DHQ: Digital Humanities Quarterly
2025
Volume 19 Number 1

Experiments in Distant Reading: Using Topic Modeling on Chinese Buddhist Texts from 500-800 CE

Marcus Bingenheimer <m_dot_bingenheimer_at_gmail_dot_com>, Temple University https://orcid.org/0000-0002-9102-9217
Justin Brody <jdbrody_at_gmail_dot_com>, Franklin and Marshall College https://orcid.org/0000-0001-7486-0175
Ryan Nichols <rnichols_at_fullerton_dot_edu>, Califoria State University, Fullerton https://orcid.org/0000-0001-9445-9821

Abstract

The article tries to answer whether the BERTopic topic modeling framework can be used to obtain topics that meaningfully distinguish two corpora of Buddhist Chinese texts from 500 to 800 CE. The first corpus consists of translated “Indian-Chinese” Buddhist texts, the second of “Chinese-Chinese” texts, i.e. texts directly authored in Buddhist Chinese. Does the application of topic modeling reveal aspects that are typical for these corpora and do these topics suggest avenues for future research into the sinicization of Buddhism that took place during that time?

For our implementation of BERTopic, we used the customized GuwenBERT, a language model trained on classical Chinese. To reduce the dimensionality of the embeddings we used the UMAP algorithm. Next, the HDBSCAN takes care of hierarchical clustering. The most relevant words of each cluster are identified with c-tf-idf. As a last step, we score each cluster by its monochromaticity – this is a measure of how likely the documents in the cluster are to be derived from either just the “Chinese-Chinese” or just the “Indian-Chinese” documents.

In order to communicate the topics we create virtual paragraphs that combine most of the top twenty terms that represent a sample of ten highly monochromatic topics. Discussing these topics from a Buddhist Studies point of view, we find that our modified BERTopic workflow does indeed return topics that are characteristic of their corpus and highlights facets that help to understand the process of how Buddhism became sinicized in the three centuries between 500 and 800 CE. Thus distant reading of latent topics in the corpus is possible. While some topics are in themselves unsurprising, others highlight new promising areas for research.

1. Introduction

In an article in these pages on “Genealogy of Distant Reading” Ted Underwood made the valid point that distant reading as a form of “macroscopic literary inquiry” did not originate with digital text [Underwood 2017]. Still, all DH methods, by definition, process digital data, and most can only be applied effectively only within a computational setting. While some methods (e.g. markup or GIS) extend analog practice (such as editing or cartography) into the digital, others are digital natives. Topic modeling is one of the computational methods that over the last twenty years have made their way from Natural Language Processing [Blei et al2003] into the Digital Humanities. It has been consistently associated with distant reading. In the introduction to a special issue on topic modeling in the *Journal of the Digital Humanities* the editors even called it “distant reading in the most pure sense” [Meeks and Weingart 2012]. [1] As digital corpora grow ever larger, so does the need for distant reading. Thus topic modeling has survived early hype and critique [Schmidt 2012] and is still going strong in DH [Du 2019] and in information retrieval in general. Originally dominated by the application of Latent Dirichlet Allocation [Blei et al2003], by now there are several different approaches to topic modeling. Comparing their respective strengths and weaknesses has resulted in a veritable sub-genre of comparative studies [Kherwa and Bansal 2019], [Vayansky and Kumar 2020], [Fu et al 2020], [Ma et al 2021], [Egger and Yu 2022], [Rüdiger et al 2022], [Chen et al 2023]. In this paper we will use a modified version of BERTopic [Grootendorst 2022], a

new, neural network based approach, to distant read two related text corpora in Buddhist Chinese, a low-resource idiom. Although an attempt has been made to use BERTopic on Chinese poetry [Fang 2023], to our knowledge this is the first time BERTopic has been used on canonical Buddhist texts in any language.

Our first question is whether BERTopic, which relies on BERT for word embeddings, does return coherent topics at all when confronted with a low-resource idiom such as Buddhist Chinese, which was not represented in the training data for BERT, and for which tokenization algorithms are available but not widely implemented [Wang 2020]. Secondly, we would like to know whether the application of the BERTopic workflow can yield meaningful topics that can be taken to express concerns in Chinese Buddhist texts created between 500 and 800 CE and might in particular be used to distinguish two different corpora, Indian-Chinese and Chinese-Chinese texts, and how they might be used in furthering our understanding of the texts. 2

Why this particular time-frame? In the centuries between 500 and 800 CE, Buddhism in China turned into Chinese Buddhism. The first Indian Buddhist texts were translated into Chinese in the second century. Starting from the late third century we are able to reconstruct an unbroken social network that connects generations of Chinese Buddhists until today [Bingenheimer 2021]. [2] Until the 6th century this network was mainly informed by Indian-Chinese texts, i.e. Indian texts translated into Chinese. Chinese-Chinese texts, i.e. texts authored by Chinese (or Korean or Japanese) writers, began as paratext (prefaces, catalogs etc.) to these translations, then developed via short essays, apocrypha and commentaries, eventually blossoming into long historiographical and doctrinal works in the 6th century. The time between 500 and 800 is special in the sense that in spite of the continuing reception and translation of Indian texts, it were the Chinese-Chinese texts produced during that period that became distinctive for later Chinese Buddhism. The philosophical works of Zhiyi 智顗 (538–597) and Fazang 法藏 (643–712), the devotional treatises of Tanluan曇鸞 (476-542) and Shandao善導 (613–681), and the recorded sayings of early Chan masters like Huineng 慧能 (638–713) and Mazu Daoyi 馬祖道一 (709–788), to give but a few examples, became the textual foundation for the dominant traditions within Chinese, and indeed East Asian, Buddhism of the second millennium: Chan, Pure Land, and Tiantai Buddhism. 3

While in the year 500 CE Buddhism in China was still very much dominated by the influx of Indian Buddhist ideas, by 800 CE Chinese Buddhists had started to put their own spin on Buddhism. Most features that would come to characterize Chinese Buddhism were in place, even if it took another few centuries for them to fully form: the dichotomy, both stable and dynamic, between meditative Chan and devotional Pure Land practice, an interest in historiography and lineage, a commitment to the lay-monastic distinction, and the necessity of apologetics in the framework of the Chinese state. While Buddhism declined in India after 800 CE and became virtually extinct there after the 12th century, Chinese Buddhism blossomed throughout the Song Dynasty (960–1279) and today remains the largest world religion in China with more than 200 million adherents [Pew Research Center 2023]. 4

The topics returned by topic modeling methods are notoriously open to interpretation. Whether or not they cohere with respect to a particular domain can only be evaluated by domain experts. In order to communicate how we read a topic to a wider audience we combine the lists of terms that represent topics into “virtual paragraphs” (Appendix A). In that way we hope to make our reading transparent for non-specialists who can easily grasp what the topic is about. Creating virtual paragraphs from topics could also be used in teaching. When practiced in conjunction with close, linear reading of classical texts, students stand to gain familiarity with concepts and semantic fields that are latent in larger corpora. 5

2. Data

Based on the largest digital archive of Chinese Buddhist texts [CBETA ver 2021] we have assembled two corpora of comparable size. [3] 6

1. Indian-Chinese: This consists of all 661 originally Indian texts in the CBETA corpus that were translated into Buddhist Chinese between 500 and 800 CE. Total number of characters: 17,959,037.
2. Chinese-Chinese: This corpus consists of all 293 texts in the CBETA corpus that were composed directly in Buddhist Chinese between 500 and 800 CE. Total number of characters: 21,136,841.

