Journal of Literature and Art Studies, March 2019, Vol. 9, No. 3, 317-328
doi: 10.17265/2159-5836/2019.03.006

Inheritance and Development of Three Pre-Qin Classics of Confucianism—An Application of Topic Modeling in Classical Chinese Text Analysis

HU Jia-jia
Beijing Normal University, Beijing, China

The *Analects*, *Mengzi* and *Xunzi* are the top-three classical works of pre-Qin Confucianism, which epitomized thoughts and ideas of Confucius, Mencius and XunKuang[1]. There have been lots of spirited and in-depth discussions on their ideological inheritance and development from all kinds of academics. This paper tries to cast a new light on these discussions through "machine reading[2]".

Keywords: pre-Qin Confucianism, the *Analects*, *Mengzi*, *Xunzi*, text analysis, machine reading, topic modeling, Mallet, Gephi

Introduction

Topic modeling provides a suit of algorithms to discover hidden thematic structure in large collections of unlabeled texts. It has been increasingly used to analyze massive text data from internet pages, new media and social net, like document classification and clustering, hot event detection and tracking, opinion mining and so on. As a Chinese information processing tool, topic modeling has been used to analyze modern Chinese texts. However, there are large volumes of ancient classical Chinese works that have been precious wealth of Chinese culture. Generations of researchers put decades of their life into "direct and close reading" of these classical works. What if some "distant reading" methods were used on these works? Will these methods bring some new insights? This paper is such a try to discuss the practicability and prospect of one of these "distant reading" a.k.a "machine reading" methods in the analysis of classical Chinese texts.

This paper has four parts. First, it introduces a statistical natural language processing approach to explore the contents of large volumes of unlabeled text, known informally as a topic model. Second, it shows how to use Mallet[3] to discover hidden thematic structure in the top-three classical works of pre-Qin Confucianism. Third, it tries to make a reasonable interpretation of the result topic model. Finally, it discusses prospects of using topic modeling in the research of classical Chinese texts.

HU Jia-jia, Ph.D., Research Center for Folklore, Classics and Chinese Characters, Beijing Normal University, Beijing, China.

[1] Mengzi and Xunzi can be refered as the name of books as well as the name of authors. In order to distinguish, this paper uses Mencius and XunKuang to refer to the authors.

[2] N. Katherine Haylyes, How We Think: Digital Media and Contemporary Technogenesis (Chicago: University of Chicago Press, 2012), 55-80.

[3] McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu. 2002.

What Is a Topic Model

A topic model is a type of statistical model to find groups of words that frequently occur together in a large corpus of texts—groups of words that readers felt comfortable calling topics. For example, if there is a topic which consists of following terms “apple, samsung, huawei, android, ios”, a reasonable guess is that this topic is about “smart phone”. Strictly speaking, these topics are probability distributions over the unique words of the corpus; those words to which the distributions assign the highest probability are those associated with the topic.

There are a family of probabilistic algorithms developed to discover hidden topics in a collection of documents. Most of them are based on an assumption of a simple probabilistic procedure by which documents can be generated. To make a new document, one chooses a distribution over topics. Then, for each word in that document, one chooses a topic at random according to this distribution, and draws a word from that topic. This is known as a generative model for documents. It bases upon the idea that documents are mixtures of topics, where a topic consists of a cluster of high frequent co-occurrence words. The generative process described here does not make any assumption about the order of words as they appear in documents. The only information relevant to the model is the number of times words are produced. This is the well-known bag-of-words assumption.

For brevity, this paper writes $P(z)$ for the distribution over topics z in a particular document and $P(w|z)$ for the probability distribution over words w given topic z. Each word w_i in a document (where the index refers to the i^{th} word token) is generated by first sampling a topic from the topic distribution, then choosing a word from the topic-word distribution. $P(z_i = j)$ is used as the probability that the j^{th} topic was sampled for the i^{th} word token and $P(w_i|z_i = j)$ as the probability of word w_i under topic j. Then generative model specifies the following distribution over words within a document $P(w_i) = \sum_{j=1}^{T} P(w_i \mid z_i = j) P(z_i = j)$, where T is the number of topics. To simplify notation, let $P(w)=P(w|z)P(z)$ where $P(w)$ refers to the multinomial distribution over words for a collection of documents.

Given the observed words distributions in a set of documents i.e. $P(w)$, standard statistical techniques can be used to invert this process, inferring what topic model is most likely to have generated the data. This involves inferring the probability distribution over words associated with each topic i.e. $P(w|z)$, the distribution over topics for each document i.e. $P(z)$, and, often, the topic responsible for generating each word i.e. z.

