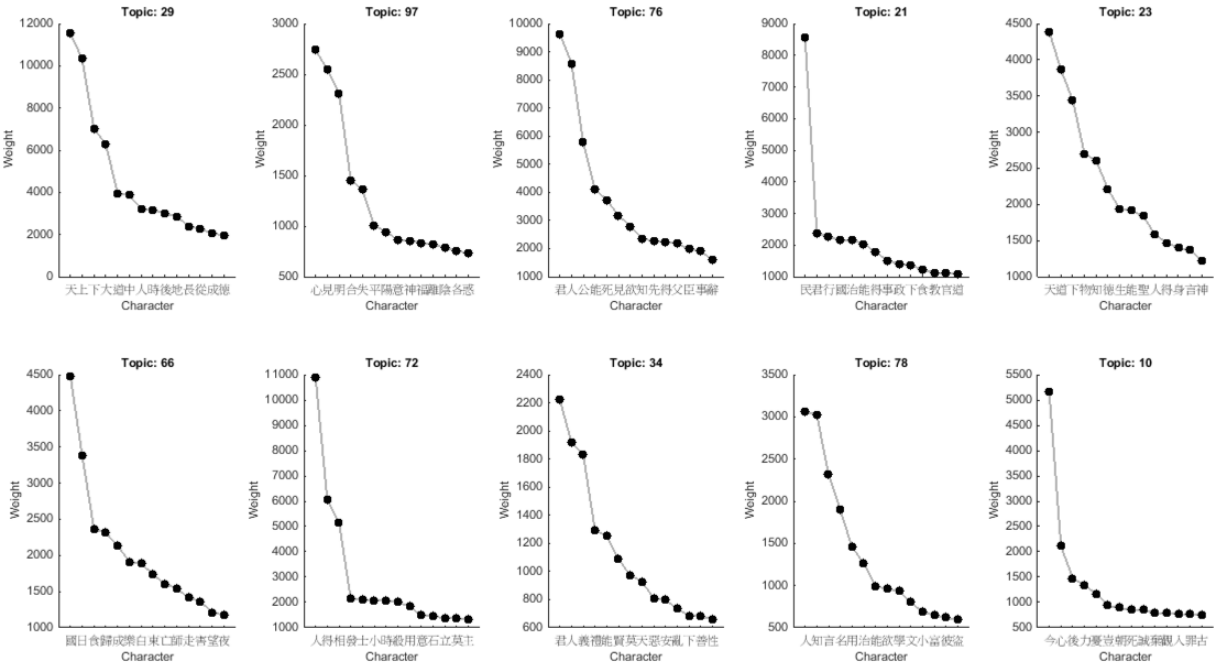
Appendix 4 Stopwords

之	是	于	元	后	哉	還	甚	求	氏	焉
不	與	在	正	作	難	絕	本	說	外	我
也	夫	非	多	因	稱	往	止	左	同	復
以	可	六	西	雖	屬	己	興	起	受	千
而	五	諸	足	始	宜	邪	耳	會	反	亦
其	將	必	又	里	聽	固	廣	定	少	九
為	使	然	高	請	終	首	益	通	常	七
曰	何	若	內	女	遠	由	應	對	過	方
者	至	及	當	右	盡	共	十	所	此	乃
子	四	未	去	敢	異	徒	則	故	太	百
有	矣	萬	北	前	進	任	無	三	謂	皆
於	自	吾	來	易	初	更	一	二	如	乎

Figure 1 Corpus Composition by Era

Era	Dates	Character Count	Percent of Corpus
Pre-Warring States	Before 480 BCE	30,447	0.53
Warring States	479-222 BCE	1,424,080	24.79
Han	221 BCE-220 CE	3,501,256	60.9
Post-Han to Song	221 CE-1044 CE	786,546	13.6
Totals		5,742,329	1.00

Figure 2 Keyword loading in Highest Loading Ten Topics in Our Corpus



Individual plots in this figure represent the distribution of highest loading keywords within the target topic. Characters along the horizontal axis represent central characters in the target topic, with the most central character nearest the vertical axis. The numbers along the vertical axis represent the number of occurrences of each character. Typically keyword weights approximate a discrete power law distribution, with the weights being inversely proportionate to the keyword's rank for any given topic. This describes topic 21 because people (mín 民) has nearly four-times the word weight as Topic 21's second-ranked character, lord (jūn 君). Contrast Topic 23, which is almost linearly distributed. See Figure 3 below.

Fig 3_Heaviest 10 Topics in Corpus

Figure 3 Highest Loading 10 Topics in Corpus

Topic #	Corpus Weight	Topic Name	Topic Keywords in Descending Order of Weight
29	0.600	Heaven, Cosmos, & the Way	天 上 下 大 道 中 人 時 後 地 長 從 成 德
97	0.475	Cognition, Perception & Cosmic Fortune	心 見 明 合 失 平 陽 意 神 福 離 陰 各 惑
76	0.471	Roles of Rulership	君 人 公 能 死 見 欲 知 先 得 父 臣 事 辭
21	0.459	Political & Social Order	民 君 行 國 治 能 得 事 政 下 食 教 官 道
23	0.452	Moral-Cosmic Attunement	天 道 下 物 知 德 生 能 聖 人 得 身 言 神
66	0.446	Ritual Sacrifice	國 日 食 歸 成 樂 白 東 亡 師 走 害 望 夜
72	0.428	Political Roles, Political Affairs	人 得 相 發 士 小 時 殺 用 意 石 立 莫 主
34	0.373	Ethical Rulership	君 人 義 禮 能 賢 莫 天 惡 安 亂 下 善 性
78	0.364	Learning & Governance	人 知 言 名 用 治 能 欲 學 文 小 富 彼 盜
10	0.348	Temporal Cognition & Planning	今 心 後 力 憂 豈 朝 死 誠 棄 觀 入 罪 古