Both corpora include material from before 500 CE. Obviously, the Indian-Chinese texts are translations and might have

been authored before 500 CE. But the Chinese-Chinese corpus too, although authored between 500 and 800 CE, contains commentaries on older texts and thus includes material (and topics) from earlier periods. This caveat notwithstanding the corpora represent two different modes of textual production in Chinese Buddhism during that period and an experiment in modeling is justified and might yield new insights. In Section 4 we will try to interpret select topics guided by the hypothesis that the topics in these two corpora differ meaningfully. It should be understood that the topic modeling workflow we implemented does not process the two corpora separately. Whether a topic is aligned with the Indian-Chinese or the Chinese-Chinese corpus is determined by its monochromaticity (s. Sec. 3). For the topic modeling pipeline described here we use these texts in the tokenized, segmented version produced by Yu-chun Wang (making use of the Conditional Random Field (CRF) model he describes in [Wang 2020]. [4] This is currently the most consistently segmented available corpus for Buddhist Chinese.

We use the term “Buddhist Chinese” here for the particular idiom (or set of idioms) in which pre-modern Buddhist texts were composed. Needless to say, across 1700 years there was a lot of variation. The translation of Indian Buddhist texts into Chinese was never standardized and there exists a many-to-many relationship in the translation of terms: One Indian word could be rendered by different Chinese characters, and the same Chinese character was used for different Indian terms. Chinese-Chinese Buddhist texts were generally closer to the classical idiom of literary Chinese than translated texts, but they share features with translated texts that sets them apart from the usual “classical Chinese” written in that period. Among these features are the morphology of Buddhist vocabulary which makes words more likely to be compounds of two more more characters, instead of the preferred one-character = one-semantic-unit equivalence of “*classical* classical Chinese.” [5] 7

The reason why there are more than twice as many Indian-Chinese than Chinese-Chinese texts is because in the eight century a large number of short *dhāraṇī* text were translated, or in fact rather transcribed. The main component of such texts is the *dhāraṇī* itself: a Sanskrit invocation or spell, which was used in the context of esoteric ritual (see also Sec. 4). Although longer than a mantra, *dhāraṇīs* could not be too long as many of them were supposed to be remembered by heart, but within esoteric Buddhism they were popular. Thus the Indian-Chinese corpus consists of a larger number of short texts, in spite of the corpus being overall smaller than the Chinese-Chinese corpus in terms of characters. 8

3. Method

Our method follows BERTopic [Grootendorst 2022], a recent topic modeling framework that has quickly attracted attention. [6] We use our own variation of the BERTopic pipeline, which we make available at https://github.com/mbingenheimer/cbetaCorpusSorted. [7] Specifically, our implementation involves the following steps: 9

1. Using a large language model trained on classical Chinese (Koichi Yasuoka’s variant of GuwenBERT), we embed every sentence in our corpus into a high dimensional vector space.
2. We perform UMAP to reduce the sentence embeddings to a 3-dimensional set.
3. We use HDBSCAN to determine clusters which we interpret as topics.
4. The words in the sentences of each topic are collected and ranked according to c-tf-idf to measure the degree to which they represent the entire topic.
5. The individual sentences are tagged as to their provenance (whether they come from a Chinese-Chinese or a Chinese-Indian source). This allows us to compute the monochromaticity of each cluster and determine which clusters are mixed versus which ones are representative of one or the other of our corpora.

The main idea of BERTopic is to take advantage of the capacity of modern neural networks to create sentence embeddings which respect *semantic similarity*, meaning that linguistic elements with similar meaning should be embedded close to each other in the target vector space. While many previous techniques are based on mapping language elements to vectors (with word2vec probably being the best-known), encoder-only models like BERT have the advantage of embedding language in a way that is sensitive to semantics. In particular the embedding of a multivalent word “bank” is dependent on the surrounding sentence. Thus homonyms can be disambiguated in the embedding. At the same time synonyms like “happy” and “joyful” that have a similar probability of occurring at a particular position in a sentence are defined with similar embeddings. As a result, any BERT model can be expected to display some degree of 10

identifying semantic similarity. Specifically, these models operate by learning how to predict missing tokens (e.g. words) from a natural language sentence. To be successful, the model has to learn the probabilities associated with various possible words at a specific point in the sentence. This will necessarily entail disambiguating homonyms – if it would represent all instances of the word bank in the same way then it could not correctly learn the context-dependent likelihoods of the words *money* and *river* occurring elsewhere in the sentence. Similarly, it would be helpful (if not strictly necessary) for synonyms to have close embeddings; this would allow them to automatically treat their influence on the rest of the sentence similarly.

11 Moreover, BERT models can be explicitly trained to respect semantic similarity – this requires a custom dataset designed for this purpose. In our case we chose Koichi Yasuoka's variant of GuwenBERT as our base model. [8] GuwenBERT is a BERT model which is trained on the Daizhige 殆知阁 dataset, which according to the creators of GuwenBERT contains "15,694 books in Classical Chinese, covering Buddhism, Confucianism, Medicine, History, ...Taoism, [and others]". [9] Whereas the original GuwenBERT was trained on the corpus in simplified Chinese, Yasuoka's variant allows for traditional Chinese characters. To our knowledge, there is no semantic similarity training set for Classical Chinese; thus we rely on the level of semantic similarity native to our base model. [10] This results in the danger of outputting collocations instead of topics: For example, our pipeline sometimes picks up on ngrams that contain "十一eleven (-somethings)" or "四十forty (-somethings)", resulting in clusters that are based on one or two characters rather than genuine, domain-specific semantic clustering. By grouping for instance, " 四十 (forty), 四十二 (Forty-two), 四十九 (forty nine), 四十八 (forty-eight), 四十四 (Forty-four), 四十一 (forty one), 四十六 (Forty -six)" in one cluster the pipeline does its job (it picks up on the signal of "forty"), but because we are (not yet) able to fine-tune it to produce more semantically relevant clusters. The more a cluster looks like a snippet from a KWIC index, the less interesting it is for topic modeling, because we are looking for semantically related terms not for the same term in different contexts.

