CA Journal of Cultural Analytics

October 12, 2017

Topic Modeling the *Hàn diǎn* Ancient Classics (汉典古籍)

Colin ALLEN[a,b], Hongliang LUO[c], Jaimie MURDOCK[b], Jianghuai PU[a], Xiaohong WANG[a], Yanjie ZHAI[a], Kun ZHAO[c]

a Department of Philosophy, School of Humanities and Social Sciences, Xi'an Jiaotong University, Shaanxi, China

b Cognitive Science Program, Indiana University, Bloomington, Indiana, USA

c Institute of Computer Software and Theory, School of Electronic and Information Engineering, Xi'an Jiaotong University, Shaanxi, China

Authors listed alphabetically

ARTICLE INFO

Peer-Reviewed By: Paul Vierthaler

Article DOI: 10.22148/001c.11882

Dataverse DOI: 10.7910/DVN/3QXX29

Journal ISSN: 2371-4549

ABSTRACT

There is a small but growing literature on large-scale statistical modeling of Chinese language texts. Ouyang analyzed a corpus of over 40,000 ancient documents downloaded from multiple sources. This was used to plot the temporal distributions of word frequencies and geographic distributions of authors. Huang and Yu modeled the SongCi poetry corpus, first converting it to tonally marked pinyin to conserve poetically important pronunciation information. Nichols and colleagues reported initial modeling of the Chinese Text Project corpus in a conference paper. (Further below, we describe differences between this corpus and the Handian.) With additional collaborators, this group has now conducted two studies that are currently unpublished but under review. In the first, they apply topic models to address scholarly questions about the relationships among important texts of Ancient Chinese philosophy. In the second, they use topic modeling to investigate the concepts of mind and body in ancient Chinese philosophy. Although we share similar scholarly objectives with these researchers, our approach in this paper is unique in that for the first time anywhere we bring the benefits of computational modeling of ancient Chinese texts to a robust public platform that is mirrored on both sides of the Pacific. Besides being just a useful portal to the texts, our approach foregrounds the interpretive issues surrounding topic models, and makes more sophisticated exploration and analysis of interpretive questions possible for experts and novices alike.

There is a small but growing literature on large-scale statistical modeling of Chinese language texts. Ouyang analyzed a corpus of over 40,000 ancient documents downloaded from multiple sources. This was used to plot the temporal distributions of word frequencies and geographic distributions of authors.[1] Huang and Yu modeled the SongCi poetry corpus, first converting it to tonally marked pinyin to conserve poetically important pronunciation information.[2] Nichols and colleagues reported initial modeling of the Chinese Text Project corpus[3] in a conference paper. (Further below, we describe differences between this corpus and the Handian.) With additional collaborators, this group has now conducted two studies that are currently unpublished but under review. In the first, they apply topic models to address scholarly questions about the relationships among

important texts of Ancient Chinese philosophy. In the second, they use topic modeling to investigate the concepts of mind and body in ancient Chinese philosophy.[4] Although we share similar scholarly objectives with these researchers, our approach in this paper is unique in that for the first time anywhere we bring the benefits of computational modeling of ancient Chinese texts to a robust public platform that is mirrored on both sides of the Pacific. Besides being just a useful portal to the texts, our approach foregrounds the interpretive issues surrounding topic models,[5] and makes more sophisticated exploration and analysis of interpretive questions possible for experts and novices alike.

The Chinese language presents interesting challenges for humanities computing. Both modern and ancient Chinese, but especially the latter, rely heavily on context for the interpretation of individual characters and words[6] and some researchers have argued that differences in Chinese morphology make some of the techniques that work well for DH work in Western languages less applicable to Chinese.[7] Words in Chinese are highly polysemous, requiring considerable amounts of context for their proper interpretation. The study of ancient Chinese philosophy is especially challenging because this ambiguity and openness to multiple interpretation seems to be deliberately exploited by the ancient masters.[8] Take, for example the character '道' which could refer to Taoism, but has up to 10 meanings in ancient Chinese texts, such as 'way' or 'road', and is also used as a verb to mean 'say'. At the same time, the long and relatively continuous history of the Chinese nation has enabled the transmission of a rich corpus of ancient texts to the present day. Computational modeling of these texts does not, as we see it, aim to remove the *human* from the *human*ities. Rather, by enabling the discovery and quantitative analysis of connections, computational methods promise at least these two benefits: (i) enhanced means of access to large sets of documents, and (ii) new sources of evidence about texts that can support the ongoing discussion of their interpretation relative to the past and the present. We are also interested in a more general theme (iii), concerning the potential broader significance for theoretical discussions of the nature of meaning and the role of language in conceptual schemes.

Our primary contribution in this paper is of type (i), to provide enhanced access to a corpus of ancient Chinese documents. Specifically, we introduce an application of the InPhO Topic Explorer[9] developed at Indiana University,

Bloomington, USA, to a large, public corpus of ancient Chinese texts, resulting from collaboration with philosophers and computer scientists at Xi'an Jiaotong University, Shaanxi, China. We also discuss potential projects and future research of type (ii) concerning the analysis of the themes in ancient Chinese philosophy and other literary sources. We present a very brief discussion of the broader significance (iii) before the conclusions section of this paper.

Selecting and Preparing the Corpus

A good understanding of Chinese intellectual culture during the classical period is important in itself, and essential for understanding the reception of Western ideas during various stages of China's history, and vice versa. As philosophers, we are particularly interested in philosophical texts, but we recognize that the boundaries between philosophy and other areas such as religion and political theory are fuzzy at best, and practically non-existent in some cultures or during certain periods of history. Thus, rather than try to demarcate "philosophy" from the rest, we decided to pursue our computational inquiry with as broad a corpus as we could locate.

A secondary consideration is that we want our work to provide a public benefit by being accessible to scholars and the public. It is less than optimal to analyze sources that only a few people—not even all scholars—have access to. For example, although the Wenyuange Edition of the Siku Quanshu archive[10] is of high quality for scholars, it is accessible only to those with subscriptions that are locked to specific IP addresses. Thus we conducted a scan of repositories of ancient Chinese documents, and found that the crowd-sourced website at zdic.net provided the best combination of quantity and access to a large number of classic texts, thanks to its permissive re-use policy under a Creative Commons 1.0 Public Domain Dedication.[11] The full website at www.zdic.net contains a dictionary of Chinese characters, a dictionary of words, dictionary of idioms and several other resources. Among them is the collection of classics identified as 汉典古籍 (*Hàn diǎn gǔ jí*) or Chinese classics—the portion we refer to as the "Handian" corpus — directly accessible at http://gj.zdic.net/, and it is this portion of the website that we chose to model. This section of the website is not without problems, however. It contains a diverse collection of different file formats, containing both traditional and simplified characters, and of varying quality because they have been crowd-sourced from many different users using many different sources, with varying degrees of scholarly care. A better-curated corpus is the Chinese Text

Project (ctext) used by Nichols, Slingerland and colleagues.[12] Although this site can be downloaded for private and academic use, its re-use policy is not as permissive as the Handian, and the online analysis tools require a subscription. Furthermore, because ctext.org is registered in Panama and hosted in the USA as well as directed towards English-speaking users, access by users in mainland China is generally slower and more difficult than zdic.net, which is registered and hosted in China. A third option that was unknown to us when we started this project is the Kanseki Repository (kanripo) hosted at Kyoto University (and on GitHub).[13] Like ctext, kanripo has high quality curation, but unlike ctext it is made freely available using a Creative Commons license. In the future we intend to build a topic explorer around this corpus.