Fig 4_Corpus weights Topics 0-99

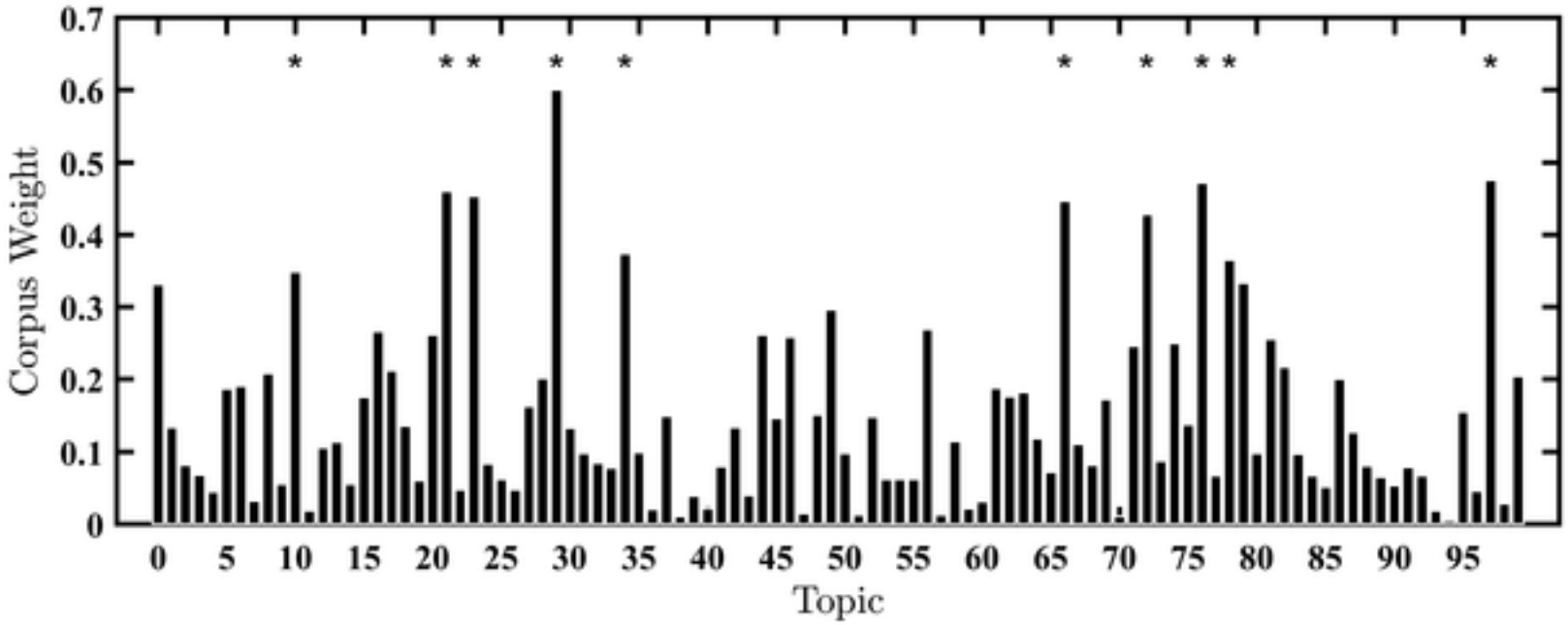


Fig 5_Topic 34 A-M-X

Term	English	*Analects*			*Mengzi*			*Xunzi*		
		Occurrences	Per 1000 characters	Term Rank	Occurrences	Per 1000 characters	Term Rank	Occurrences	Per 1000 characters	Term Rank
君	nobleman	160	21.0	2	253	14.3	5	547	13.5	4
人	person	219	28.7	1	611	34.6	1	1241	30.6	1
義	righteousness	24	3.1	63	107	6.1	25	315	7.8	13
禮	ritual	75	9.8	9	68	3.8	40	343	8.5	10
能	able	69	9.0	12	135	7.6	12	519	12.8	5
賢	virtuous	25	3.3	60	74	4.2	37	152	3.7	44
莫	do not	18	2.4	89	58	3.3	53	257	6.3	18
天	day/heaven	49	6.4	23	293	16.6	4	598	14.7	3
惡	evil	39	5.1	38	80	4.5	36	190	4.7	30
安	peace	17	2.2	94	23	1.3	167	190	4.7	29

Figure 6 Topic 27 Keywords & Weights

Chinese	Pinyin	English	Word Weight
馬	mǎ	horse	0.049
白	bái	white	0.04
物	wù	thing	0.035
生	shēng	to be alive; to give birth to	0.033
汝	rǔ	you	0.031
無	wú	without, nothingness	0.028
見	jiàn	see	0.022
指	zhǐ	finger; denote, point	0.022
色	sè	color	0.019
列	liè	to break up; to rank	0.019