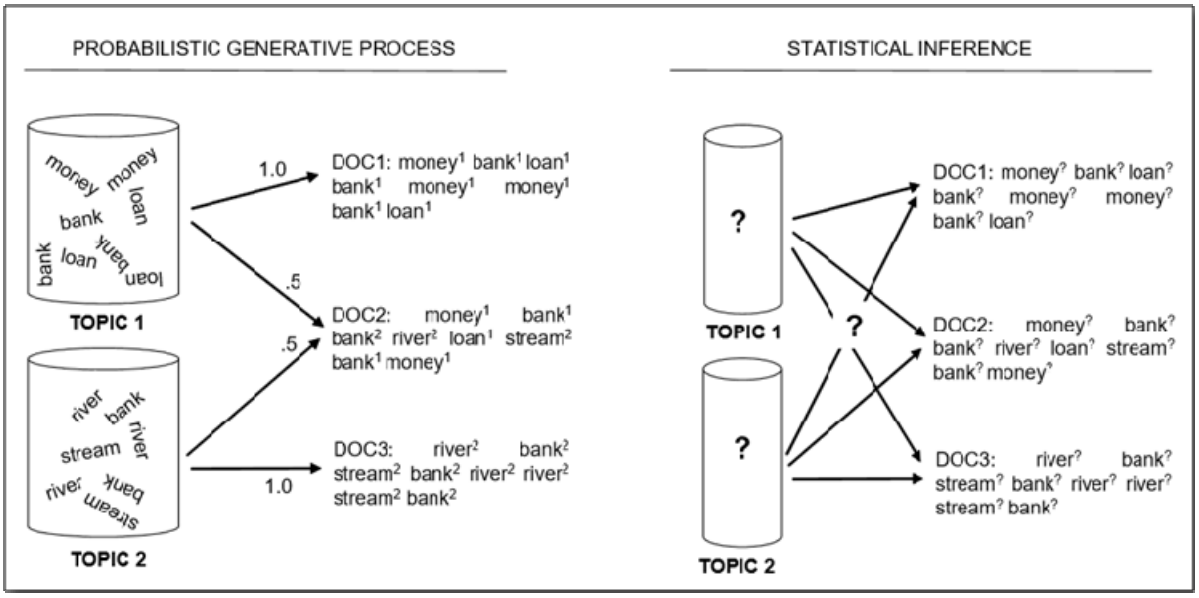

Figure 1. Illustration of the generative process and the problem of statistical inference underlying topic modeling.[4]

Figure 1 illustrates the generative process of three documents under two topics and the statistical inferring process. Therefore, each document is related to one or more topics and a word may appear in more than one topic. Using contextual clues, topic models can connect words with similar meanings and distinguish between uses of words with multiple meanings (like "bank" in topic 1 and topic 2).

As a digital humanities researcher, one does not need to know every detail of the statistic inferring process. There are many statistical Natural Language Processing software packages with topic modeling tools, such as Paper Machines[5], Tethne[6] and Mallet[7]. This paper uses the Mallet. "Mallet (MAchine Learning for LanguagE Toolkit) is from the University of Massachusetts Amherst (UMass Amherst) for a Natural Language Processing JAVA software package developed for text classification and clustering and topic modeling, information extraction and other machine learning to text.[8]"

How to Use Mallet to Build a Topic Model of Classical Chinese Texts

Plain Texts

The topic modeling tool in Mallet deals with plain text (.txt) files. This paper acquires the top-three classical works of pre-Qin Confucianism from the "Chinese Text Project (CTEXT)[9]". With over thirty thousand titles and more than five billion characters, CTEXT is credited with providing accurate and accessible copies of ancient (in particular pre-Qin and Han dynasty) Chinese texts in an organized and searchable format. "The Chinese Text Project Application Programming Interface (CTP API) provides methods for integrating content and

[4] Steyvers, M., & Griffiths, T. Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), Latent semantic analysis: A road to meaning. Hillsdale, NJ: Laurence Erlbaum. 2007.
[5] http://papermachines.org.
[6] https://github.com/diging/tethne.
[7] http://mallet.cs.umass.edu/.
[8] McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu. 2002.
[9] http://ctext.org/.

functionality of the CTP with other web-based or web-aware sites and applications.[10]" This paper used its "gettext[11]" function to obtain texts needed through their "CTP urns", which are unique identifiers describing textual items such as books or parts of books.