12 We pass each sentence of each document into our base model, thereby obtaining a high-dimensional vector representation of each sentence. These sentences are tagged according to the corpus they originated from. It is understood within data science that high dimensional datasets are difficult to work with; many mathematical algorithms designed for low dimensional datasets break down in high dimensions – this is known as the *curse of dimensionality*. Thus our set of high dimensional sentence representation is then passed through the UMAP algorithm [McInnes et al 2018]. which reduces the representations to being 3-dimensional. UMAP is designed so that representations which are close (in some sense) in the high dimensional space will get transformed into representations which are still close in 3-dimensional space. This ensures that clusters that are observed in the 3 dimensional space do actually correspond to representations which are close in the high-dimensional space. If our embeddings display a high degree of semantic similarity, this should mean that sentences whose representations are clustered close together actually represent sentences with related meanings.

13 Having reduced the sentences representations to three dimensions, we use HDBSCAN [McInnes et al 2017], a clustering algorithm, to identify and extract clusters. Each cluster is comprised of a set of sentences. To uncover what topic each cluster represents, the sentences are broken into words. Each word in the cluster is given a score determined by the frequency of the word in the cluster, multiplied by the information content (or rarity) of the word in the corpus. This metric is known as c-tf-idf. We take the top 20 scoring words as indicative of the cluster's content. As usual with topic modeling, the total number of topics varies with parameterization.

14 Finally we score each cluster by how *monochromatic* (in our case: purely derived from the "Chinese-Chinese" corpus) and how large it is (i.e. how many sentences are represented in the cluster). Our intuition is that large clusters are more representative of the corpus. Monochromaticity (MC) is expressed by a score between 0 and 1, with values converging to 0 indicating a cluster largely derived from the Indian-Chinese corpus, convergence to 1 indicates a cluster mostly derived from the Chinese-Chinese corpus, and clusters with MC around 0.5 are derived equally from both corpora.

15 Below an attempt to visualize the results of this process:

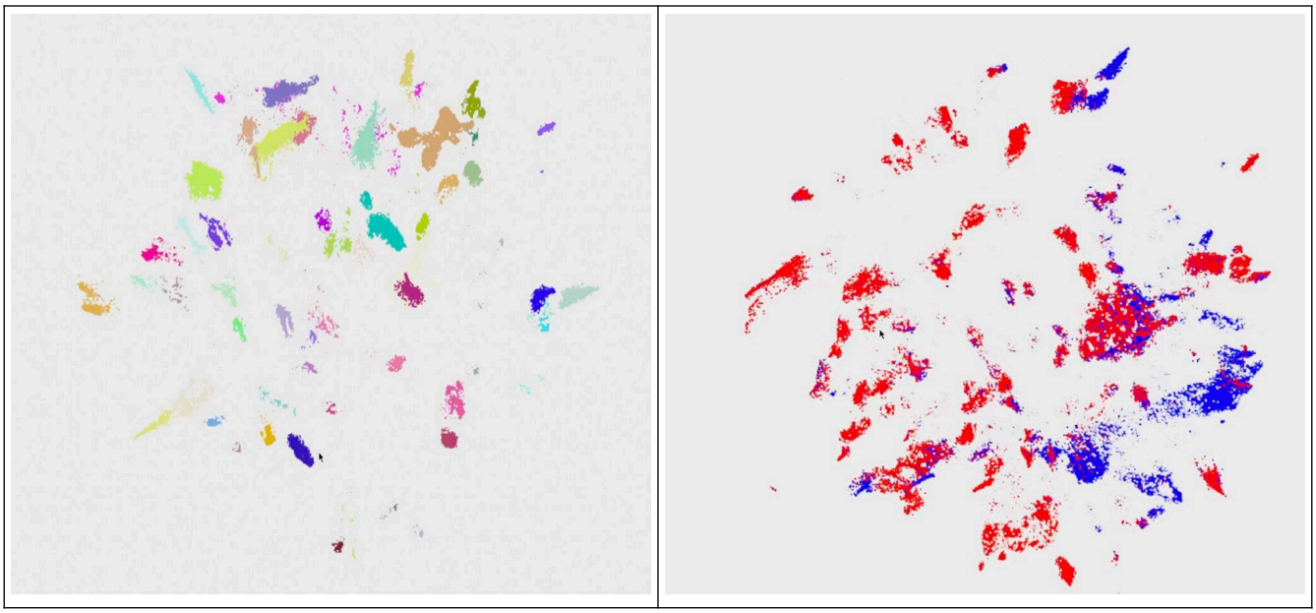

Figure 1. Visualizations of the topics (left) and their chromaticity (right). Both images are projected from 3-dimensional space, which itself is a reduction from a 1024-dimensional vector space. On the left, individual topics are shown with different clusters in (randomly) different colors. On the right, each point in the embedding space is colored red if it corresponds to a sentence from a Chinese-Chinese text and blue if it corresponds to an Indian-Chinese sentence. The images are taken from different viewing angles, thus the displayed clusters do not align.

4. Discussion

16 Orienting ourselves by monochromaticity and cluster size allows us to focus on the "**coherent**" topics that have the best chance at being "**meaningful**" for distinguishing between Indian-Chinese and Chinese-Chinese corpora. We found it useful to distinguish between these two orders of sense-making: We define "coherent" here not by one of the several computed coherence metrics for topic modeling (such as Umass, C-V etc.). In the case of BERTopic, that is already taken care of during the dimensionality reduction via HDBSCAN. Rather, we use human evaluation to decide whether the terms in a topic semantically relate to each other in a way that allows a reader with domain knowledge to string them together in virtual paragraphs (Appendix A) without forcing associations. Domain knowledge remains indispensable for such exercises. On trying out the classic tf-idf formula on Woolf's The Waves, Stephen Ramsay [Ramsay 2011, 12] wrote: "Few readers of The Waves would fail to see some emergence of pattern in this list". It should be added that, conversely, for someone who has not read the text "patterns" are highly unlikely to emerge. A modified form of tf-idf remains part of BERTopic and other topic modeling frameworks, and although the topic coherence as assessed by computable metrics arguably has improved, the need for domain knowledge for "the 'lighting up' of an aspect (das 'Aufleuchten' eines Aspekts) remains. [11]

17 Beyond seeing a coherent pattern in the word lists that represent topics, a second-order of sense making is relevant. "Meaningful" topics must be meaningful not only in themselves (i.e. coherent) but also relevant in the wider context of a research question. Do the topics make heuristic sense for a researcher in Buddhist studies? Are they insightful to the distant reader? In our case, do they actually distinguish the Indian-Chinese from the Chinese-Chinese texts or are they just different? Obviously, coherence here is a condition of meaningfulness. If the top-twenty words that represent a topic do not cohere, there is no topic to discuss in a wider context.