Fig 7_Weightiest 10 topics in A_M_X

Figure 7 Weightiest 10 topics in each of *Analects*, *Mencius* & *Xunzi*

Topic	Keywords	Text Weight in *Analects*
61	孔 問 仁 言 人 禮 行 聞 道 貢	0.307
76	君 人 公 能 死 見 欲 知 先 得	0.130
63	禮 君 人 喪 士 父 樂 母 侯 廟	0.069
78	人 知 言 名 用 治 能 欲 學 文	0.074
21	民 君 行 國 治 能 得 事 政 下	0.040
34	君 人 義 禮 能 賢 莫 天 惡 安	0.043
5	大 祭 食 門 婦 先 入 既 服 出	0.038
33	人 大 天 知 王 得 世 — 心 已	0.026
29	天 上 下 大 道 中 人 時 後 地	0.034
82	公 王 德 成 事 民 告 用 聞 既	0.029
Topic	**Keywords**	**Text Weight in *Mencius***
21	民 君 行 國 治 能 得 事 政 下	0.102
61	孔 問 仁 言 人 禮 行 聞 道 貢	0.121
33	人 大 天 知 王 得 世 — 心 已	0.122
99	王 人 下 孟 取 相 或 士 他 好	0.114
76	君 人 公 能 死 見 欲 知 先 得	0.089
29	天 上 下 大 道 中 人 時 後 地	0.041
18	下 王 詩 天 亡 士 得 侯 善 臣	0.043
86	文 利 學 用 大 古 賢 義 能 今	0.036
63	禮 君 人 喪 士 父 樂 母 侯 廟	0.027
10	今 心 後 力 憂 豈 朝 死 誠 棄	0.027
Topic	**Keywords**	**Text Weight in *Xunzi***
34	君 人 義 禮 能 賢 莫 天 惡 安	0.256
78	人 知 言 名 用 治 能 欲 學 文	0.084
29	天 上 下 大 道 中 人 時 後 地	0.057
71	人 治 事 法 世 行 功 明 主 亂	0.058
21	民 君 行 國 治 能 得 事 政 下	0.053
23	天 道 下 物 知 德 生 能 聖 人	0.052
76	君 人 公 能 死 見 欲 知 先 得	0.038
18	下 王 詩 天 亡 士 得 侯 善 臣	0.039
17	國 法 民 兵 賞 力 利 刑 重 上	0.035
63	禮 君 人 喪 士 父 樂 母 侯 廟	0.025

Fig 8_Topics Differentiating A_M_X

Figure 8 Topics Differentiating *Analects, Mencius* and *Xunzi* from one another

Document	Text Weight	Topic	Corpus Weight	Name	Topic Keywords in Descending Order of Weight
Analects	0.029	82	0.22	Virtuous Rule & Moral Suasion	公 王 德 成 事 民 告 用 聞 既 實 能 先 政
Analects	0.038	5	0.19	Sacrifice, Ritual, Mourning	大祭 食 門 婦 先 入 既 服 出 飲 相 小 哭
Mencius	0.027	10	0.35	Past & Future Cognition	今 心 後 力 憂 豈 朝 死 誠 棄 觀 入 罪 古
Mencius	0.114	99	0.2	Language of *Mencius*	王 人 下 孟 取 相 或 士 他 好 長 舍 章 羊
Mencius	0.036	86	0.2	Benefit & Moral Excellence	文 利 學 用 大 古 賢 義 能 今 國 商 良 富
Xunzi	0.052	23	0.45	Moral-Cosmic Education	天 道 下 物 知 德 生 能 聖 人 得 身 言 神
Xunzi	0.058	71	0.25	Legal Order	人 治 事 法 世 行 功 明 主 亂 亡 得 相 用
Xunzi	0.035	17	0.21	Institutional Rulership	國 法 民 兵 賞 力 利 刑 重 上 勝 官 戰 爵

Fig 9_Topic 5's Text Weights Sheet1

Text	Text Weight of Topic 5
Yílǐ 儀禮	0.175
Lǐjì 禮記	0.17
Zhōulǐ 周禮	0.052
Dàdàilǐjì 大戴禮記	0.04
Analects (Lúnyǔ) 論語	0.038
Báihǔtōngdélùn 白虎通德論	0.034
Mùtiānzǐzhuàn 穆天子傳	0.033
Kǒngzǐjiāyǔ 孔子家語	0.032
Shìmíng 釋名	0.029
Ěryǎ 爾雅	0.027