For our initial goals, the benefits of accessibility, especially to Chinese users, outweighed the concerns about corpus curation quality. Such concerns are also partly mitigated by the topic modeling methods (described in more detail below). Because topic models treat documents as unordered "bags of words", they are relatively robust in the face of the "noise" provided by the variable quality of the texts. The techniques we describe here can be applied to more scholarly editions of the same texts. By demonstrating the power of the approach with the Handian corpus, we hope to encourage curators of scholarly editions to incorporate similar methods and make their efforts publicly available. We have made the products of our research available for all at our Indiana University website in the USA, mirrored at the Xi'an Jiaotong University website in China.[14]

In November of 2016 we crawled and downloaded the four sections of the Han classics from the gj.zdic.net site. These sections, which are derived from the Siku Quanshu (the library of the Qianlong Emperor in Four Sections) are the 经部 (*Jīng bù*, classics section), containing Confucian classics, 史部 (*Shǐ bù*, history section), containing historiographic works, 子部 (*Zǐ bù*, "masters" section), containing writings of the philosophical schools, and 集部 (*Jí bù*, anthology or *belles-lettres* section), a section of miscellaneous anthologies, including poetry, drama, and other works of literature. Each of the sections contains a multi-level tree of further subsections terminating in text files. For example, within the *Jīng* section are three subsections, labeled 十三经 (13 *jīng,* thirteen classics), 十三经注疏 (13 *jīng zhù shū*, thirteen classics annotations), and 经学史及小学类 (*jīng xué shǐ jí xiǎoxué lèi*, history of classical studies and traditional Chinese

philology) -- and these are further subdivided.[15] We found that some of the files were index files listing the contents of the directories, so we discarded these.

We developed a custom mixture of automated and semi-automated methods to extract the original texts from the downloaded HTML pages. Next we cleaned the corpus by regularizing the characters and their encoding method. Because of the mixture of traditional and simplified characters in the corpus, we decided to map all characters to simplified characters. This entails a loss of visual, aural and etymological information, important for interpretation by knowledgeable readers, but of no direct use to the algorithms beneath the topic modeling process. (In the future we will provide additional support for both traditional and simplified characters within the Topic Explorer.)

After this preliminary processing, we found that quite a few files were empty — some representing documents lost to history, others not present for other reasons. So, we removed these files leaving 18,414 files for analysis.[16] These files contain over 125 million individual characters. Chinese does not use spaces to separate words, but some words comprise multiple characters. Hence, text modelers face a choice of whether to model the corpus character-by-character or whether to segment the text into words. Because the vast majority of ancient Chinese words are written as single characters, the character-by-character option may have been a reasonable choice for this corpus. It was our judgment, however, that segmentation of the texts into words rather than characters would improve interpretability of the models.[17] Software to address the word segmentation problem in modern Chinese exists, but these solutions either use statistical, machine-learning approaches that have limited accuracy for classical Chinese,[18] or they are dictionary-based. Thus it was necessary for us to find and deploy a dictionary of ancient Chinese that we constructed from different sources.[19] Applying this dictionary to our corpus segmented the 125 million characters into over 104 million word tokens, comprising approximately 15,000 unique word types. (Segmenting with a modern dictionary resulted in approximately five times as many unique word types.)

The most common word in the Handian corpus is 之 (*zhī,* it/this/for) at just over 2.3 million occurrences and the most common two-character word was 以为 (*yǐwéi*, think) at just over 83,000 occurrences, 256th most frequent in the overall list. Very high frequency words are relatively uninformative and they tend to

overwhelm the available methods for corpus analysis, both because of the additional time to process so many characters in a corpus of this scale, and because the highly frequent terms tend to dominate more meaningful terms in the trained models. Therefore, it is normal to develop a "stop list" of such words to remove them from the corpus.[20] Our stop list of 142 words,[21] is slightly larger than the 132 words listed by Slingerland et al.,[22] and the two lists overlap in 59 words. The relative disjointness of the two can be explained by the differences in size and scope of the two corpora and the different objectives of the two projects. For example, we found it useful to filter out more of the frequently occurring number words.

LDA Topic Modeling

Based on our previous experience working with large text collections within the InPhO team at IU, we chose to apply LDA (Latent Dirichlet Allocation) Topic Modeling to the Handian corpus. (LDA is named for the 19th C. mathematician Gustav Dirichlet who laid the foundation in probability theory for the technique.) LDA Topic modeling has become popular within DH in recent years, although the interpretation of this kind of model remains a matter of considerable discussion.[23] It treats documents initially as "bags of words"—that is, all grammatical structure and information about word order within sentences or documents is ignored, and the document's initial profile is simply the frequency with which of all the words appear in it. Topic modeling aims to find latent (hidden) structure among these "bags of words", by re-representing each document as a mixture of topics. A topic may also be thought of as a writing context, as we now explain.

We understand topic models to provide a theory about writing. Authors of documents combine different subjects of discussion. Different authors working within similar cultural contexts have overlapping interests in various subjects, but they combine the available topics differently. When writing about good behavior, for example, one may be concerned with the good behavior in the public sphere of business or politics or religion, or in the family or social community, or as a topic within moral philosophy. An author is more or less likely to use a given word when writing about each of these subjects. For example, the words 'sister' or 'father' are more likely to be used when the author's subject is family than when writing about business. Other words may have very similar likelihoods of being used in these contexts. For example, the word "virtue" might be equally

likely to be used by authors discussing family or business matters. Discussion of good behavior may span the contexts of nature, family history, legal cases, theology, mythology, etc. Across a large corpus of documents we may expect to see these themes arising in different combinations—both when different authors are writing within similar cultures, and when one author writes at different times in his or her career. Furthermore, writers write for different contexts and audiences: letters to friends or family or superiors, philosophical dialogues, public speeches, etc. Each of these contexts also changes the likelihood of the author selecting certain words, and the same word in different contexts may produce slight or major variations in meaning.

LDA topic modeling provides a method for automatically identifying topics within a set of documents. At the end of a training process:

(a) each topic is represented as a total probability distribution over all the words in the corpus — that is, every word is assigned a probability in every topic, and the sum of all the word probabilities within one topic is equal to one; and
(b) each document is represented as a total probability distribution over the topics — that is, every topic is assigned a probability in every document, and the sum of the topic probabilities within one document is likewise equal to one.