Chinese Word Segmentation

As discussed above, the input data of a topic modeling algorithm is the observed words distributions in a set of documents $P(w)$. In English texts, words are distinguished by spaces naturally. However, there are no natural boundaries between words in Chinese texts. Therefore, it is necessary to split sequences of consecutive Chinese characters into sequences of Chinese words by spaces before topic modeling. This process can be implemented in computers known as "Chinese word segmentation" in the field of Chinese information processing. It is generally known that pre-Qin classical Chinese texts mainly composed of monosyllabic words and every monosyllabic word is a single Chinese character. So, unlike modern Chinese, the word segmentation of pre-Qin classical Chinese texts only needs a special dictionary to identify polysyllabic names of people, places, official titles and other name entities firstly, and then treats every single character as a monosyllabic word. The special dictionary and word segmentation algorithm used in this paper is provided by Dr. Donald Sturgeon[12]. Table 1 shows part texts of the *Analects*, which have been removed all punctuation marks and segmented into words by spaces.

Table 1

Segmentation Results in Classical Chinese Texts

孟懿子問孝子曰無違樊遲禦子告之曰孟孫問孝於我我對曰無違樊遲曰何謂也子曰生事之以禮死葬之以禮祭之以禮
孟武伯問孝子曰父母唯其疾之憂
子遊問孝子曰今之孝者是謂能養至於犬馬皆能有養不敬何以別乎
子夏問孝子曰色難有事弟子服其勞有酒食先生饌曾是以為孝乎

Comparative Documents

Topic modeling is a statistical method, which requires large numbers of documents to produce useful results. These documents should be comparable objects. This means they should have similar length and similar origins so that some similar word use can be found in these documents. What is more, individual documents should be long enough for meaningful statistics, that is to say, they should be long enough to have "about-ness". In analysis of modern Chinese, typical minimum 100 words are required per document. Considering classical Chinese works are more refined than modern Chinese, the required length of each document can be properly reduced.

There are 503 chapters in the *Analects*, 260 chapters in *Mengzi*, 596 paragraphs in *Xunzi*. Considering chapters in the *Analects* are usually shorter than those ones in *Mengzi* or *Xunzi*, this paper treats every three consecutive chapters in the *Analects* as a single document, every chapter in *Mengzi* as a single document, and every paragraph in *Xunzi* as a single document. As a result, there are totally 1,024 documents which are filed in a folder named *lmx* as the input data of topic modeling.

---

[10] http://ctext.org/tools/api.

[11] http://ctext.org/plugins/apilist/#gettext.

[12] Dr. Donald Sturgeon is the creator and administrator of the Chinese Text Project. He is currently based at Harvard University as a postdoctoral fellow at the Fairbank Center for Chinese Studies and lecturer at the Department of East Asian Languages and Civilizations. The special dictionary and word segmentation algorithm was got from his course at Harvard University "Digital Methods in Chinese Study". There is also a similar course on the website http://digitalsinology.org/.

Stop Words List

In topic modeling, there are some most frequent words which may appear in many topics but provide little information, like "子曰" in the *Analects*. Mallet allows users to customize a list of stop words that they want to completely ignore. None of stop words in the list will be assigned to any topics in the process of modeling. There is no single "perfect list" of stop words. To make a proper list of stop words, this paper firstly calculates word frequency in the three classical Chinese works as Table 2 shows. Referring to this table, a list of stop words is customized to exclude high-frequency words that provide little meaningful information in topic modeling.

Table 2

High-frequency Words in the Analects, Mengzi and Xunzi

Words	The *Analects*	*Mengzi*	*Xunzi*
曰	0.05046	0.028557	0.007161
之	0.040128	0.055714	0.053567
不	0.038861	0.03219	0.033261
也	0.035462	0.036686	0.036869
子	0.034529	0.005449	0.001599
而	0.022064	0.021768	0.032222
其	0.017931	0.01748	0.01633
者	0.014598	0.018939	0.022397
以	0.014065	0.018939	0.021345
有	0.012198	0.013698	0.010112
矣	0.012065	0.007504	0.00742
……	……	……	……

Command Lines

Mallet is run by command lines. The current version is mallet-2.0.8[13]. Shawn Graham, Scott Weingart, and Ian Milligan have written an excellent tutorial on Mallet topic modeling[14]. The command lines used in this paper to build a topic model are listed in Table 3.