Fig 10_Text Weights in A/M/X

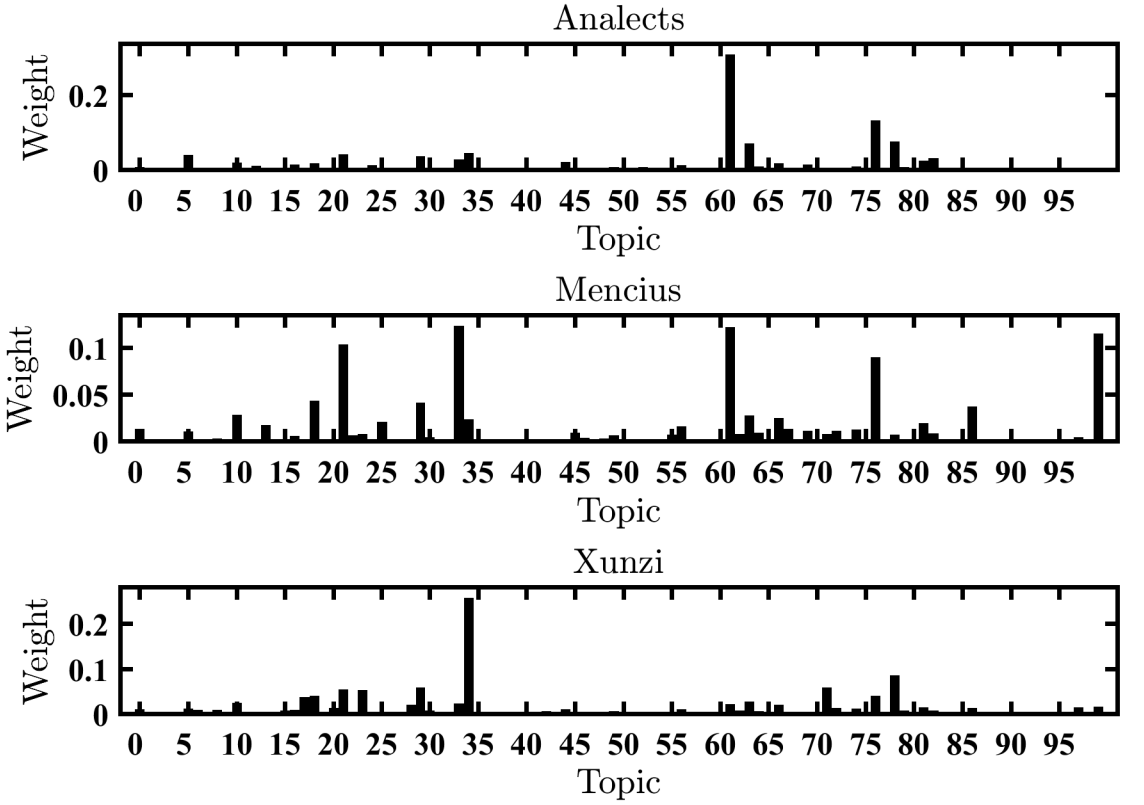
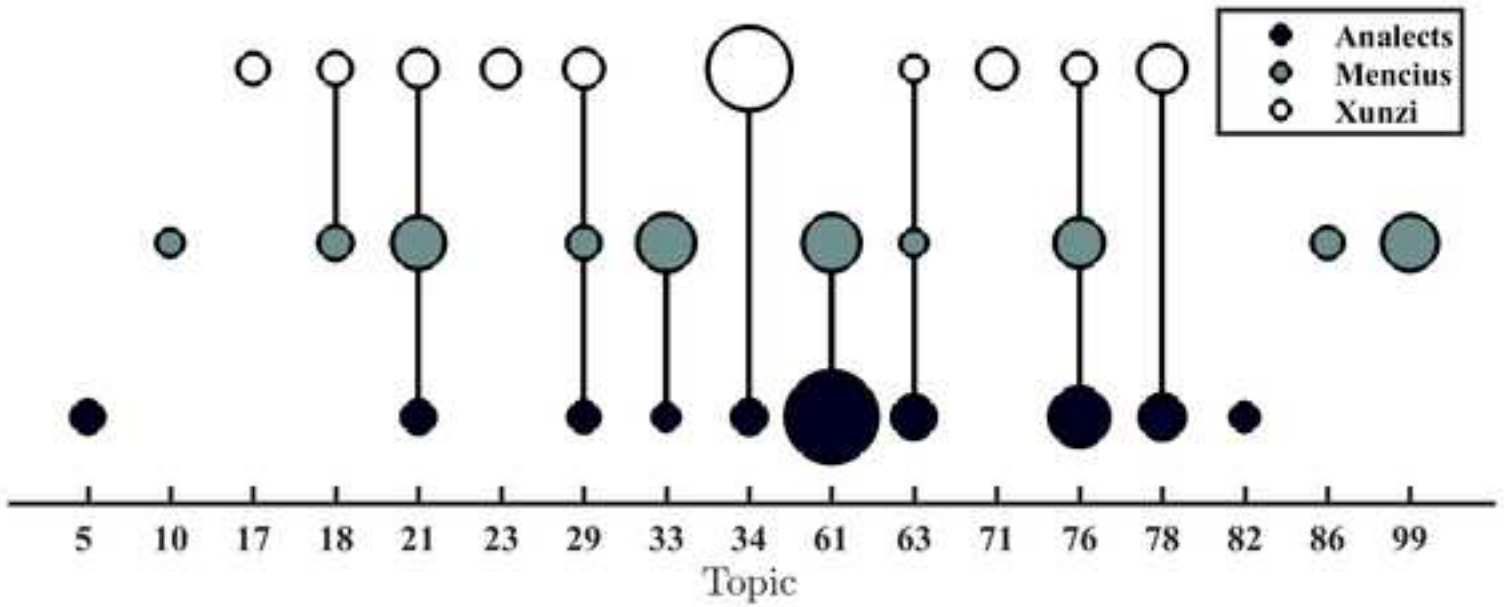


Fig 11_Intersecting Topics at A_M_X

Figure 11 Formal Interpretation Matrix of Intersections of *Analects, Mencius* and *Xunzi* with Topic Keywords (∩=intersection of sets)

Document	Topic's Weight in Text (Text Weight)	Topic	Topic's Weight in Corpus (Corpus Weight)	Name	Topic Keywords in Descending Order of Weight
(Mencius ∩ *Xunzi* ∩ *Analects)*	0.04/0.06/0.03	29	0.6	Heaven & the Way	天 上 下 大 道 中 人 時 後 地 長 從 成 德
(Mencius ∩ *Xunzi* ∩ *Analects)*	0.1/0.05/0.04	21	0.46	Political & Social Order	民 君 行 國 治 能 得 事 政 下 食 教 官 道
(Mencius ∩ *Xunzi* ∩ *Analects)*	0.03/0.03/0.07	63	0.18	Ritual, Family & Governance	禮 君 人 喪 士 父 樂 母 侯 廟 親 主 命 事
Xunzi ∩ *Analects*	0.25/0.04	34	0.37	Ethical Rulership	君 人 義 禮 能 賢 莫 天 惡 安 亂 下 善 性
Xunzi ∩ *Analects*	0.08/0.07	78	0.37	Learning & Governance	人 知 言 名 用 治 能 欲 學 文 小 富 彼 盜
Mencius ∩ *Analects*	0.12/0.3	61	0.19	Language of *Analects*	孔 問 仁 言 人 禮 行 聞 道 貢 仲 學 知 路
Mencius ∩ *Analects*	0.12/0.03	33	0.08	Human, Heaven and Political Order	人 大 天 知 王 得 世 一 心 已 義 且 今 見