The model starts with random probabilities assigned to the word-topic and topic-document distributions. It is trained by a process of adjusting the word-topic and topic-document probability distributions. The word-topic and topic-document distributions are controlled by two parameters (technically "hyperparameters" or "priors") that are set to ensure that there is sufficient variation in the probabilities assigned to the topics in the documents and to the words in the topics. The number of topics is chosen by the modeler. Our group typically trains multiple models with different numbers of topics, and we compare the different models to each other – see below for further discussion of the rationale for doing so. For the present study we trained models with 20, 40, 60, 80, and 100 topics. In general, with too few topics, each topic becomes very general and hard to interpret. With too many topics, some of the topics are specialized on just a few documents, making them less useful for finding common themes. While there exist methods within computer science for estimating an optimally efficient number of topics for a given corpus, users of the models may prefer a coarse-grained scheme (fewer topics) for some purposes while other users may prefer a more fine-grained

scheme (more topics) for other purposes.[24] Furthermore, working with multiple models simultaneously, fosters the kind of "interpretive pluralism" that characterizes humanities computing.[25]

The process by which we built these models using the InPhO topic-explorer package consists of four steps: initialization of the corpus object, preparation of the corpus by filtering words according to their frequency, training the corpus models, and launching the Topic Explorer's Hypershelf interface.[26]

Using the Topic Explorer & Notebooks

The Topic Explorer provides both a map-like visualization of the topic space (described further below) and a "Hypershelf" that allows users to experiment with the trained model to explore the corpus in any standard web browser. We call the latter interface a Hypershelf because although the browser initially presents documents from the corpus in a single linear order, it can be rearranged by the users to reflect their interests, and any document can be opened to view the full text. Thus, the Hypershelf initially provides a top level "distant reading"[27] view of the corpus, but allows the user to zoom down to the original text. This supports a two-way interaction in which interpretation of the texts helps the user to interpret the topics in the model, while interpretation of the topics in the model can help the user to interpret the texts. (We provide an example of this interplay below.)

The Hypershelf has two main modes: a document-centered view and a topic-centered view. Beginning with the document-centered view, the user may either select a document at random or use the search box to enter a few characters. These characters are automatically matched to the document labels, and the user can select a document from the drop-down list (Figure 1 shows initial options for the characters for *Lúnyǔ*, the Analects: 论语). Typing further characters narrows the search. The forward slash character '/' marks the hierarchical directory structure of the corpus.

Once a document is selected, and a number of topics for the model is chosen, the browser window is filled with a row of multi-colored bars (Figure 2), each block of color corresponding to a topic. The top row represents the topics assigned to the document by the computer during the final training cycle, according to the key at the right. Hovering over any of the colored sections displays a list of the highest probability words for that topic (see Figure 3). It is important to remember, however, that very many words are assigned non-zero probabilities in every topic, so these first few words do not exhaust the context provided by the

Figure 1. Autocompletion of document names within the InPhO Topic Explorer Hypershelf.

topic. Each subsequent row represents the topic distribution of another document from the corpus, scaled such that the length of the bar indicates similarity to the top document.[28]

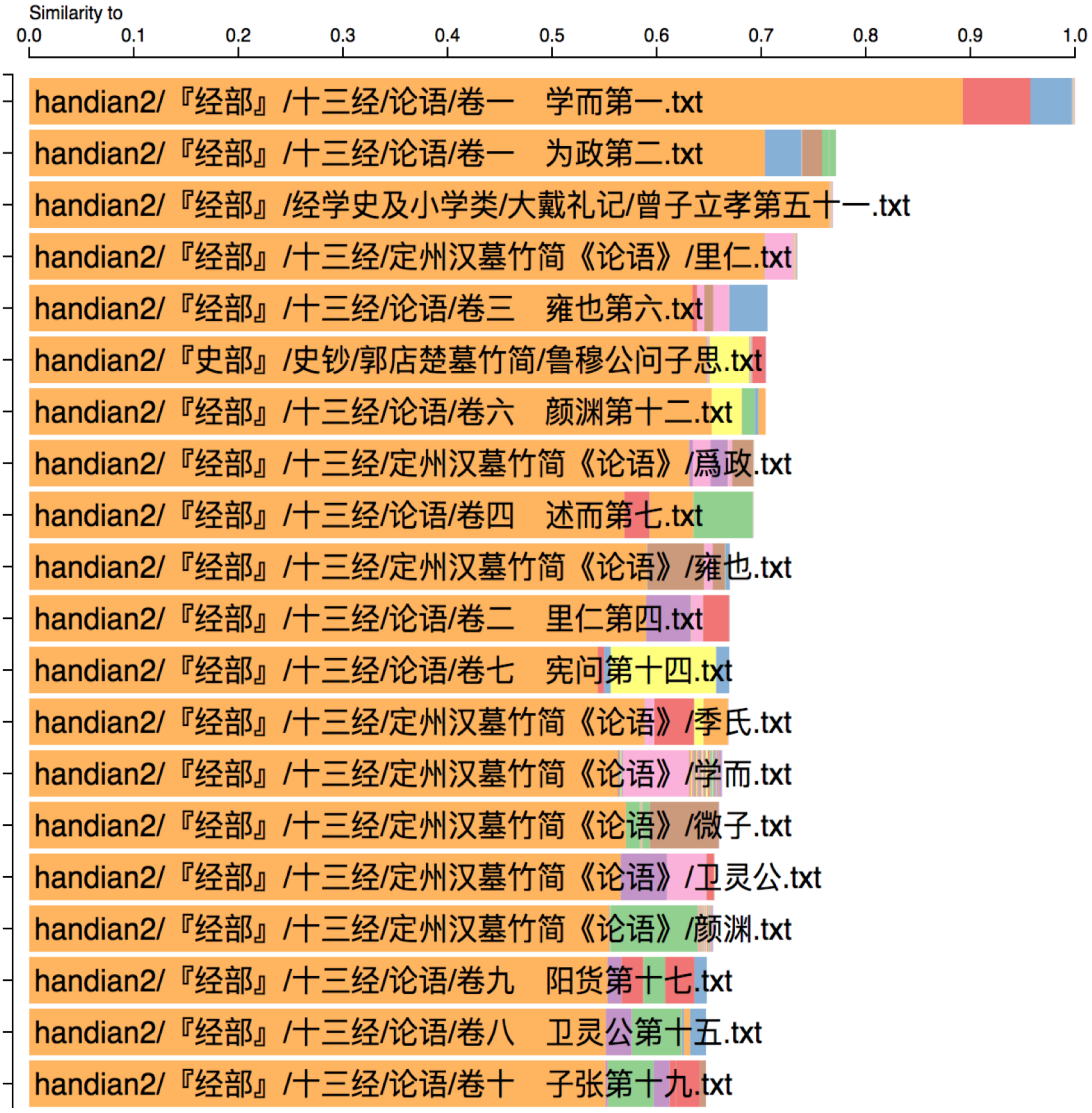


Figure 2. Similarity of documents to Book 1, Chapter 1 (卷一　学而第一) of the Analects (□□) in the 60-topic model. Each bar represents a document, and the colors represent the distribution of topics assigned to that document. (Different topics may be assigned to the same color.). The length of the bar indicates overall similarity to the document on the first row. Similarity is based on the Jensen-Shannon Distance between the overall topic mixtures of each document. See Figure 3 for additional details about the Topic Explorer display. See Appendix 6 for more information about the distance measure.

Initially, similarity between documents is shown with respect to the entire topic mixture associated with the focal document, but clicking the mouse on any of the topics re-sorts the list according to overall proportion of each document that the model assigned to the selected topic. This capacity of the HyperShelf allows users to rearrange the documents according to their interest in a particular topic (Figure 3).