18 The particular iteration from which the ten topics below are taken yielded c. 750 clusters, but there is no natural upper or lower limit for the amount of clusters produced in any given pipeline. How did we select the ten clusters we discuss below, that we have identified as coherent and meaningful? Like with the overall amount of clusters, "ten" too is an arbitrary limit that takes into account what can be comfortable presented to DHQ readers. Arguments could be made for longer discussions of five topics, or a more concise presentation of twenty. Our selection "algorithm" was: Descending

from the most monochromatic clusters (which are most likely to distinguish Indian-Chinese from Chinese-Chinese texts) select the first ten clusters that are both “coherent” and “meaningful” in the sense sketched above. Different researchers of Chinese Buddhism might come up with slightly different sets but that is the not yet the point. Distant reading allows for hermeneutic difference just as close reading. Are there many more clusters that are coherent and meaningful? This is hard to quantify, because “meaningful” implies a concrete research questions. In our case, reading through these lists of topics, our intuition is that about one in ten clusters seems coherent, but of these, only those with high monochromaticity should be expected to *meaningfully* distinguish the Indian-Chinese from the Chinese-Chinese corpus.

19 Below, we will discuss the heuristic value of ten topics suggested by BERTopic. These are presented as ‘virtual paragraphs’ in Appendix A. What signals does a distant reading of our two corpora discern? Among the topics that align strongly with the **Indian-Chinese** corpus the least surprising ones are two that relate to the introduction of tantric, esoteric Buddhism to China: *maṇḍala* (A1) and *mantra* (A2). *Maṇḍalas* were used in Indian esoteric Buddhism in visualization practice, as part of rituals, and as an art form. These three aspects are related as sophisticated *maṇḍala* paintings are not only used in rituals, but also as models for meditators who learn to visualize them as part of their meditative practice. In India, the earliest Buddhist uses of *maṇḍala* images are attested for the 6th century, well within our time-frame. In China, a host of texts on how to use *maṇḍalas* in rituals were translated in the 8th century (e.g. in the Taishō edition (T): T0850, T0852a, T0852b, T0862, T0911, T0912, T0959, T1001, T1004, T1040, T1067, T1167, T1168B, T1184). Texts containing *mantras* or the longer *dhāraṇī* spells (e.g. T0402, T0899, T0901, T0902, T0903, T0905, T0907, T0918, T0933, T0944A, T0952, T0956, T0962, T0963, T0964, T0967, T0968) were already mentioned above as the reason why the Indian-Chinese corpus has twice as many texts but an overall lower character count than the Chinese-Chinese corpus. Neither poetry nor prose, *mantra* and *dhāraṇī* texts are a distinctive genre in themselves. That distinctiveness, not surprisingly, appears in the BERTopic output. Topic A2 illustrates a ‘mantra’ topic; the terms *mantra* and *dhāraṇī* echo through several other topics and indeed also appear together in A5. *Maṇḍalas* and *mantras/dhāraṇīs* are the visual and aural elements of esoteric ritual and meditation practice that was introduced to China in the 8th century (and from there to Japan in the early 9th century). This was the last great transmission of a distinct Buddhist tradition from India to China and topic modeling picks up a clear signal of this process. This proves at the very least that BERTopic works for our corpus, and can identify “typical” Indian-Chinese topics.

20 But there are other, more subtle topics, such as A3, “Yama Heaven,” where things get more interesting to think with. Indian cosmography posits a number of heavens above (and hells below) our “middle earth” (*madhyadeśa*). [12] In Chinese Buddhism between 500 and 800 CE, the writings of Tanluan, Daochuo and Shandao laid the foundation for the Pure Land school in which practitioners aim to be reborn in the heavenly paradise of Amitābha Buddha. Their writings in turn are based on Mahāyāna Indian sūtra texts translated before 500 CE. The “Yama Heaven” topic signals that in the Indian texts translated 500-800 CE the traditional view of “layered” heavens still persisted, and was not yet subsumed into the otherworldly Pure Land of the West that became so extraordinarily influential in East Asian Buddhism in the second millennium. This generates new avenues for research: The texts connected to the topic (e.g. T21n1340, T13n0416, T16n0675, T18n0892, T14n0455, T17n0721) can now be further explored e.g. to see how the heavenly realms in Mahāyāna Indian-Chinese texts differed from Amitābha’s Pure Land extolled in the Chinese-Chinese works in our period.

21 Another Indian topic related to place is A4 “The palace 宮 of Śākyamuni.” Like early Buddhist iconography the topic is in a way aniconic: it describes the city where Prince Siddhartha, the Sage (muni) of the Śākya clan grew up, but names only his father, Śuddhodana, not Śākyamuni Buddha himself. That the monochromaticity of the topic converges to 0 suggests the Buddha legend features highly in Indian-Chinese texts. And indeed, the topic appears not only in a dedicated Buddha legend epic (T0190), but also in (Mūlasārvastivādin) Vinaya texts (T1442, T1443, T1450), and encyclopedic collections (T2121, T2122). The latter are compiled in China, but contain extensive quotations from translated Indian texts. The topic marks an influx of Indian texts which speak of the historical Buddha in a Chinese Mahāyāna environment, where many other texts propagate the existence of a multitude of Buddhas “innumerable as grains of sands in the Ganges.” It is a reminder that the proliferation of Buddhas did not overwrite interest in the story of Śākyamuni, the historical Buddha.