Table 3

Command Lines Used to Build a Topic Model of Three Pre-Qin Confiucianism Works

bin\mallet import-dir --input **lmx** --output **lmx.mallet** --keep-sequence --token-regex "[\p{L}\p{M}]+" --extra-stopwords**lmx-stopwords.txt** --remove-stopwords bin\mallet train-topics --input **lmx.mallet** --num-topics **10** --output-topic-keys **lmx-topic-keys.txt** --output-doc-topics **lmx-doc-topic.txt** --word-topic-counts-file **lmx-word-topic.txt** --optimize-interval 10

The first command counts words distributions over documents in folder lmx without considering the stop words listed in lmx-stopworsd.txt. The result is stored in a mallet readable file lmx.mallet, which is also the input of the second command. The second command uses the words distributions $P(w)$ as input, and outputs three text files. (1) The topic-keys file (lmx-topic-keys.txt) lists topics repeatedly occur in the set of documents, as Table 4 shows. Each topic is represented by a serial number followed by its proportion in the set of documents and a group of co-occurrence key words that describe what the topic is about. (2) The doc-topics file (lmx-doc-topic.txt) describes how documents and topics are related i.e. the distribution over topics for each document $P(z)$ as Table 5

[13] http://mallet.cs.umass.edu/download.php.
[14] http://programminghistorian.org/lessons/topic-modeling-and-mallet.

shows. (3) The word-topic-counts-file (lmx-word-topic.txt) shows how topics and words are related, i.e. the probability distribution over words associated with each topic $P(w|z)$ as Table 6 shows.

Table 4

10 Topics of the Top-three Classical Works of Pre-Qin Confucianism

Topic No.	Possibilities	Key words
0	0.18696	君人臣諸侯大夫朝國事天子公命禮士賢敢受友位出見
1	0.25889	人國天下利亡貴危欲義信好仁爭富愛脩常彊民埶
2	0.1509	水民善獸取山馬地詩教木生田作火天下政射深海
3	0.15128	天下舜民堯湯禹道天世紂桀文王君歸治己人周公武行
4	0.28661	道事明知心功君物賢天治成主聽亂誠德蔽易官
5	0.19536	食父親長事人母孝養弟兄老妻居衣富飲耕身貴
6	0.6192	人言君子知仁行欲見道惡善心聞好學求士義志小人
7	0.1312	樂禮文生聲死養情飾鼓憂喪色章玉祭目衣終和
8	0.22628	王民刑齊服事行殺兵政威臣罪敵國死教楚強地
9	0.35319	人禮義亂知法性治君子生名王理分道情正異惡辨

Table 5

Proportion of 10 Topics in Each Document of the Top-three Classical Works of Pre-Qin Confucianism

	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
lunyu1.txt	0.004848	0.058581	0.003913	0.003923	0.0593	0.082867	0.690334	0.003403	0.005868	0.086961
lunyu2.txt	0.078574	0.302242	0.00372	0.028385	0.007066	0.152746	0.187851	0.003235	0.128854	0.107327
lunyu3.txt	0.049079	0.00581	0.003386	0.003395	0.185966	0.071709	0.552497	0.002944	0.117287	0.007926
lunyu4.txt	0.004393	0.006083	0.003546	0.003554	0.006734	0.216057	0.484475	0.003083	0.005317	0.266758
lunyu5.txt	0.003557	0.062003	0.021897	0.002878	0.005453	0.136898	0.62061	0.135678	0.004305	0.00672
lunyu6.txt	0.008672	0.012008	0.099764	0.007017	0.152441	0.009061	0.631692	0.006085	0.010496	0.062764
lunyu7.txt	0.004196	0.00581	0.003386	0.003395	0.006432	0.049268	0.552497	0.182478	0.117287	0.075251
lunyu8.txt	0.006118	0.008471	0.004938	0.00495	0.009379	0.660846	0.052985	0.167907	0.07285	0.011557
lunyu9.txt	0.007612	0.010541	0.046861	0.00616	0.093103	0.007954	0.595248	0.005342	0.009214	0.217965
……	……	……	……	……	……	……	……	……	……	……

Table 6

Word Counts in 10 Topics of the Top-three Classical Works of Pre-Qin Confucianism

Word No.	Word	Topic: word counts
0	學	6:152[15] 9:21 2:15
1	習	9:18 6:1
2	朋	5:16 7:1
3	遠	6:59 4:44 2:15 3:4 7:1 5:1
4	來	6:21 8:20 5:20 0:8 9:2
5	樂	7:217 6:42 1:23 3:5
6	人	6:735 9:371 1:285 0:152 5:80 3:43 4:38 2:16
7	知	6:320 9:173 4:143 1:42 5:10 7:4
8	慍	6:5 3:1
9	君子	6:349 9:136 7:1

[15] 6:152 means that the word 學 appeared 152 times in topic 6.