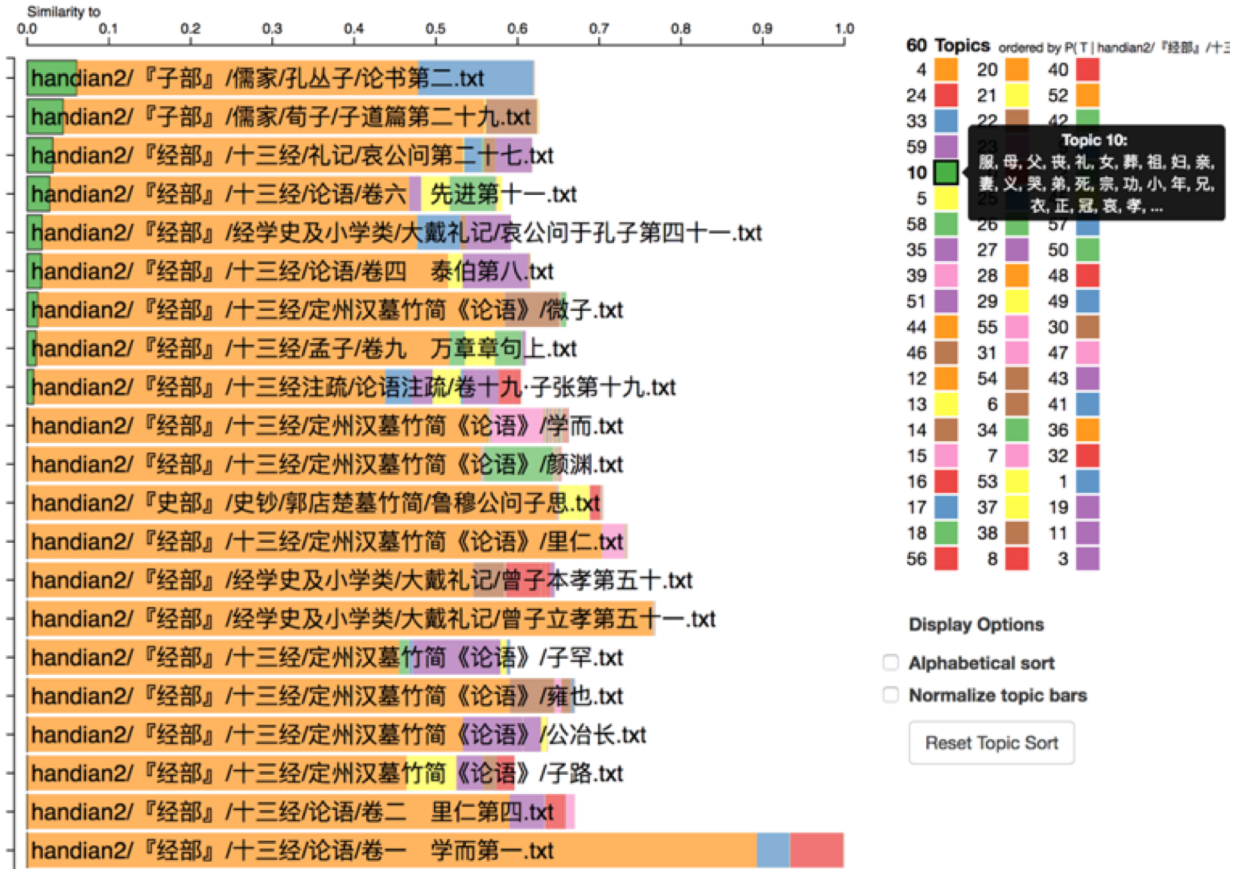


Figure 3. Highest probability words for each topic appear in the topic key at right when the cursor is positioned over the key, or over the corresponding topic in any of the document rows. Clicking in either location causes the Hypershelf to re-sort the list of results according to proportion of that topic assigned to each document. Here we show the reordering of results after selecting Topic 10, a topic about ritual, as the comparison dimension for documents most similar to Book 1 Chapter 1 (卷一　学而第一) of the Analects (□□) in the 60-topic model. (Note that this topic only represents a small portion of this selection of documents.)

From this point the user may click the "Top documents for Topic…" button below the key on the right (not shown in the Figure) to select the documents from the entire corpus that have been assigned the highest proportion of the selected topic. Alternatively, the user may choose to refocus on any of the other documents in the display by clicking on the arrow icon to the left of a row. (This icon appears when the mouse hovers nearby.) The user may read the full document by clicking on the "page" icon, which appears to the left of the arrow icon.

Results

We successfully trained topic models on the corpus of over 18,000 classical Chinese documents and made them available to explore interactively online.[29] We believe our choice to do word segmentation rather than single character modeling is justified by the contribution that the two-character words make to the interpretability of the topics, as well as by our investigation of 阴阳 (yinyang) within the corpus, as described below.

Topic models for humanities computing cannot be evaluated against a "gold standard" of correct performance, because no such standard exists. Neither could such a standard exist if one takes seriously the idea that the process of interpretation at the core of the humanities applies to the models as much as the texts (see Discussion section below), and is as variable as the interests of the users themselves. Ultimately, a topic-modeling approach succeeds or fails according to the ability of users to use the models for their own purposes, be it self-education, pedagogy, exploratory research, or systematic analysis of the texts. In future work we intend to assess how users respond to the topic models, and to conduct more complete analyses of relationships among the texts using the models. Here we present an example of how a particular individual used the Topic Explorer modeling and visualization results for self-guided investigation and serendipitous discovery — a process we refer to as "guided serendipity".[ii]

Our subject, one of the Chinese coauthors of this paper, began this project with only a basic familiarity with ancient Chinese philosophy acquired from an undergraduate course. He decided initially to investigate the concept of yinyang. This concept is important to many aspects of Chinese culture and philosophy, and has come to represent many dualities such as feminine/masculine, moon/sun, passive/active, dark/light, and so forth. Using the capacity of the Topic Explorer for topic-mediated term search, this term was queried in the 60 topic model.[30] Documents are retrieved according to their overall similarity to the topics selected by the term. The practical import is that because searches are topic-mediated, the documents retrieved need not contain the actual query term.

The first document identified in this way is from the *Zǐ Bù* ("masters") section of the corpus, which contains writings of the philosophical schools. It is from the 医家 (*Yī Jiā*, traditional medicine) subsection of the *Zǐ Bù* section, specifically the 素问 (*Sù Wèn*, or 'Basic Questions'), an important book from the Warring States

Period, which is a very famous dialog between the mythologized “Yellow Emperor” Huángdì (黄帝) and his minister, Qí Bó (岐伯), supposed to have lived in the third millennium BCE. The specific chapter located in this way is 天元纪大论篇第六十六 (*Tiān Yuán Jì Dà Lùn piān dì liù shí liù*, the 66th chapter of On the "Tianyuanji"). It describes the change of the seasons and the life process with the theory of yinyang and the Five Phases, which formed a basic framework of traditional Chinese medicine.

In the fourth row is a chapter from the *rú jiā* subjection, with concerns Confucianism. The chapter labeled ‘参两篇第二’ (*Cān Liǎng piān dì èr*, the 2nd chapter of Canliang, "Reflections upon Duality") from the volume labeled ‘张子正蒙’ (*Zhāng Zi Zhèng Méng*, or Master Zhang and Dispelling Ignorance) is part of the work Zheng Meng, which is very significant within the Confucian tradition. It was written by Zhang Zai (1020-1077), an important thinker of the Song Dynasty from Shaanxi province. The chapter relates yinyang theory to the astronomical calendar and the classical theory of Five Phases: Wood, Fire, Earth, Metal, Water (also referring to Jupiter, Mars, Saturn, Venus and Mercury respectively) used to explain the laws governing speed and direction of planetary motion.