22 Topic A5 reflects an ongoing concern with the “propagation of the Dharma.” That this topic lights up in Indian texts

translated 500-800 CE is perhaps an indication of the pressures Buddhism faced in India. While in China the Sui (581–618), Tang (618–907), and Song (960–1279) dynasties were (mostly) a long golden summer for Buddhism, in India autumn had set in by the 7th century. The invasion and rule of North India by the Alchon Huns in the sixth and early fifth centuries (c. 460-530 CE) was adversarial to Buddhism, and Xuanzang who traveled to India some hundred years later (629-645 CE) found many pilgrimage sites in decline. Besides the last blossoming of Buddhist philosophy in the works of Dharamakīrti and Candrakīrti (7th century), which were not translated into Chinese, the development and transmission of tantric esoteric Buddhism was the last major doctrinal development in Indian Buddhism. As Indian Buddhism slowly lost ground to powerful Hindu movements (Vaishnavism, Shaivaism, Bhakti etc.) the need to teach and propagate the Buddhist teachings to laypeople remained a concern in Buddhist literature.

23 Two highly monochromatic topics are associated with monastic life and its rules, one (A6) tending to the Indian-Chinese corpus, the other (B1) to the Chinese-Chinese corpus. Whereas A6 draws on the more technical discourse of Buddhist canon law, the Vinaya, and its history, B1 is about precepts, the rules that Buddhists ought to follow. Returning to the texts we realize that topic A6 is connected to the translations of Yijing 義淨 (635–713), who went to India and returned with the Mūlasarvastivāda Vinaya and commentaries. Although in the end the authoritative version of the monastic rules in East Asian Buddhism relied on a different Vinaya tradition, A6 can be said to reflect an ongoing exchange between India and China in terms of canon law. Although associated with translated Indian texts the topic does not so much reflect a new concern within Indian Buddhism, but rather an ongoing Chinese interest in the Vinaya. Indeed, later historiographers have asserted the development of a “Vinaya School” 律宗 in seventh and eighth century China. In contrast, B1 is marked as a Chinese-Chinese topic by the peculiar term jieti 戒體, the “essence” of the precepts, which had great traction in the Chinese Vinaya tradition, but for which there is no ready Indian equivalent. Distant reading A6 against B1, one can discern an important polarity which reflects two competing Vinaya traditions in East Asia. Next to the mainstream Indian Vinaya, there were the “Bodhisattva precepts” of an Eastern Mahāyāna Vinaya that formed around the, probably apocryphal [Funayama 1996], *Fanwang jing*梵網經. Both traditions were present in China during and beyond our timeframe in China as well as in Japan and Korea, but are rarely addressed as distinct. The two topics thus do not merely reflect historical reality, but highlight a fundamental distinction in the development of Buddhist norms. In addition, the association of the Fanwang jing vocabulary of B1 with Chinese-Chinese texts supports the claim that this sūtra was indeed apocryphal and topically related to Chinese, not Indian concerns.

24 Regarding topic B2, “Early Translators,” one might at first be tempted to think that any “translation” topic (there are other, less monochromatic ones), might be due to paratext, such as the translator byline that precedes most fascicles of an Indian-Chinese text, but obviously this topic is aligned with the Chinese-Chinese corpus. Moreover, the topic is specifically about “early” translators, active from the second to the fourth centuries, before our timeframe. It therefore cannot reflect paratext, but rather a specific Chinese concern with Buddhist historiography, a fundamental difference between Chinese and Indian Buddhism in any period. Chinese Buddhists made use of Chinese historiographic genres to create catalogs, biographies, annals and more. Compared with India, the historical record for Chinese Buddhism of the first millennium is extremely rich and detailed. Although this might be better understood as a general cultural trait, not specific to Buddhism, topic B2 points to the role Buddhist historiography played in the formation of a distinct Chinese Buddhism.

25 Another “typical” Chinese topic is “Pillars of the state”(B3) which clusters the titles of Chinese government officials, none of which would appear in an Indian text, where the administrative hierarchy was much less sophisticated. Which texts are responsible for the topic? The first two texts associated with it are anthologies of apologetic writing (T2110, T2103) where Buddhists are in debate with officials, friendly or adversarial. Next is the encyclopedic Fayuan zhulin 法苑珠林 (T2122), which, like T2110, too was compiled by Falin 法琳 (571-639). Then there are scriptural catalogs (T2156, T2157) and biographies (T2051 (again Falin), T2060). Thus a closer look at the sources of the terms that represent Topic B3 encourages us to reevaluate Falin’s role in the sinicization of Buddhism [Jülch 2014].

26 A final monochromatic topic associated with the Chinese-Chinese corpus is B4 which seems to built around collocations of *sheng* 乘 ‘vehicle’ (Sk. *yāna*). The topic is meaningful in that it reflects a major concern in medieval Chinese Buddhism, namely the categorization of Buddhism in different traditions. After c. 400 CE Chinese Buddhists increasingly understood themselves as followers of Mahāyāna, a distinctive new development within Indian Buddhism. While in India

Mahāyāna Buddhism developed further into esoteric or tantric Buddhism (Vajrayāna, Tantrayāna), Chinese Buddhism continue to define itself as Mahāyāna, although, as we saw above, was exposed to esoteric Buddhism in our time frame. In the *sheng* 乘 imaginaire earlier Indian mainstream Buddhism was cast as the 'smaller (or 'lesser') vehicle' Hīnayāna, in contrast to the 'larger (or 'greater') vehicle', the Mahāyāna. Topic B4 reflects this coming to terms with different counts of yānas, such as one (the single transcendent truth applicable to all), two (Hīnayāna and Mahāyāna), or five yānas (the teachings of non-Buddhist humans, deities, śrāvaka Buddhists, pratyekabuddhas, and bodhisattva Buddhists). Although not used in academic narratives of Buddhist history, these distinctions are still of interest to modern Chinese Buddhists and BERTopic is able to identify this concern in the texts produced between 500 and 800 CE.

5. Conclusion

Topic modeling is a real upgrade on divination. Traditional divination systems set up a randomized procedure that selects symbolic tokens from a given set. It is then left to the diviner to interpret the tokens for the question at hand. Whereas with divination systems the number of elements in the set is usually fixed (78 cards in the Marseille Tarot, 64 Yijing hexagrams, 12 zodiac signs etc.), topic modeling builds its topics as it goes along and can return any number of them. However, like divination, it relies heavily on the interpreter to perceive the internal coherence as well as the overall meaning of a topic. The output of topic modeling is certainly grounded in more sophisticated probabilistic methods than the randomness of lot drawing, but its susceptibility to parametrization and the fact that different approaches yield somewhat different topics, lend the procedure a vagueness and ambivalence not unlike the hermeneutic play of divination. In the BERTopic pipeline a degree of randomness, specifically in the UMAP dimensionality reduction, means that even with the same data and parameters, the top twenty words in a topic may be slightly different. As Benjamin Schmidt (2012) warned at the beginning of the DH topic modeling boom, topics are neither necessarily coherent nor stable. There is no surefire method to separate a "true" topic from a "fata morgana topic". And whether a topic is meaningful in the context of a research question can (for now) only be decided by humans. This being said, our experiments with different BERTopic workflows, give us confidence that the latent topics are stable at least within this framework. Moreover, as argued in Section 4, our experiments indeed turned up coherent and meaningful topics. These topics mark differences between the Indian-Chinese and the Chinese-Chinese corpus, in spite of BERTopic working on a low-resource idiom (Buddhist Chinese). The caveats are: First, there is no guarantee that we have identified all relevant topics – the rate of recall is unknown and perhaps unknowable. Second, there is no guarantee to precision either, as other approaches might turn up different, but equally plausible topics. 27