Fig 12_topicIntersection



Responses to Referees

Referee 1, General Comments

Overall, I am impressed with the scholarship in this article. Your discussion of the utility of topic models for the study of early and medieval Chinese literature is convincing. You have clearly done a good job addressing some of the problems involved in using topic models for humanistic research and have made good efforts to avoid bias. The first three sections, which I view as three parts of a single whole, are very good and only need some refinement. Primarily, I would like to see **greater engagement with the few extant pieces of scholarship.** I know that you address a different question than previous scholarship by focusing on Mencius, Xunzi, and the Analects, but how do you situate yourself vis-a-vis this earlier work? **Are you extending or refining the use of topic models in applying them to your question?** Further, as JAS is aimed at a general Asian studies audience, you should **elaborate on the possible impact your work** will have outside the narrow field in which it engages.	We have added reference to more topic modeling papers--both papers that introduce topic modeling to humanists and papers that use topic modeling to advance issues in humanities. Thanks to Referee 1's comments, we have also taken effort to more clearly situate what we are aiming to do with our topic model. We expended more effort to add to our use of secondary literature about Chinese thought to provide context to the discussion of contents of topics. For example, with respect to Topic 5 we set its contents in the context of a debate about the role of sacrifice to gods and spirits in *Analects*, citing Wilson (2014) on one side and Feng Yu Lan (1953) on the other. While we do add a few remarks about the overall significance of topic modeling at the beginning of the paper, we believe traditional scholars will continue to express resistance to use of machine-learning to supplement traditional scholarship. This is why we believe that further elaboration about the possible impact of this work, or topic modeling in general, on Asian Studies at large to be speculative.
In some cases, your argument could use some clarifications, and I really think you should **include more figures that visualize your topic models**. One of the great things about topic modeling is its capacity to be visualized in clear and intuitive manners. Take more advantage of this! The points I believe need clarification are included in my detailed comments below.	We took this suggestion very seriously and have added a host of additional figures and tables to the paper.
I am much more **hesitant about section 4**, for a number of reasons. First, I believe **it could**	After reading these remarks, and holding some discussion amongst ourselves, we

be dropped from the article with no impact to the first three sections. If you would like to retain it, I think there needs to be a strong justification for its inclusion. The primary connection, continued use of topic modeling, is not enough. As it is not used in this section as an interpretive device, but instead as a method of dimensionality reduction in preparation for a classifier, I see these as completely different methodologies. There also needs to be more reference to machine learning scholarship. Why use neural nets and not some other classification algorithm? As classification is very new to Asian Studies scholarship, make sure to provide extensive citations to relevant literature.	resolved to remove the section about dating and the *Shangshu* from the paper. This not only streamlined the paper, but doing so opened up the opportunity for a short additional publication about that issue and freed up some word count to respond to other concerns of Referees 1 and 2. We evidently needed to hear this point from a referee as insightful as Referee 1 to overcome our own attachment to that section. Thank you for the observation. Note that several subsequent remarks by Referees 1 and 2 are addressed with 'moot'. This is because those comments presumed the inclusion of Section 4 about dating the *Shangshu*.
More critically, I do not think the results of your model can be interpreted in the way in which you describe. Your topic-document based classifier performs admirably. If this article were an exploration of how to build an era-classification model with decent performance, I would go along with it. However, as the conclusion is that 4 chapters are reclassified into adjacent periods, which is a demonstration of how classification can lead to new insights, I have to disagree. Your classifier shows 93 percent accuracy, with good precision and recall. In the end, you classify all 58 chapters of the Shang shu, and 4 chapters are predicted to belong to different eras. This is exactly what you should expect from a model with 93 percent accuracy: 4 misclassifications (54/58 being 93 percent). While it may be true that some chapters are misclassified by traditional scholarship, your model does not offer interpretable results; those four chapters just represent the error in the model.	Moot
Also, you should extend your discussion of how you tested the model in the first place (you mention a 70/30 train/test split, but not if	Moot

you run this multiple times with random variation, or if you have included your new labels in testing the model).	
There are various ways to address these issues of course, though most of them would involve significant research. I would recommend extending your training set and seeing if you still get the same changing classifications (I know you mention the possibility/difficulties of extending the training set in your article). Also, what happens if you re-label these four chapters with your new classification and then rerun your model? Does the model continue to perform at the same level? Do other chapters now get reclassified? Also, consider ways of increasing the performance of your model. Over-fitting is a danger, but don't let worry of it handicap your model, as extensive cross-validation should help prevent it. You might also consider randomly selecting various subsets of the topics and rerunning the model.	Moot
I also realize that some of my criticisms may be answered if I could see the code you use to implement your models, but keep in mind that **most JAS readers will not be interested in digging in to the code**. The text of the article should make these things clear.	Understood. To address this and meet some criticism from Referee 1, we greatly expanded, and slowed down, our introduction, in particular, our remarks about the topic modeling process.
Finally, realize that a likely criticism some traditionally minded scholars will have is that you are **too hard on traditional methods**. You mention that you do not see this as replacing older methods, but that only comes at the very end of the article in the last sentence. The role of topic modeling as a supplement to traditional research should be emphasized much earlier. Try to make your discussion of the biases inherent in both traditional and digital scholarship a bit more nuanced.	This is an astute observation that anticipated some concerns from Referee 2. In response, we have altered the tone of our discussion of traditional methods, deleted some content about the contrasts between methods, and intentionally adopted a metaphor about using machine-learning techniques to begin a new thread in a conversation with traditional methods. This established that we see the two methods as mostly complementary and only occasionally competitive.