Pursuing the idea that the Five Phases Theory provides the backbone of a broad system of thought encompassing many areas, our subject re-sorted the Hypershelf by clicking the arrow to the left of the fourth row, to refocus on this document. He then inspected the topics and identified topic 54 as seemingly most relevant to his interests. Clicking on that topic reorders the results according to the proportion of the topic allocated to each document in the list.

The top document identified in this way is from the 梦溪笔谈 (*Mèng Xī Bǐ Tán*, or Dream Brook Sketchbook), named 卷七· 象数一 (*Juàn Qī· Xiàng Shǔ Yī*, On Numerology with Images). This document, which is very abstract and obscure, tries to explain the fact that almost everything in this world is changing using the theories of yinyang and the Five Phases. By inspecting the documents near the top of the list, our subject noticed that many of them are about 周易 (*Zhōu Yì*, or the Book of Changes), or from the 三命通会 (*Sān Mìng Tōng Hui*, the title of a text about fortune telling and divination), or from the Sù Wèn, and from other areas. For example, the document 卷二·论大运 (*Juàn Èr·Lùn Dà Yùn*, On Long-

term Fortune) is part of the *Sān Mìng Tōng Huì* about the Five Phases theory, It relates the fortune of an individual to the yinyang property determined by the individual's birthdate. Also present are numerous documents from the yī jiā section. For example, the document 八正神明论篇第二十六 (*Bā Zhèng Shén Míng Lùn Piān Dì Èr Shí Liù*, the 26th chapter of *On Bā Zhèng Shénmíng*, roughly "On Eight Times and Mental States") is from the *Sù Wèn* section, it discusses acupuncture in the context of qi (气), a very important concept concerning life force or vital energy in traditional Chinese medicine, and connects qi to yinyang. In this chapter we can know how qi will influence the right place to put the needle and the appropriate depth of the needle, while qi will be influenced by yinyang, which is in turn determined by the season, weather, time and some other factors. Our coauthor reports that before using the Topic Explorer his concept of the Five Phases Theory was ambiguous, but in the interplay between topics and documents he learned many new details about the Five Phases Theory and its relationship to yinyang. This helped our subject to understand that the Topic Explorer could help him identify in which parts of Chinese culture the yinyang theory was prominent, namely divination, Confucianism and traditional Chinese medicine. For an expert, these interrelations may be well known, but for a learner, the capacity to rapidly relate the concepts in this way serves a very valuable function. His understanding of the complexity of the concept of qi was also broadened, leading to a plan to work further on this concept in future work with the topic models. This example shows how one individual's understanding of the connectedness of concepts from traditional Chinese medicine, astronomy, and religion was deepened by interaction with both the high-level overviews provided by the topic model and the close reading of specific texts directed by following the models. Although just one case, we believe that this case is not unique: the Hypershelf interface of the Topic Explorer supports spontaneous exploration and guided serendipity, customized to the user's particular interests.

We turn now from the Hypershelf to an interactive visualization of the entire topic space which is also provided by the Topic Explorer software package. This visualization can be explored interactively at InPhO websites. Figure 4 shows a map and cluster analysis of the topics across all five models. The map is generated using the isomap procedure applied to the Jensen-Shannon Distance measure (the same measure used to assess document similarity in Figure 2 etc.).[31] Isomap is a technique for reducing a high dimensional space (in this case the probability space of the word distributions in the models) to fewer dimensions—i.e., two, for this

visualization. Such dimensionality reductions are useful for identifying structure in the models. Whereas standard Multi-Dimensional Scaling (MDS) approaches are linear,[32] isomap detects non-linear structure in the data.

Groups of topics are clustered and colored automatically using k-means clustering with an arbitrary choice of ten clusters. Although these clusters are very broad, some general themes emerge—for example, the light green and dark red clusters above are related to history and geography, the light blue region in the middle left has some topics about historical accounts of war, while the light orange region below it contains mostly topics related to historical accounts of politics. The dark green cluster in the bottom is mainly about Confucianism and we can also find some topics about astronomy and Daoism in this region. The dark blue cluster in the middle right covers topics about literature and poetry,

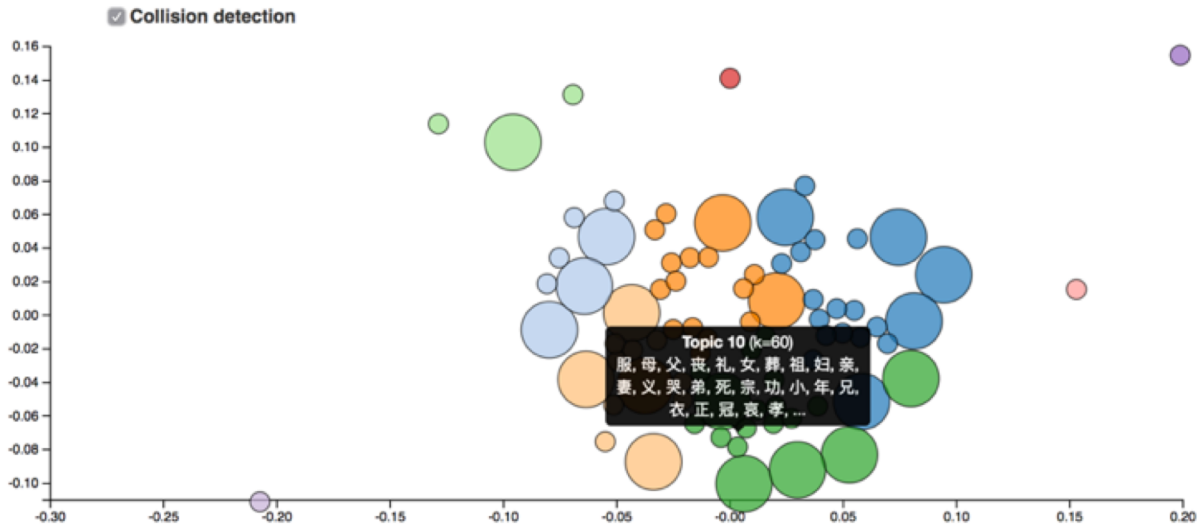


Figure 4. Topics from the 20- and 60- topic models arranged and clustered by the isomap algorithm. Values on the axes have no intrinsic meaning, but reflect relative scaling of the information theoretic distance between topics. Circle size is *inversely* proportional to the number of topics in the model: larger circles representing topics from the 20-topic model, smaller from the 60-topic model. In this example, the "Collision detection" box has been clicked to minimize overlap between circles, for easier reading, although this distorts the distances in the isomap. Proximity of topics within and between models, as represented by the locations of the circle centers, indicates semantic similarity. Hovering the mouse over any circle shows the highest probability words for any topic, as shown here for Topic 10 in the 60-topic model. See main text for further details, and the next two figures for applications of the map to locate topics using terms.

traditional Chinese medicine, and Buddhism. Although the focus of the dark orange cluster in the middle of this map is a little vague, we still can tell that its topics primarily concern history.