To counter these limitations to a degree, we are working on a follow-up study in which we use machine translated versions of the texts instead of the Chinese originals. The approach might confirm or undermine the stability of topics. Furthermore automating the production of "virtual paragraphs" might lead to greater involvement by the wider community of scholars interested in the corpus. In the end, to us the glass is half full: Our modified BERTopic workflow yields coherent topics that meaningfully distinguish Indian-Chinese and Chinese-Chinese texts in our time frame. Some of the topics are more surprising than others, but even those that at first glance seem obvious, such as the "translation" topic (B2) associated with the Chinese-Chinese corpus, turned out to be more complex than expected. Since all topics can be traced back to texts, it is possible to follow up and study them further. Thus the distant reading via topic modeling can nudge our close reading and research in directions it would not have taken otherwise. 28

Acknowledgments:

Research on this paper was funded by the *Anatomy of Agency* project (Nichols, PI; 1003) via the "Launching Experimental Philosophy of Religion" grant (Ian Church, PI; 61886), itself supported by the John Templeton Foundation. We give special thanks to Robert Buswell, Nicholaos Jones, Song Wang, and Michael Radich for comments and correspondence leading to the improvement of this project. 29

The research was in part conducted at the Asia Research Institute (ARI) at the National University of Singapore (NUS). The authors are grateful for the support of ARI and the Department of Chinese Studies at NUS. 30

Appendix A: Topics as virtual paragraphs

In order to communicate our findings to those who have neglected their study of medieval Chinese Buddhist texts in recent years, we have in the following created “virtual paragraphs” for ten of the more monochromatic and coherent topics that were stable across different iterations of our BERTopic pipeline. The current workflow outputs more than 750 topics, so this is only a small sample. All the Chinese terms are as they appear in the topic and follow their English translation. Their syntactic relationship is invented freely, but informed by our domain knowledge, i.e. they are used in the context of medieval Buddhism. The idea is to use such artificial paragraphs to demonstrate to a non-specialist readership that these topics indeed constitute intelligible semantic fields. The numbering does not imply any ranked order.

A. Indian-Chinese Corpus

A1 Maṇḍala曼拏羅 Maṇḍalas 曼拏羅 depict a variety of Buddhist deities from various tantric families 種族 such as Vajrapāṇi 金剛手 or Vajrasattva 金剛薩埵, and Amoghapāśa 不空羂索 or Amogharāja 不空王/施暮伽王 Bodhisattva. By drawing maṇḍalas, together with the secret 祕密 transmission of the esoteric teachings 密迹/密跡主 and the help of *dhāraṇī* 陀羅尼 spells and mantras 真言 practioners can achieve adamantine 金剛 samādhi 三昧(耶). Maṇḍalas are often used in rituals on or around altars 壇.

A2 Chanting 誦 Mantras 真言 Three times circulambulating 三匝 altars, stupas, or images one chants (e.g.) the Mantra 真言 of the King of Fury 奮怒王. One recites rapidly 緊捷誦, sometimes using the 108 一百八 beads of the rosary 念珠. One recites and memorizes 誦持 a thousand times 一千 遍, imprinting the mind-seal 心印, until one can recite the secret 誦密 spells in one’s mind 誦心 / 祕密心, thereby creating and gaining blessings 加持.

A3 In Yama Heaven 夜摩天 The King of Yama Heaven 夜摩天 is *Musulundha 牟修樓陀. [13] There the goddesses 天女 play 遊戲 in the parks 園林, singing and dancing 歌舞, between ponds 池 and groves 林. The mountains 山峯 are full of many-colored 雜色 birds 鳥. The gods in Yama Heaven sport 娛樂 in an lovely 可愛 environment filled with the seven precious materials 七寶 such as silver 白銀, beryl 毘琉璃, crystal 頗梨. There also are bees 蜂.

A4 The palace 宮 of Śākyamuni A palace 宮 in a walled city 城 with high ministers 大臣. Where the great king 大王 Śuddhodana 淨飯王 lives with his wifes 妻/婦 and crown prince 太子. In the city lived merchants 商人/商主, householder women 女, and elders 長者. That time 是時 a seer 仙人 came to the palace to tell 告 the king father 父王 Śuddhodana about the fate of the prince 王子.

A5 Propagating 演說 the Dharma Teachings are explained 演說 by a good friend 善知識, a guiding teacher 導師, endowed with rhetorical skill 辯才. It is done at Buddhist temples 佛剎 to an audience of sentient beings 眾生, including good women 善女人, who have approached 親近 the teacher. It is done in detail 廣大, reaching everywhere 十方/普入, for all possessed with wisdom 智慧, in order to tame the minds 調伏 of sentient beings, benefit 利益 them, and make them to rid 捨離 themselves of suffering. The teaching of a omniscient 一切智 Buddha or Bodhisattva demonstrates 示現 even that which cannot be explained 不可說.

A6 Sangha 僧伽 life The members of the Sangha 僧伽, monks 苾芻, nuns 苾芻尼, and others seeking nirvāṇa 求寂 begin their life of training 學處 under a preceptor 鄔波馱耶. Different from lay people 俗人, they have left their secular family 俗家, and are allowed only few possessions such as their alms bowl 鉢 and some bedding 臥具. If they commit a pārājika offense 波羅市迦, however, they are expelled. Some famous monks had known the Buddha even before he went forth, such as Ānanda 阿難陀 and Chandaka 闡陀. There also were trouble makers such as the monk Upananda 鄔波難陀 or the nun Sthūlanandā 吐羅難陀, who were without shame 嫌恥 or remorse 惡作.