Referee 1, Specific Comments

Page 5. **Define "loading."** Most JAS readers will not know the connotations of this term. Consider adding it in a footnote.	The re-written manuscript eschews that term in order to consistently use the term “weight.” This term is helpful because it

	can easily be used across the different types of data output by a topic model, including word (or character) weight, text weight, and topic weight. Though we felt that inclusion of some technical discussion is necessary, we have moved all of that discussion to footnotes. Note that the other examples of topic modeling in Asian Studies do not include much technical information. Not coincidentally, they are fraught with serious technical problems that render subsequent interpretation seriously misguided. We refrained from calling these papers out in our submission. Nonetheless, we insist in including (highly optional) technical information in the footnotes.
Page 8: "Appearances were deceiving..." Consider **removing this and the following sentence**.	Resolved.
"... we enlisted the help of over 60 experts in classical Chinese thought and language to perform independent code topics." Rephrase the last part to something like "to independently code topics." This is an excellent idea! I would also move your later sentence "We refer to these results periodically... up to just after this sentence. I would **consider rephrasing it to drop mention of the reader.** Just state that it was an analytical step taken to validate our interpretations and avoid bias. **No need to mention that it is done to assuage reader concerns**. This is a good thing to do in an article espousing the use of topic models, but I don't think it should be something that readers should consider necessary to do if they want to use topic models.	Resolved. We clarify that the principal reason for which we solicited others' expertise, even though three of the six authors on the paper are themselves published authors in the area of pre-Qin Chinese thought, is to reduce our own interpretive biases.
Page 10: ... to suggest how methods such as topic modeling..." could be **shortened** to "to suggest how topic modeling..."	Resolved.
Page 11: In the paragraph that starts "this debate is unlikely to be resolved..." I have a few hesitations. The machine generated	This is a great point. In our excitement, we overemphasized ways in which our data is new or different.

topic modeling represents a transformation of data that is already available. While this transformation fundamentally alters the nature of the data, I would be **very hesitant to call it a "new source of data."** You should also change the end of the paragraph as well "...which does not suffer from traditionally scholarly methods or biases" isn't exactly true. You could frame it as "does not suffer from the same biases as traditional scholarly methods," but be very clear that it still suffers from biases. **Corpus selection** represents a big one. **Stopword selection** is another huge one. You also briefly mention Ted Underwood's and Andrew Goldstone's assessment of topic modeling earlier in the paper. Bear in mind that many JAS readers will be approaching topic modeling with a skeptical eye. Make sure to **be as nuanced as you can when you discuss possible biases**.	We have also clarified ways that topic modeling suffers from its own biases in a new paragraph that details human decisionmaking in the preprocessing, processing and interpretive phases of a topic modeling project like ours.
The paragraph "Topics discussed in this section..." should be extended a bit. Whenever you get into **technical terminology**, be aware that it **may trip up some readers**, so be as clear as you can. Secondly, make sure to **be explicit in your selection criteria**. What about their probability distributions leads you to discuss them? Do you deal with just the top three topics in each document? Topics with a text weight above .03? If it changes between the three texts, how and why?	This is another excellent point from Referee 1, and it is made in an inquisitive, open-ended way we appreciated. We have clarified our choice-making about topics for analysis and about the number of topics in our model (100). In short, in Section 3, we discuss topics that set each of the three documents (*Analects*, *Mencius*, *Xunzi,* A/M/X) apart from the other two documents. In Section 4, we discuss topics that lie at the intersections of those three documents; these are topics that they share. Together, this two-pronged approach allows us to see what they have in common and how they differ. To make this more obvious, we have included a couple additional figures, for example, one with the top 10 topics in each of these documents.
Pages 12-13: I think in its current form, **this article misses a few good opportunities for visualizations**. You make several claims when discussing **Topic 5** that would	Among the half-dozen new visualizations for this resubmission is one tailored to Topic 5 per Referee 2's comment here. This was a helpful suggestion that adds rhetorical

be further elucidated in a figure that visualizes topic 5's text weight in each text across the whole corpus. This way you can show the reader, rather than just tell him or her, that topic 5 loads highest into the Analects, Zhouli, Liji, and Yili. This would make the argument that "unsupervised topic modeling is able to precisely track this scholarly insight..." much stronger.	force to our argument.
Page 13: The first mention of **frequency ranks** occurs halfway down this page (and recurs through the article). If you are going to use this as an interpretative framework, **introduce it sooner and make it clear what you are doing**. Jumping back and forth from topics/their weights to straight frequency can be a bit hard to follow. As it stands, it is easy to miss in the midst of the paragraph. Also, **be careful in conflating word rank (frequency) with "importance."** This makes sense to me, but make sure you justify this move. You should also be careful, as the length of a text can influence these rankings significantly. It might make more sense to **refer to occurrences per thousand characters** (as it is not outside the realm of possibility that the 50th and 100th most common words are only a small number of occurrences apart, due to Zipf's law). As you remove stop words, this danger gets even higher.	We have created a visualization that zooms in on character-level data in a topic. Columns in this chart represent word weight of keywords in the topic, occurrences in A/M/X, occurrences/1000 characters in A/M/X and word ranks in A/M/X. This and several other revisions and new visualizations are designed to allow the humanist reader to look under the hood and at minimum identify all the major parts (=types of data) output by the topic model.
Page 14: The last sentence begins and ends with the same basic point. **Consider cutting** one or the other	Resolved.
Page 19: Go in to more detail in **how you construct the intersections** between these works. Do you use a weight threshold? As all topics load into all documents (though many with near zero weight), you should be more specific. Are you looking at the top ten topics per document and then discussing the topics that are common between the works? This is the impression I get from later in the paper, but it should be more explicit here. This might be another area in	We have gone into some more detail on this semi-technical point, again, in a footnote. We also have created a new visualization (about which two of us were genuinely excited) to represent intersections for the non-technical reader. Note that several types of our visualizations are novel, so far as we know, to the world of topic modeling. Note also that at the beginning of Sections 3 and 4 we have clarified exactly what topics