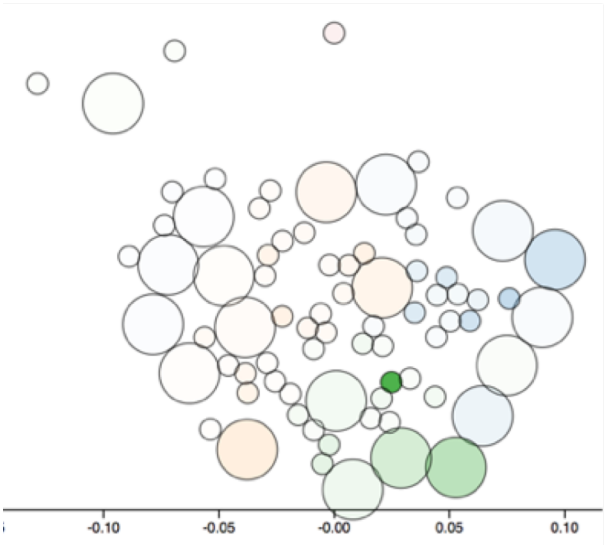
The map allows one to assess the overall similarity of topics in the different models. One important function this serves is to aid understanding of the relationships among topics in the different models. When presenting topic modeling, we are often asked whether the relationship between models with different numbers of topics is hierarchical: e.g., is each topic in the 20-topic model represented by two closely-related topics in the 60-topic model? Inspection of these figures, and the interactive map online reveals that these relationships are not strictly hierarchical, although semantically similar topics from the models with different numbers of topics do tend to cluster together, yielding a quasi-hierarchical structure with the topics from the more coarse-grained 20-topic model being associated with multiple topics from the 60-topic model. Another important function served by the isomap is that it allows users to easily compare the interpretability of the topics in the different models. We argue that there is no "one best number of topics". A beginner may prefer a coarse-grained model, while experts may prefer (and be able to make sense of) the distinctions afforded by a finer-grained model. Furthermore, to go beyond the visualizations to quantitative analysis of hypotheses about the corpus, is it is important to know which results are robustly found in the different models, rather than perhaps being artifacts of a particular value for the number of topics.[33]

The interactive online version of the topic isomap supports exploration of the models in a variety of ways. Hovering the mouse over the elements of the map shows the highest probability words for the topics, as shown in Figure 4, and allows the user to click through to the Hypershelf, showing documents most similar to that topic. Alternatively, entering a word or words in the search box above the map adjusts the colors in the map to show the relative weight of the term across all the topics. More saturated colors indicate a higher probability of the term being generated by that topic. Figure 5 shows a comparison of the terms 孔 (the first character of '孔子' representing Master Kong, i.e., Confucius; Figure 5a), and 佛 (representing the Buddha, Figure 5b) within the 20- and 60- topic models. Both are implicated in many of the topics, as indicated by the high number of topics receiving some color. The query about Confucius identifies a concentration in the dark green cluster of the isomap, with topic 60:4 having the highest saturation, while the query about Buddha identifies a concentration in the dark blue region of the map, with topics 60:50 and 20:11 having the highest saturation. Clicking on any of the topics identified in this way takes the user to

the Hypershelf with the top documents for that topic already loaded, and inspection of those documents show the results to be highly related to Confucianism and Buddhism, as one would desire.

Figure 6 shows a similar comparison for the terms 气 (qi) and 阴阳 (yinyang). Here the distributions are quite similar, although the topics related to qi are more concentrated on the right side of the diagram whereas topics related to yinyang are distributed a bit more across central parts of the map. The relative confinement of topics related to qi corresponds to the fact that the Isomap algorithm has placed health and traditional medicine topics on the right side of the map in the dark blue cluster.

(5a) 孔, Kong



(5b) 佛, Buddha

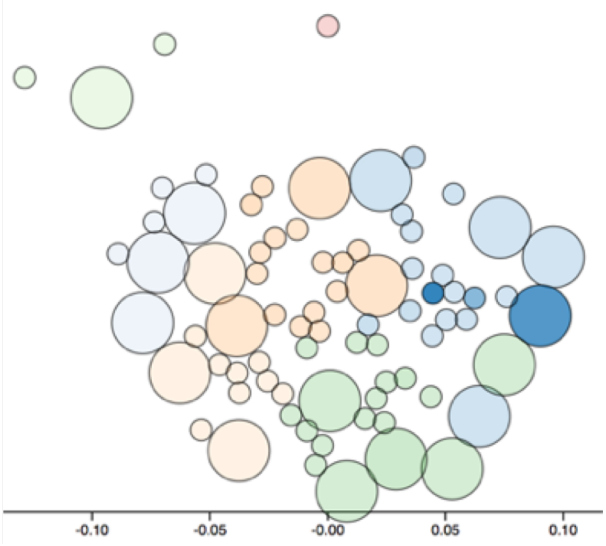


Figure 5. See Figure 4 for explanation of underlying map. Here, colors are saturated according to relevance to word entered in the search box: (5a) 孔, for Kong (Confucius), and (5b) 佛, Buddha/Buddhism. Both are significant for many of the topics in the models, but孔selects more highly for topics in the dark green cluster, while 佛 selects more highly for topics within the blue cluster. See main text for further discussion.

Among the most central topics in the map (i.e., those near to the 0,0 origin in the map) are these from right side of the dark orange cluster: [34]

20:19 天, 德, 神, 风, 光, 阙, 灵, 山, 道, 清, 皇, 明, 命, 万, 元
40:20 钱, 百, 民, 万, 千, 年, 田, 户, 岁, 官, 盐, 州, 税, 数, 给
60:14 钱, 百, 万, 民, 年, 千, 田, 官, 户, 州, 岁, 盐, 税, 数, 给
80:43 钱, 百, 万, 民, 年, 千, 官, 田, 户, 盐, 州, 岁, 税, 给, 数

100:19 钱, 盐, 年, 万, 官, 州, 百, 运, 税, 岁, 司, 利, 法, 行, 数

These topics from the 40-, 60-, 80-, and 100-topic models are highly convergent, with highest probability terms relating to taxation and state rules and regulations. The significance of centrality in the map requires further analytical and interpretive work. What we know at present is that different runs of the model produce somewhat different maps, although government-related topics tend to be central across these different runs. This stable feature of the models may be seen as reflecting both the large number of government documents in the Handian corpus, and the central importance of the civil service in China for the preservation and transmission of classical Chinese culture and values.

In the models presented here, the 20:19 topic is an apparent outlier, relating to literature and poetry, however there is another topic nearby in the isomap, located just to the left of the center that has a high degree of word overlap with the other four, i.e., 20:4官, 州, 钱, 本, 司, 诏, 路, 年, 百, 民, 令, 日, 臣, 法, 万. The convergence among the topics and the map helps give some confidence in the models and the clustering technique. It is important to keep in mind, however, that the isomap plot is generated using the full term distribution, not just the first 15 terms shown here, and the reduction of multiple dimensions to just two further distorts the true relationships among the models; hence it cannot be taken as a complete guide to the underlying models. A complete assessment of the topics and their related documents would need to go beyond simple inspection of the top terms in each topic and the isomapped locations.