B. Chinese-Chinese corpus

B1 Keeping the precepts 戒/禁戒/戒法 All Buddhists take refuge in the three 三歸 jewels. Lay-people take five, eight or ten precepts 五戒, 八戒, 十戒. To become a monastic one must take the precepts 受戒, the

complete set of precepts 具戒, as found in the prātimokṣa 戒本, the list of precepts, as they have been explained 說戒 by the Buddha. The ordination takes place in a marked ordination site 戒場. Once one has taken the pure precepts 淨戒 one must keep the precepts 持戒 is bound by the precepts 結戒. One may not break the precepts 犯戒/破戒. Otherwise one cannot experience their essence 戒體 that develops in one as one practices. One should never give them up 捨戒 again.

B2 Translators and translation 譯 At the time when the Indo-Scythians 月支 ruled Northern India, the early translations 譯 of Buddhist texts began in China in the final decades of the Latter Han 後漢 Dynasty with An Shigao 安世高/安公. It continued later in the Kingdom of Wu 吳 (222-80) with Zhi Qian 支謙 (fl.222–252 CE) and under the Western and Eastern Jin 西晉/東晉 (266–316 / 317–420) with Dharmarakṣa 竺法護 (c. 233-310). They wrote their translations on sheets of paper 紙, wrapped in bundles 帙, or glued together into scrolls 卷. Later catalogs 錄 like that of Sengyou 僧祐 (445–518) recorded not only first translations but also alternative translations 異譯 of the the same text and translation of which the translator is unknown 失譯.

B3 Officials as "pillars of the state" Officials are the "pillars of the state" 柱國: Next to the civil administration with its Grand Counselors 侍中, the Director of the Department of State Affairs 尚書令, the Vice Directors 僕射 and ministers 尚書 and members of the Secretariat 中書, the Minister of Education 司徒, the vice ministers 侍郎侍郎, the provincial governors 太守, the secretaries 內史 and aides 長史, and the inspectors 刺史 of the censorate, there are the military 軍事 officials such as the military provincial governors 都督, the field marshals 大將軍 and generals 將軍, the 總管 regional commanders and commanders 司馬 and adjutants 參軍.

B4 How many vehicles/yānas 乘? There is the Hīnayāna 小乘 and the Mahāyāna 大乘 both 大小乘. There are Hīnayāna scriptures 小乘經 and Mahāyāna scriptures 大乘經. But is that all? Aren't there perhaps five vehicles 五乘 or three vehicles 三乘, instead of these two vehicles 二乘? Or is there perhaps only the teaching of one single vehicle 一乘教/一乘法, a single Buddha vehicle 一佛乘.

Appendix B: Workflow - Technical Details

Below the specific parameters that we used in our instantiation of the BERTopic pipeline. All computational work was performed on a Gentoo Linux workstation with a Xeon W-2295 CPU, an Nvidia A6000 GPU and 512GB of memory. 32

Our corpus was already pre-segmented, we therefore did not use automatic segmentation tools like jieba. As explained above, the segmentation was performed by researchers at the Dharma Drum Institute of Liberal Arts as part of a project on using conditional random fields for automatic segmentation (for details see Wang 2020). The sentence embeddings are then produced using Yasuoka's version of GuwenBERT. This itself was wrapped in a SentenceTransformers instance to extract the embedding of the [CLS] token to get an embedding that represents the sentence as a whole [Reimers et al 2019]. This produces 2037769 embedded sentences, each of which in 1024-dimensional space. 33

From here, we apply UMAP to create 3-dimensional embeddings which respect relative closeness of embedded sentences by projecting the 1024-dimensional points onto a 3-dimensional manifold. For this, we use the CUDA-enabled version of UMAP provided by the NVIDIA cuML library [Raschka et al 2020]. We used the default parameters for the reduction, only specifying that the output should be 3-dimensional. 34

For the clustering, we used hdscan, again using the CUDA-enabled version from cuML. We set three parameters for the clustering algorithm, using 750 for the minimal number of clusters 100 for the minimal number of samples. The first parameter indicates that we want the algorithm to produce at least 750 clusters, while the second indicates that very small clusters (greater than 100 sentences) should not be counted. 35

Notes

[1] The fact that the method has survived the now defunct journal, is an indication of the fluid, evolving nature of DH as a field.

[2] Data for this network is available as the *Historical Social Network of Chinese Buddhism* https://github.com/mbingenheimer/ChineseBuddhism_SNA.

[3] Our corpora are available here: https://github.com/mbingenheimer/cbetaCorpusSorted.

[4] A large part of the CBETA corpus of Buddhist texts is available in this tokenized form here: https://github.com/DILA-edu/word-segment (accessed 2024-09).

[5] The high-brow, civil examination style of classical Chinese that was modeled on the thirteen Confucian classics and other works of the Zhou, Qin, and Han dynasties. While this style of writing was used until the early 20th century in certain genres, other, more vernacular, styles existed that were closer to spoken Chinese where many words are multimorphemic.

[6] As of 2023-11 a web of science search returned 49 articles and 13 proceeding papers with “BERTopic” in the title or abstract.

[7] The pre-segmented version was originally produced by Yu-chun Wang and forked from the larger repository (which includes more texts from the Taishō canon) at https://github.com/DILA-edu/word-segment. The version we used for the experiment is available (together with our code) at https://github.com/mbingenheimer/cbetaCorpusSorted/tree/main/bertopic/word-segment-main.

[8] At: https://huggingface.co/KoichiYasuoka/roberta-classical-chinese-large-char (accessed 2023-12-10). GuwenBERT at: https://huggingface.co/ethanyt/guwenbert-base (accessed 2023-12-10).

[9] https://huggingface.co/ethanyt/guwenbert-base (accessed 2023-12-10). Slightly different figures are given at: https://www.kaggle.com/datasets/wptouxx/daizhige (accessed 2023-12-10). The Daizhige dataset is floating around on the web, slightly unmoored. One accessible version is on Github: https://github.com/garychowcmu/daizhigev20 (accessed 2023-12-10). As one of the largest datasets of classical Chinese it deserves to be better curated.

[10] We did experiment with fine-tuning the model based on a Korean model which had been fine-tuned for semantic similarity, but the results were not an improvement. This will be a topic of future research.

[11] This is how Wittgenstein describes the moment when we recognize a hitherto unseen pattern (Wittgenstein Philosophical Investigations ¶118, ¶140).

[12] See [Kirfel 1920] for an overview. For a distant corpus based reading of the role of gods and deities in the classical Chinese tradition see Nichols et. al (2020).

[13] Musulundha as Sanskrit for 牟修樓陀 is suggested by [Lin 1949, 61].