which a figure would be useful.	are in play and why.
Page 21: The crux of your argument seems to be that the Xunzi and Analects share more of their top ten topics than the Mencius and the Analects. In the cases where there is homology between Mencius and Analect, the Xunzi is also fairly similar (but just below your threshold for intersection?). In the few cases where Xunzi does not contain much of a topic shared by the Mencius and Analects, it is in topics related to style rather than semantics. I had to re-read this paragraph a few times to get this, so **consider rewriting this section**. It also might be pertinent to look at the intersections of the Xunzi and the Mencius.	We have rewritten this so make the train of our reasoning much more transparent to the reader. I'm afraid that the length of the paper (after deletion of the *Shangshu* section but addition of much more content) prevents of us from including new data about intersections of *Mencius* and *Xunzi*.
Page 22: "second, as **primarily stylistic**, 61 and 33..." This seems more important than its billing in this paragraph suggests.	On further consideration, we have withdrawn our remark about 33 being primarily stylistic. 61, however, remains primarily stylistic in our opinion. We explain both 33 and 61 in greater detail in the resubmission.
"In most cases our data confirm scholarly consensus, specifically, via the following topics." This **sentence seems out of place**.	Resolved.
Page 24: "We reverted to a version of the Shang Shu that included all stopwords." Excellent! Stopwords play a critical role in most text classification tasks (both era and authorship attribution).	Moot. While most of the remaining comments did not factor into our revision of the submission, since we deleted this section, we want to thank the Referee for such helpful and thoughtful comments about that section.
"A feedforward neural network..." There needs to be a more extended discussion of why neural nets are the appropriate choice here (rather than SVM/kNN/Naive Bayes or some other method). Was this choice empirically determined and chosen because it worked the best? Given that these approaches are very new in Asian Studies, extended engagement can only help. Make sure to also engage with the appropriate literature! I also wonder, is computational efficiency a concern on a 58 document problem?	Moot.

Page 25: "A term-document matrix..." This is another instance where you will have to be very clear. You do describe what a tdm is, but make sure to also address what vectorization is.	Moot.
"Shang shu has 1834 characters..." This should be "unique characters"	Moot.
"To increase the accuracy of the model and cope with redundancy..." I know that using topic models is the main connective tissue in the article, but I find myself wondering if you tried other methods of dimensionality reduction or feature selection (PCA or maybe a logit regression to select discriminating features)? Topic modeling is fine, but you should justify this choice with reference to the literature (consider works like "Authorship Attribution with Latent Dirichlet Allocation" by Yanir Seroussi, et. al). You are obviously working on era not authorship, but they are similar classification problems.	Moot.
Why G-measure and not F-score? I've mostly encountered F-score in the literature. G-measure is fine, I was just wondering what rationale led to this choice.	Moot.
"To further compare the topic-based classifier's performance, we also trained a similar model on the Shang Shu topic-document matrix, which serves as a baseline for comparison." Do you mean term-document in the second instance?	Moot.
Page 26: Did you calculate the performance metric over a single run? Or did you average multiple runs with randomly selected 70/30 training/testing sets?	Moot.
"Our dating work on the Shang Shu..." How is your classifier trained here? Are you training it on 57 chapters and testing it against the remaining chapter? Why are you confident that with 93 percent accuracy that these 4 chapters are not just the misclassifications that show that your model is 93 percent accurate?	Moot.

To what extent can you confirm your model is predicting era rather than some other latent characteristic?	Moot.
Page 27: "Notice, however, that the classifier never jumps two categories..." With only 4 errors, could this just be random chance?	Moot.
Page 28: "More important to us, we hope that our preliminary distant reading..." This is an important sentence and one that needs more time! Most of the conflict you will encounter with this piece will be from scholars thinking you want to throw out traditional research techniques in favor of this one. Make it clear early on that this is not your intention. Leaving it until the last sentence of the article makes it seem a little underwhelming.	
Figures: I would consider adding several figures to the topic modeling discussion, in addition to the single one I mentioned above. First, **visualize the topic distribution for all three texts** you are dealing with. Also, **visualize the topic weights across the entire corpus** and label the more prominent topics. Finally, consider a figure that **visualizes your intersecting topics in each of the texts**. You touch on all of these things, but presenting them in a figure would increase their impact.	We have generated all three visualizations requested by Referee 1 here. We have added a chart that visualizes the topic distribution for topics with high topic weights in A/M/X. We have added a new visualization of topic weights across the corpus, highlighting high corpus weight topics in it. As mentioned above, we have also generated an exciting new visualization depicting the intersection of topics between A/M/X.

Referee 2

This is an ambitious piece and it presents a groundbreaking methodology. However, it is **far too technical for the audience of JAS**. While the presentation of topic modeling is straightforward, I found the **discussion of neural networks much harder to follow**.	As we have deleted this section on the advice of Referees 1 and 2, this comment is now moot.
The discussion of particular topic models could be confusing, as it **required constant**	It is customary in science-related papers to append figures at the end of the document.