(6a) 气, qi

(6b) 阴阳, yinyang

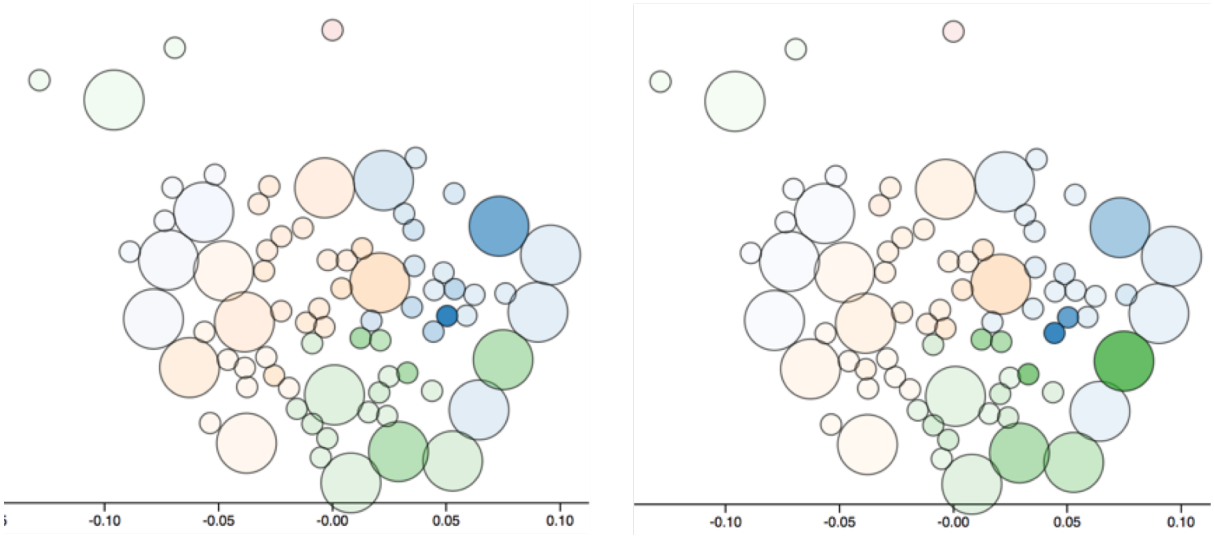


Figure 6. See Figures 4 and 5 for explanation of layout and coloring scheme. Here we compare distribution of topics for two terms：(6a) 气, qi, and (6b) 阴阳, yinyang. The distribution of these topics in the corpus is rather similar, and is convergent particularly at topic 60:56, which is the most saturated of the blue circles in (6a). That topic relates to Chinese traditional medicine. The slightly more saturated blue circle in (6b) a non-medical topic that is heavily represented in the chapters from the Book of Changes that discuss yinyang.

Discussion

LDA topic models are not themselves interpretations of the documents — indeed they stand in need of interpretation themselves[35]—but they may assist scholars in exploring and interpreting large collections of materials. Ultimately there is no substitute for reading the documents, but the Topic Explorer interface, through its Hypershelf and Topic Isomap components, can guide scholars and learners alike to documents that they might not have otherwise encountered or thought to look for, resulting in a particularly productive form of guided serendipity.

Topic models are interesting to think about from the perspective of theories of meaning. While they do not capture exact meanings—"John loves Mary" and "Mary loves John" are viewed as identical documents under the "bag of words" assumption—they are quite successful at capturing something like the general gist or context of the words being used. Scholars of Chinese literature have emphasized the high degree of context sensitivity for the meanings of words in the Chinese language,[36] but a strength of the topic modeling approach is that the same word is placed in multiple contexts, helping with the process of

disambiguation. At the same time, because the models have a solid grounding in information theory, the use of metric measures such as the Jensen-Shannon distance is feasible for many applications. This provides new forms of evidence for humanistic discussions.

Although the corpus we used may be missing some potentially important documents, it is large enough that the topic models we derived from this corpus prove to be adequate for various purposes. Improved curation of the corpus nevertheless remains an important goal of our group for the future, and will be reflected in future iterations of the Handian Topic Explorer mirror site. Future work will allow us to address questions about topical relationships among the documents in the Handian corpus and about historical and geographical shifts in the topic distributions as represented in the corpus model, and ultimately to analyze the behavior of individual authors.

Finally, and more speculatively, philosophers of mind and cognitive science have sometimes been tempted by the idea that meaning or semantic content assignment is a kind of measurement process rather than the assertion of a relationship to a determinate abstract proposition.[37] Computer scientists have started to provide the means to convert this idea into quantitative models[38] to which measures such as Jensen-Shannon distance can be applied. Thus, the Digital Humanities are poised to have a significant impact on philosophical and practical discussions of the nature of meaning.

Conclusions

Topic models present a powerful new tool for computer-assisted interpretation in the humanities. We have described some particular issues faced for using topic models with ancient Chinese texts, and we have detailed the process of training LDA topic models on the Handian corpus of over 18,000 classical Chinese texts using the InPhO Topic Explorer. The results of these efforts and the software we have developed have been made publicly available via the Hypershelf interface at mirror sites at Xi'an Jiaotong University and Indiana University. This interface allows users to visualize the results of the modeling process. We have provided some preliminary description and analysis of the topics discovered by the algorithms using the more advanced notebook features of the Topic Explorer.[39] These preliminary investigations have revealed some interrelationships among Confucian, Taoist and Buddhist themes, and the penetration of these themes in

many aspects of traditional Chinese culture, from medicine to government. By following the threads among specific texts, guided by these topic models, scholars may exploit these new tools to enrich their understanding and interpretation of China's rich cultural heritage.

Acknowledgements

The software described in this paper was originally developed at Indiana University with generous support from the National Endowment for the Humanities and IU's Office of the Vice President for Research. Its extension to ancient Chinese owes much to IU's support and to the research funding provided by the College of Humanities and Social Sciences, the office of Dean Yanjie Bian and the Philosophy Department at Xi'an Jiaotong University. The authors of this paper, we would like to thank Xiaoliang Wang and Wenjing Yuan at XJTU for their initial guidance concerning ancient Chinese philosophy. We also acknowledge the prior programming efforts of Robert Rose, Doori Lee, Jessie Pusateri, and Adithya Nagaraj-Tirumale at Indiana University. We are grateful to Henry Rosemont, Jr. for comments on an earlier draft of the manuscript, and to an anonymous referee for comments on the first submitted version. All errors of interpretation remain our own.