Choice of Topic Number

In most topic modeling methods, users should choose the number of topics. The choice can affect the interpretability of the results. A solution with too few topics will generally result in very broad topics whereas a solution with too many topics will result in uninterpretable topics that pick out idiosyncratic word combinations.

At the first run of topic modeling, this paper sets the parameter of num-topics, which defines the number of topics in the second command line, as 10. This is the default value of num-topics in Mallet. The question is how to know it is an appropriate choice. This paper tries to use relations of key words among different topics to find the proper number of topics.

As shown in Table 4, a keyword may appear in more than one topics (like "天下" and "君子"). There may be two reasons. One is that the word is a polysemy. The other one is that the topics, which share one or more same words, are related. Therefore, if viewing each key word as a node that has a direct link to its topic, which can also be viewed as a node, there would be a network among topics linked with their key words. Figure 2 is such a network among the 10 topics listed in Table 4. It is drawn by the network analysis and visualization software package Gephi[16]. The topic network in Figure 2 is drawn in a force-based layout. Typically, spring-like attractive forces are used to attract pairs of nodes of the graph's edges towards each other, while simultaneously repulsive forces like those of electrically charged particles are used to separate all pairs of nodes. As a result, nodes are pushed away from other nodes but edges are kept as short as possible.

[16] https://gephi.org/.

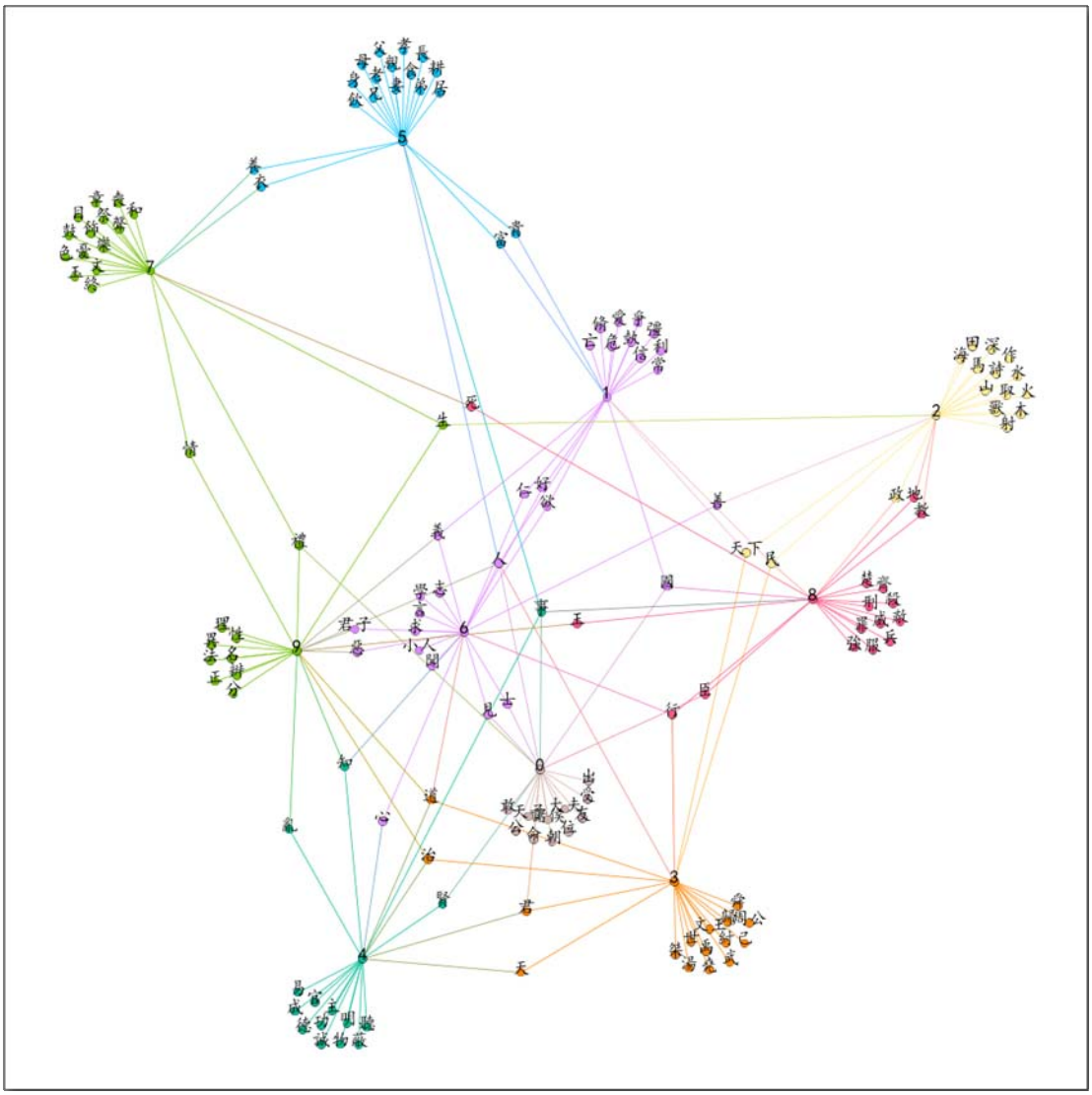

Figure 2. The topic network of 10 topics.