reference to the appended topic results.	However, given Referee 1's comment here, we have taken liberty to include low-resolution version--of the sort MS Word permits us to save within a file--in the body of our resubmission. This appears to be in violation of JAS's online submission management system. However, *please note that these low-res versions are strictly to ease Referee 2's reading of the MS*. We have high-res versions saved separately and uploaded through the submission portal that are of publication-quality as instructed. As a result, please keep in mind that *our manuscript's length is actually much shorter than it will appear*.
There was far **too little evidence of reading in the secondary scholarship** on early Chinese texts, which could have helped situate the particular assumptions / hypotheses and results and show why the topic model results were interesting. In particular, I found the **command of the traditional scholarship to be lacking**, thus undermining the persuasive force of the new methodology. Very **little secondary scholarship was cited for early thought** (and Brooks / Brooks is a particularly troubling work to cite in the absence of other, more mainstream scholarship). It seems to me that the authors **do not know the field of early Chinese thought and history particularly well,** and	Though we bristled at this remark, it is understandable at one level and we have taken pains to include apt references to secondary literature, especially in the context of discussion of the contexts of our topics. However, we would add the following. Contrary to the implication of Referee 2, we did not cite Brooks and Brooks "in the absence of other, more mainstream scholarship". As to the composition and dating of Analects, we also cited in the earlier submission Cheng (1993, 313-23), Makeham (1996), Qu Wanli (1983, 382-9) and others. We do not understand what Referee 1 could mean by this remark. Regarding Referee 2's comment that our command of the traditional scholarship was lacking and that we did not know the field of early Chinese thought and history particularly well, we are incredulous at this short-sighted comment. We invite the editor to judge that for himself. We do understand, however, the utility of adding material that will assist in guiding the our colleagues through our discussion of the contents of topics; we have taken steps to add to references to this literature demonstrative of the utility of topic modeling.

	However, absence of evidence in this case was mistaken for evidence of absence. We do not feel a need to weigh down a paper that aims to be at the cutting edge of Asian Studies scholarship (and that is already at the word limit for JAS) with padded references intended solely to establish our credibility.
I would hesitate to recommend this piece of publication, since it will likely fall upon **deaf ears** (i.e., folk who will not see the worth of computational approaches when the basic domain expertise is not in evidence).	About three years ago we identified JAS as the first place we wanted to seek publication precisely because of what this paper represents. We believe that in years to come topic modeling will continue to be of considerable and growing interest to a wide base of researchers in the Asian Studies community, especially new PhD students applying tech-saavy techniques to study classic texts. Further, we hope that the steps we have taken to confirm in the minds of readers of this (anonymous) resubmission our credibility in the field were sufficient.
More technical questions: what does it mean to topic model texts of very different scales (Lun yu vs. Han shu)? How does that impact the model? How are texts **being chunked or processed?** Again, the Lun yu would have very small chunks but the Han shu?	Our new version of Section 2 "Topic Modeling" walks the humanities reader through the process more slowly and in greater, but non-technical, detail. In answer to the questions, the length of the texts makes no difference to the topic models or the composition of topics. Topics are produced from the entire corpus rather than strictly from singular texts. (We can use the text weight of topics to understand how each topic is distributed into various texts. In the resubmission, we walk through this in more detail too.) We thought about chunking the corpus, but not for this project. Preprocessing texts for topic modeling (a process we describe in more detail in Section 2 of the resubmission) does not involve chunking. Preprocessing for our texts was light.
Choices made by the authors would presumably impact the results. **There is some lack of transparency**, also, in regard	In the submission we were explicit about the role of our survey of 60 experts in the field. Our survey experts are not credited

to the interpretive subjectivity of the labeling of topics. **It feels sometimes that the labels were given and then assumed to be vetted by domain experts**, thus stabilizing their meaning. I don't think that this is the case; the labels remain quite ambiguous.	with the English labels for topics; those represent a necessary heuristic with which to efficiently talk about the thematize the topics. We created those labels ourselves. Our independent experts, of course, never saw those English labels. Readers are given the words in each topic and are free to agree or disagree with our labels, just as, in more traditional scholarship, authors provide a primary text and an interpretation, with readers being free to disagree. One aspect in which we do think more transparency is warranted is with regard to the overall distribution of topics among the various texts in our corpus. We can only provide selected statistics in the body of the paper, and we could imagine readers wanting to be able to independently assess our claims about the distribution of these topics. Upon publication we will be happy to include additional materials online including our various data, including our survey data. We expected no less.
In general, I don't think that this piece is written for a general scholarly audience, and it might be a better fit for a DH journal instead. This is interesting work, and I would like to see more of this in the field, but I think that the authors need to take into account their audience and write accordingly.	Contrary to Referee 2's comment, this paper is not suitable for a general digital humanities journal. The paper is also not written for that audience, which is already very familiar with topic modeling and its applications. This paper is directed at a wide range of Asian Studies scholars, as we hope our many revisions make clear. This group includes Asian Studies scholars with an interest in understanding ancient and medieval Chinese thought; in new methods of research; in gaining skills to enhance their own research; Confucianism; and in improving their understanding of the influence of early Chinese thought on the geographical and cultural area on which they work.

Acknowledgements

This project received financial support from a Canadian Social Sciences and Humanities Research Council (SSHRC) Partnership Grant to University of British Columbia's Centre for Human Evolution, Cognition and Culture and its Cultural Evolution of Religion Consortium; The Chinese Challenge, a Templeton World Charities Foundation (TWCF) project, for financial resources and to Justin Barrett for the vision to support this new research methodology; the College of Humanities and Social Sciences, California State University, Fullerton; University of North Carolina at Chapel Hill; and the Interacting Minds Centre, Aarhus University. Part of this research was performed while Nielbo was visiting the Institute for Pure and Applied Mathematics (IPAM), which is supported by the National Science Foundation; part of the research was performed while Slingerland was an Andrew W. Mellon Foundation Fellow at the Center for Advanced Study in the Behavioral Sciences (CASBS), Stanford University; part of the research was performed while Nichols was a fellow at UBC's Centre for Human Evolution, Cognition and Culture. We thank these institutions. We thank the many experts in ancient Chinese thought and language who replied to our questionnaire on our topics, and to Steve Angle at *Warp, Weft and Way* for allowing us to post our request for expert coders at the blog. We give special thanks to Robban Toleno and Brenton Sullivan for help, and to Donald Sturgeon for permission to use his CTP online textual corpus.