Notes

[1] Jian Ouyang, "Visual Analysis and Exploration of Ancient Texts for Digital Humanities Research," *Journal of Library Science in China* 42, no. 222 (2016), 66-80.
[2] Zixuan Huang and Jinsong Yu, "Topic Model based SongCi Corpus Construction and Research on Computer aided SongCi Writing," *International Journal of Knowledge and Language Processing* 3, no. 2 (2012), 1-19.
[3] http://ctext.org/
[4] Ryan Nichols, Kristoffer L. Nielbo and Uffe Bergeton, "Topic Modeling the Ancient Chinese Corpus: The Textual Contexts of High and Low Gods in Chinese Thought." Conference presentation at Cultural Evolution of Religion Research Consortium, Montreal, Canada, May 2015. Ryan Nichols, Edward Slingerland, Kristoffer Nielbo, Uffe Bergeton, Carson Logan and Scott Kleinbahn, "Modeling the contested relationship between *Analects*, *Mencius*, and *Xunzi*: Preliminary evidence from a Machine-Learning Approach," submitted. Slingerland, Nichols, Nielbo and Logan, "The Distant Reading of Religious Texts: A "Big Data" Approach to Mind-Body Concepts in Early China," submitted.
[5] Rockwell & Sinclair *op cit.* David M. Blei, "Probabilistic Topic Models," *Communications of the ACM* 55 (2012), 77–84. Ted Underwood, "Theorizing Research Practices We Forgot to Theorize Twenty Years Ago," *Representations* 127 no. 1 (2014), 64-72.
[6] Henry Rosemont, Jr. "Translating and Interpreting Chinese Philosophy," *The Stanford Encyclopedia of Philosophy (Summer 2016 Edition)*, Edward N. Zalta (ed.), URL http://plato.stanford.edu/archives/sum2016/entries/chinese-translate-interpret.
[7] Qi Zhao Q, Zengchang Qin and Tao Wan, "Topic Modeling of Chinese Language Using Character-Word Relations," in: Bao-Liang Lu, Liqing Zhang and James Kwok (Eds.), *Proceedings of the 18th International Conference on Neural Information Processing,* ICONIP 2011. *Lecture Notes in Computer Science*, vol 7064. (Berlin, Heidelberg: Springer, 2011).
[8] Rosemont *op. cit.*
[9] Jaimie Murdock and Colin Allen, "Visualization Techniques for Topic Model Checking," *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI).* (Austin, TX: AAAI Press, 2015), http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/10007.
[10] http://online.eastview.com/projects/skqs/
[11] http://www.zdic.net/aboutus/ (retrieved November 1, 2016).
[12] Nichols et al. 2015 and submitted, Slingerland et al. submitted.
[13] Christian Wittern maintains the repository at kanripo.org and https://github.com/kr-shadow/kr-shadow.github.io.
[14] http://inphodata.cogs.indiana.edu/handian/ and http://inpho.xjtu.edu.cn/handian/.
[15] The 2nd-level sections of the Handian corpus are listed in appendix 1 of the supplemental materials available at 10.7910/DVN/3QXX29
[16] After training the topic models, we discovered, with the help of the models, that two folders of English translations of the Analects (20 files) and Lao Tze (84 files) had slipped through our net. We decided not to remove these and retrain the models because the presence of these files has a minimal effect on the overall results.
[17] Zhao, Qin & Wang *op. cit.* Ouyang *op. cit.*
[18] For paper on automatic segmentation of modern Chinese see: Ke Deng, Peter K. Bolb, Kate J. Lic and Jun S. Liu, "On the unsupervised analysis of domain-specific Chinese texts." *Proceedings of the National Academy of Sciences* 113 (2016), 6154-6159; Mengqiu Wang, Rob Voigt and Christopher D. Manning, "**Two Knives Cut Better Than One: Chinese Word Segmentation with Dual Decomposition.**" Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Volume 2: Short Papers (June 2014), 193-198.) http://www.aclweb.org/anthology/P/P14/P14-2032
[19] Corpus cleaning methods are detailed in appendix 2 of supplemental materials. The ancient words dictionary is available at https://github.com/inpho/topic-explorer/blob/master/topicexplorer/lib/ancient%20words.dic and embedded within the Topic Explorer using the "och" (old Chinese) extensions.
[20] Gerard Salton, Andrew Wong and Chungshu Yang, "A Vector Space Model for Automatic Indexing," *Information Retrieval and Language Processing* 18 no. 11 (1975), 613-620. Christopher Fox, "A stop list for general text," *SIGIR Forum* 24 nos. 1-2 (1989), 19-21. Steven Bird, Ewan Loper and Edward Klein, *Natural Language Processing with Python*, (Sebastopol, CA, O'Reilly Media Inc., 2009).
[21] The stop list is provided with the data repository included with this paper and in our Github distribution.
[22] Supplemental materials appendix 3 shows the 20 most common words, and their frequencies, and describes the process by which we developed our list of 187 words for ancient Chinese in Appendix 3. For the stop word list used by Slingerland et al. see their "Modeling the contested relationship", submitted.
[23] Blei, "Probablistic Topic Models."
[24] Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov and David Mimno, "Evaluation Methods for Topic Models," in *Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09* (New York, NY, USA: ACM, 2009), pp. 1105–12. Margaret Roberts, Brandon Stewart and Dustin Tingley, "Navigating the Local Modes of Big Data: The Case of Topic Models," in *Data Analytics in Social Science, Government, and Industry* (New York: Cambridge University

Press, 2015). Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber and David M Blei, "Reading Tea Leaves: How Humans Interpret Topic Models," in *Advances in Neural Information Processing Systems*, 2009, pp. 288–96.

[25] Rockwell and Sinclair *Hermeneutica*.

[26] Details of the modeling process provided in supplemental appendix 4. The topics discovered by training on the Handian corpus are represented in appendix 5, which shows the 15 highest probability words from each topic in the 20, 40, 60, 80, and 100 topic models.

[27] Franco Moretti, *Distant Reading* (New York: Verso, 2013).

[28] Jensen-Shannon Distance is due to Dominik M. Endres and Johannes E. Schindelin, "A New Metric for Probability Distributions," *IEEE Transactions on Information Theory* 49, no. 7 (2003), 1859-1861. It is based on Kullback-Leibler divergence introduced by Solomon Kullback and Richard A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics* 22, no. 1 (1951), 79–86. See supplemental appendix 6 for additional information.

[29] The models may be explored interactively at http://inphodata.cogs.indiana.edu/handian/ or at http://inpho.xjtu.edu.cn/handian/. See also supplemental appendix 5.

[30] Cameron Buckner, Matthias Niepert and Colin Allen, "From encyclopedia to ontology: toward dynamic representation of the discipline of philosophy," *Synthese* 182 (2011), 205-233.

[31] Joshua B. Tennenbaum, Vin de Silva and John C. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction," *Science* 290, no. 5500 (2000), 2319-2323. We applied the isomap algorithm to the JSD measure. See Appendix 6 for more information about Jensen-Shannon Distance.

[32] Readers may be familiar with Principal Components Analysis (PCA) which is one specific method in the MDS class.

[33] See Murdock et al. *op. cit.* for an application of this check to topic models of Darwin's reading.

[34] We do not provide English translations of all of these characters because in many cases an exact translation is difficult to provide. Rather we invite non-Chinese speakers to inspect the number of repeated characters in these lists. Topic labels refer to number of topics in the model and the (arbitrarily-assigned) topic number; e.g., "20:19" refers to topic number 19 in the 20-topic model.

[35] Blei, "Probabilistic Topic Models"; Rockwell and Sinclair *Hermeneutica*; Underwood, "Theorizing Research Practices."

[36] Rosemont, "Translating and Interpreting Chinese Philosophy."

[37] For example, Paul M. Churchland, *Scientific Realism and the Plasticity of Mind* (Cambridge: Cambridge University Press, 1979), and Robert Matthews, *Measure of Mind* (New York: Oxford University Press, 2007).

[38] For a review see Michael N. Jones, Thomas M Gruenenfelder and Gabriel Recchia, "In Defense of Spatial Models of Lexical Semantics," in Laura Carlson, Christoph Hölscher and Thomas Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (Austin, TX: Cognitive Science Society, 2011), 3444–3449.

[39] The 'topicexplorer notebook' functionality is briefly described in Appendix 4. The Jupyter interface can be used for simple functions such as retrieving top words for each topic in a model, closest topics to a word or set of words, closest document to a topic or set of topics, as well as providing a full iPython programming environment for more advanced analyses.