Gephi provides a benefit function called “modularity” to measure the quality of a division of a network into communities. Nodes in a community tend to connect more to other nodes in the same community. In Gephi, different communities can be marked with different colors, as Figure 2 shows. Intuitively, each topic and its key works should be a single community, that is to say, the internal relationships of atopic are closer than its outer relationships with other topics.

In Figure 2, the topic network consisting of 10 topics is divided into 8 communities, which may give a reference for the choice of topic number. This paper uses the same input data but sets the parameter of num-topics as 8 to build a new topic model of the top-three classical works of pre-Qin Confucianism. Table 7 lists the result of 8 topics and their key words. Figure 3 shows the topic network of these 8 topics drawn in Gephi with a force-based layout and divided into 8 communities by modularity function.

Table 7

8 Topics of the Top-three Classical Works of Pre-Qin Confucianism

Topic No.	Possibilities	Key words
0	0.34582	人國天下事利道危明賢主君義刑民行功貴信亡治
1	0.20329	王民國齊仁兵政地天下楚敵取水教服樂殺強利守
2	0.5967	人言知君子仁道行欲見心善惡學求好義聞小人過己
3	0.22353	君人臣諸侯命禮朝大夫受友天子事敢士國公服出位賢
4	0.13845	天下舜民堯君湯天禹桀紂賢世行聖周公歸文王武海人
5	0.38866	禮人亂治道知法性義名王生君子天正明成分物理
6	0.23248	食人父親長事孝母生弟兄死敬身養妻獸失居愛
7	0.1232	樂養心生聲蔽知氣天目物動情口色形文鼓飾美