Works Cited

Bingenheimer 2021 Bingenheimer, M. (2021) “The historical social network of Chinese Buddhism”, *Journal of Historical Network Review*, 5: 233-257. Available at: https://doi.org/10.25517/jhnr.v5i1.119

Blei 2012 Blei, D. M. (2021) “Topic modeling and digital humanities”, *Journal of Digital Humanities* , 2.1. Available at: https://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/. (Accessed November 2023.)

Blei et al2003 Blei, D. M, Ng, A. and Jordan, M. (2003) “Latent dirichlet allocation”, *Journal of Machine Learning Research* , 3: 993-1022.

CBETA ver 2021 CBETA v. 2021: Chinese Buddhist Electronic Text Association (2021) “There are different, evolving versions of this digital corpus. We were using the texts as made available here”: Available at https://github.com/DILA-edu/CBETA_TAFxml.

Chen et al 2023 Chen, Y., Zhao Peng, S.H.K., and Chang, W. C. (2023) “What we can do and cannot do with topic modeling: A systematic eeview”, *Communication Methods and Measures*, 17-2: 111-130.

Du 2019 Du, k. (2019) “A survey on lda topic modeling in digital humanities”in: Abstract for *Digital Humanities* (Utrecht). Available at: https://doi.org/10.34894/H9UYPI.

Egger and Yu 2022 Egger. R. and Yu, J. (2022) “A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts”, *Frontiers in Socieology*. Available at: https://www.frontiersin.org/journals/sociology/articles/10.3389/fsoc.2022.886498/full.

Fang 2023 Fang, Y. (2023) “Theme classification of the complete Song Ci from the perspective of the digital humanities”, *Lecture Notes on Language and Literature*, 13-24. Available at: https://clausiuspress.com/assets/default/article/2023/08/17/article_1692280542.pdf.

Fu et al 2020 Fu, Q., Yufan, Z., Jiaxin, G., Yushu, Z., and Xin, G. (2020) “Agreeing to disagree: Choosing among eight topic-modeling methods”, *Big Data Research*. Available at: https://doi.org/10.1016/j.bdr.2020.100173.

Funayama 1996 Funayama, T. 船山徹 (1996) “Gikyō Bonmōkyō seiritsu no shomondai 疑経『梵網経』成立の諸問題”, *Bukkyō shigaku kenkyū 佛教史學研究*, 39-1: 54-78.

Grootendorst 2022 Grootendorst, M. (2022) “BERTopic: Neural topic modeling with a class-based TF-IDF procedure”. Available at: https://arxiv.org/abs/2203.05794. (Accessed 11 March 2022).

Jülch 2014 Jülch, T. (2014) *Bodhisattva der Apologetik - die Mission des buddhistischen Tang-Mönches Falin*. München:

Herbert Utz Verlag.

Kherwa and Bansal 2019 Kherwa, P. and Bansal, P. (2019) “Topic modeling: A comprehensive review”, *ICST Transactions on Scalable Information Systems*, 7. Available at: https://eudl.eu/doi/10.4108/eai.13-7-2018.159623.

Kirfel 1920 Kirfek, W. (1920) *Die Kosmographie der Inder nach den Quellen dargestellt*. Bonn und Leipzig: K Schroader.

Lin 1949 Lin, Li-kouang 林藜光 (1949). *L’aide-mémoire de la Vraie Loi (Saddharma-smrtyupasthâna-sùtra). Recherches sur un Sûtra développé du Petit Véhicule*. Paris: Adrien-Maisonneuve.

Ma et al 2021 Ma, P., Qing Zeng-Treitler, Nelson, Stuart J. (2021) “Use of two topic modeling methods to investigate covid vaccine hesitancy”, *Proceedings 14th International Conference on ICT, Society and Human Beings (ICT 2021), the 18th International Conference Web Based Communities and Social Media (WBC 2021)*: 221-226.

McInnes et al 2017 McInnes, L., Healy, J., and Astels S. (2017) “HDBSCAN: Hierarchical density based clustering”, *Journal of Open Source Software* , vol. 2.11: 205.

McInnes et al 2018 McInnes, L., Healy, J., and Melville, J. (2018) *Umap: Uniform manifold approximation and projection for dimension reduction*. Available at: https://arxiv.org/abs/1802.03426.

Meeks and Weingart 2012 Meeks, E. and Weingart, S.B. (2012) “The digital humanities contribution to topic modeling”, *” Journal of Digital Humanities* , 2.1. Available at: http://journalofdigitalhumanities.org/2-1/dh-contribution-to-topic-modeling. (Accessed November 2023).

Nichols et al 2020 Nichols, R., Slingerland, E., Nielbo, K. L., Kirby, P., and Logan, C. (2020) “Supernatural agents and prosociality in historical China: Micro-modeling the cultural evolution of gods and morality in textual corpora”, *Religion, Brain & Behavior*, 1-19. Available at: https://doi.org/10.1080/2153599X.2020.1742778.

Pew Research Center 2023 Pew Research Center (2023) “Measuring religion in China”. Available at: https://www.pewresearch.org/religion/2023/08/30/measuring-religion-in-china/. (Accessed November 2023).

Ramsay 2011 Ramsay, S. (2011) *Reading machines: Toward an algorithmic criticism*. Urbana: University of Illinois Press.

Raschka et al 2020 Raschka, M., Antons, D., Joshi, A.M., and Salge, T. (2020) “Machine learning in Python: Main developments and technology trends in data science, machine learning, and artificial intelligence”. Available at: https://arxiv.org/abs/2002.04803.

Reimers et al 2019 Reimers, N. and Girevich, I. (2019) “Sentence-BERT: Sentence embeddings using Siamese BERT-Networks”, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019)*.

Rüdiger et al 2022 Rüdiger, M., Antons, D., Joshi, A.M., and Salge, T.O. (2022) “Topic modeling revisited: New evidence on algorithm performance and quality metrics”, *PLoS ONE* , 17(4): e0266325. Available at: https://doi.org/10.1371/journal.pone.0266325.

Schmidt 2012 Schmidt, B. (2012) “Words alone: Dismantling topic models in the humanities”, *Journal of Digital Humanities* , 2.1. Available at: http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt.

Underwood 2017 Underwood, T. (2017) “A genealogy of distant reading”, *DHQ: Digital Humanities Quarterly*, vol. 11-2.

Vayansky and Kumar 2020 Vayansky, I. and Kumar, S.A.P. (2020) “A review of topic modeling methods”, *Information Systems*, 94.

Wang 2020 Wang, Y.D. (2020) “Word segmentation for classical chinese Buddhist literature”, *Journal of the Japanese Association for Digital Humanities*, 5-2: 154-172. Available at: https://doi.org/10.17928/jjadh.5.2_154.