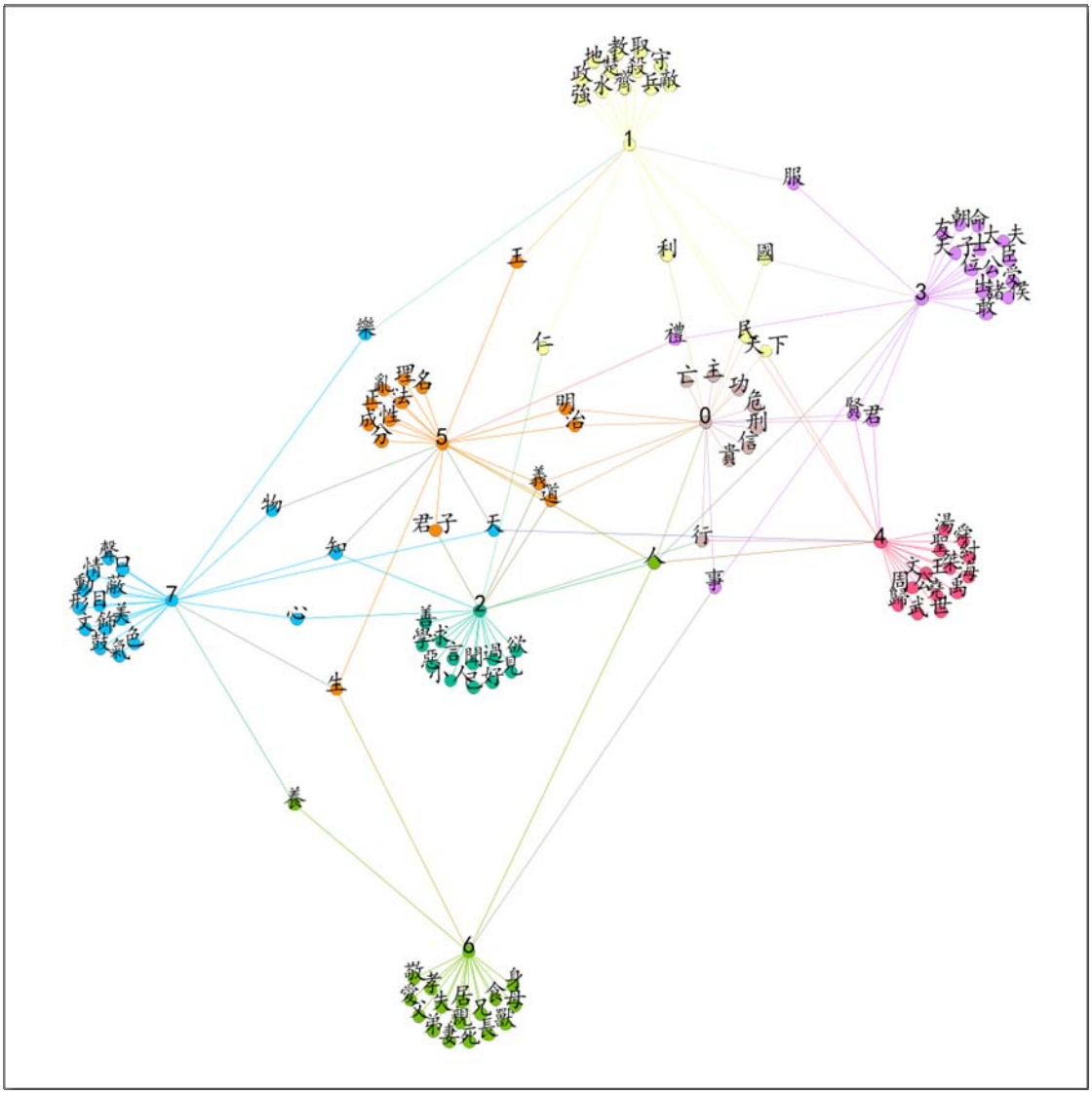

Figure 3. The topic network of 8 topics.

How to Explain the Results of Topic Modeling

In topic modeling, topics are not assigned based on meaning, they are purely a function of statistical word use. What is meaningful depends on interpretation. So this paper firstly gives each topic a name, see Table 8. Table 9 shows the distribution over topics for each book according to the doc-topics file (lmx-doc-topic.txt). For a better explanation, the order of 8 topics is changed and the data is illustrated, see Figure 4. It is interesting to find the data is consistent to some conclusions of the study on Pre-Qin ideology and culture.

Table 8

Interpretation of 8 Topics of the Top-three Classical Works of Pre-Qin Confucianism

Topic No.	Possibilities	Key words	Topic name
0	0.34582	人國天下事利道危明賢主君義刑民行功貴信亡治	刑賞治亂 the Doctrine ofRewards and Punishments
1	0.20329	王民國齊仁兵政地天下楚敵取水教服樂殺強利守	王道 Kingcraft
2	0.5967	人言知君子仁道行欲見心善惡學求好義聞小人過己	修身 Morality Cultivation
3	0.22353	君人臣諸侯命禮朝大夫受友天子事敢士國公服出位賢	禮制 the Order of Ritual
4	0.13845	天下舜民堯君湯天禹桀紂賢世行聖周公歸文王武海人	聖王 Sage Monarch
5	0.38866	禮人亂治道知法性義名王生君子天正明成分物理	禮法 **Law and Rite**
6	0.23248	食人父親長事孝母生弟兄死敬身養妻獸失居愛	齊家 Family Responsibilities
7	0.1232	樂養心生聲蔽知氣天目物動情口色形文鼓飾美	儀節 Etiquettes

Table 9

Proportion of 8 Topics in the Top-three Classical Works of Pre-Qin Confucianism

	Morality cultivation	Family responsibilities	The order of ritual	Sage monarch	Kingcraft	Law and rite	The doctrine of rewards and punishments	Etiquettes
The *Analects*	0.489453	0.096942	0.146219	0.037501	0.046617	0.090655	0.060782	0.03183
Mengzi	0.280071	0.150299	0.111612	0.106658	0.162123	0.081292	0.072423	0.035522
Xunzi	0.189727	0.073885	0.085206	0.050338	0.07193	0.239358	0.216453	0.073103

First of all, personal morality cultivation is the foundation of Confucianism. The related topic takes up a large proportion in all the three classical works, especially in the *Analects*. In the ideological system of Confucius, individual morality underlies the governance of society and politics.

The ideological differences of Confucius, Mencius, and XunKuang are reflected in their political thoughts. Confucius and Mencius tried to achieve and maintain a good social environment by keeping and restoring the order of ritual, which underlies the patriarchal clan system and the enfeoffment in early Zhou Dynasty. Whereas, in XunKuang's opinion, the law is as important as the rite. Mencius put more attentions on the discussion of ancient sage monarch and kingcraft, while XunKuang relied on rewards and punishments to quell disorder and strengthen authority. What is more, their different political opinions rooted from different times of their lives. In

pace with increasing disputes in warring states, how to build a centralized empire became more and more concerned.

All of the three classical works put little attention on the topic of "etiquettes (details to practice a rite)", which reflects the distinguish between Rite (禮 Li) and Etiquette (儀 Yi) by Confucianism. From the late spring and autumn period, the focus of Confucianism had been on the nature of rite and its relationship with the ruling system.

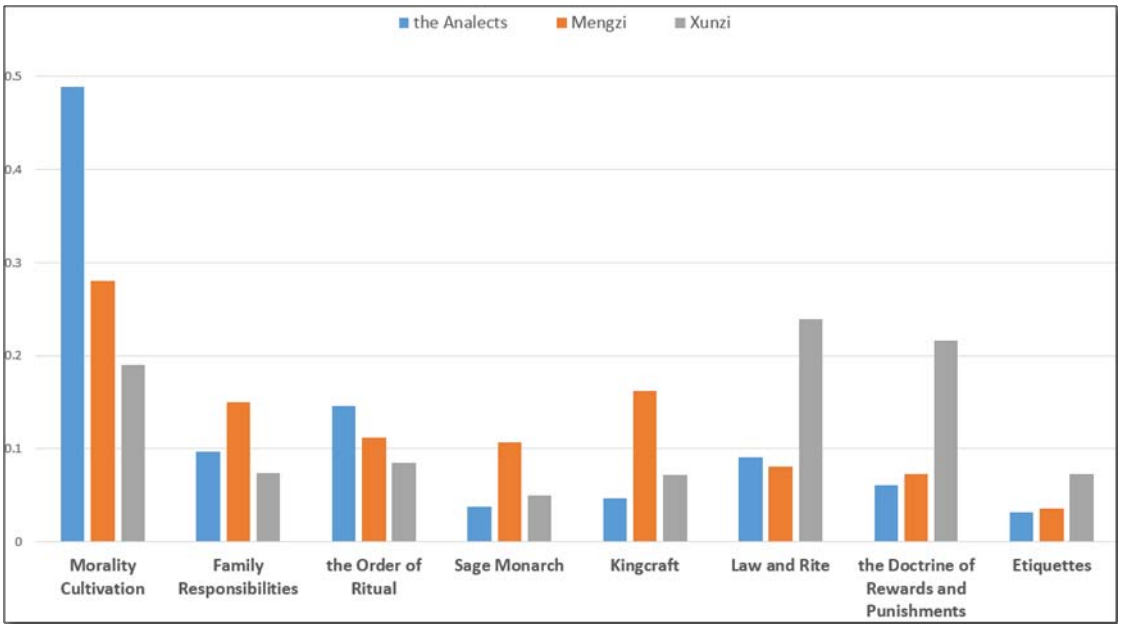

Figure 4. Proportion of 8 topics in the top-three classical works of pre-Qin Confucianism.

Conclusion

Topic modeling is a kind of machine learning. For most users, its inner statistical process is in a black box, which may lead to a lack of confidence. This paper uses mallet to build a topic model of the top-three classical works of pre-Qin Confucianism and gets some objective data. From author's point of view, these data indeed make some sense. Of course, different researchers may have different opinions. However, usually what we can't explain or what we can't achieve agreement on may just those questions need our more research.

As we have proved proper use of topic modeling in classical Chinese works can produce some reasonable data, we would like to use it to do some real "distant reading" which can tell us things about more text than we can read. The first step is to use topic modeling to analyze classical works of Miscellaneous Schools such as "Lv Shi Chun Qiu" and "Huainanzi". The Miscellaneous Schools claimed their works had adopted doctrines of various schools like Confucianism, Mohism, Daoism, Legalism, School of Names, and so on. Topic modeling may tell us how much influence these schools have on the thought of Miscellaneous Schools.

References

Brett, M. R. (2012). Topic modeling: A basic introduction. *Journal of Digital Humanities, 2*, 1.

Graham, S., Weingart, S., & Milligan, I. (2012). Getting started with topic modeling and MALLET. *Programming Historian, 2*, 12. http://programminghistorian.org/lessons/topic-modeling-and-mallet

McCallum, A. K. (2002). MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu

Riddell, A. B. (2012). How to read 22198 journal articles: Studying the history of German studies with topic models. https://ariddell.org/static/how-to-read-n-articles.pdf

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, and W. Kintsch (Eds.), *Latent semantic analysis: A road to meaning* (pp. 1-15). Hillsdale, NJ: Laurence Erlbaum.