


Developing Computational Models for Formalizing Concepts in the British Colonial India Corpus

Shanmugapriya T <shanmu_dot_shanmugapriya_at_utoronto_dot_ca>, University of Toronto Scarborough 
<https://orcid.org/0000-0002-1524-429X>

Abstract

The concepts embedded in humanities materials are unstructured and possess multifaceted attributes. New insights from these materials are derived through systematic qualitative study. However, for the purpose of quantitative analysis using digital humanities methods and tools, formalizing these concepts becomes imperative. The functionality of digital humanities relies on the deployment of formalized concepts and models. Formalization converts unstructured data into a more structured form. Concurrently, models function as representations created to closely examine the modeled subject, while metamodels define the structure and properties of these models. In this case, the absence of formalized concepts and models for studying the British India colonial corpus hampers the computational application to address humanities research questions quantitatively. The texts are intricate, and the format is non-standard, as colonial officials documented extensive information to govern and control the colonized people and land. In this scenario, the British India colonial corpus cannot be effectively utilized for topic-specific research questions employing advanced text mining without formalizing the concepts within it. This article addresses the questions of what the most effective approach is for identifying multifaceted concepts within the non-standard British colonial India corpus through models and how these concepts can be formalized using formal models. It also explores how metamodels can be developed based on this experiment for a similar corpus.

1. Introduction

The concepts embedded in humanities materials are unstructured and possess multifaceted attributes. New insights from these materials are derived through a systematic qualitative study. However, for the purpose of quantitative analysis using digital humanities (DH) methods and tools, it becomes imperative to formalize these concepts. The functionality of DH is contingent upon the deployment of formalized concepts and models. Formalization converts unstructured data into a more structured form and is applied to various fields such as computer science, medicine, and mathematics. Concurrently, models serve as representations created to closely examine the modeled subject, while metamodels define the structure and properties of these models. Fundamentally, metamodels provide a framework for the analysis and construction of models that are applicable to and beneficial for a predetermined class of problems [Tomasi 2018, 173]. Formalization and models constitute essential elements in both applied and theoretical DH. Michael Piotrowski defines applied DH as concerned with constructing formal models for phenomena studied by their “mother disciplines” (e.g., digital history and digital literary studies) and their methodology [Piotrowski 2022a, 3]. Theoretical DH, on the other hand, studies the general properties of formal models in the humanities at a higher level of abstraction. The objectives of these sub-fields involve theoretical DH creating and studying metamodels, while applied DH apply these metamodels to research questions [Piotrowski 2022a, 3]. Documentations outlining the advantages, disadvantages, challenges, and constraints of algorithms are crucial for constructing metamodels.

Piotrowski discusses formalization and models in the humanities and DH. He contends, based on the works of Gilles-Gaston Granger, that formalization in humanities not only extracts structure from the unstructured through hermeneutic interpretation but also invents “new structures” [Piotrowski 2022a, 11]. Similarly, some disciplines, such as linguistics in

humanities, have been using models to study grammar [McCarty 2004]. On the contrary, formalization and modeling in DH are used to computationally operationalize “scholarly arguments” [Ciula et al. 2018, 347]. In order to derive “new structures” [Piotrowski 2022a, 11] especially for domain-specific humanities materials using computational study, it is imperative to formalize and model the materials. The absence of formalized concepts and models for studying the British India colonial corpus, in this regard, hampers the computational application to address humanities research questions quantitatively.

Since “mapping and surveying have been seen as instrumental in extending” the colonizers control over land to establish their government [Ehrlich 2023, 193]. This resulted in transcending the survey from military and geographical to granule details of “natural history, political economy, and every conceivable species of inquiry into native society” [Ehrlich 2023, 196]. The documents hence are intricate, and the format is rather non-standard as they are “chaotic, if not anarchic, character” [Edney 1997, 162]. In this scenario, the British India colonial corpus cannot be effectively utilized for topic-specific research questions employing advanced text mining without formalizing the concepts within it. For instance, in the corpus curated to study water-related features in the Madras Presidency^[1] in British colonial South India using DH methods, data about water is not explicitly provided in most texts unless they are specifically dedicated to hydraulics. Subsequently, extracting information requires strenuous close reading due to the presence of heterogeneous concepts in these texts.

For instance, *Madras District Manual Coimbatore Volume II* (1898) contains diverse data about population, religion, caste, marriage, and many other details, along with information about water features in Coimbatore. While it serves as a comprehensive manual for the Coimbatore district, the challenge lies in extracting specific water-related data from this text. Disregarding other potentially relevant data is not feasible, as it may have connections with water-related information. Identifying such connections or potential concepts for topic-driven questions poses a challenge. Apart from this text, several others, such as *Handbook Of The Madras Presidency* (1879), *Madras District Gazetteers Coimbatore District* (1880), and *Report On The Administration Of The Madras Presidency During The Year 1869–70* (1870), encompass a variety of concepts, including attributes relevant to water. During the curation of these texts, keyword searches were employed, using terms like Coimbatore, water, canal and river etc. While this method revealed information related to the specified keywords, relying solely on a corpus curated through simple keyword searches presents challenges. As pointed out by Oberbichler and Pfanzelter, the complexity of language “characterized by ambiguity and concepts that are difficult, if not impossible, to trace by computational methods and thus keyword searches alone” [Oberbichler et al. 2021, ¶18]. Therefore, formalizing the multifaceted concepts within the texts is imperative for further computational study. In this article, the questions I ask are as follows: What constitutes the optimal approach for identifying multifaceted concepts within the non-standard British colonial India corpus through models, and how can these concepts be formalized using formal models? How can metamodels be developed based on this experiment for a similar corpus? The aim of this article is to propose proficient concept-based models to formalize heterogeneous concepts from the non-standard British colonial India corpus using computational models, and subsequently, to establish metamodels through experimental methods for the construction and analysis of analogous corpora.

This article is divided into six sections following this introduction. Section 2 focuses on the conceptual theoretical framework for constructing computational models to formalize concepts. It elucidates how models can be constructed using the infrastructure of the texts through concept diagrams and illustrates the model through a theoretical example. Section 3 outlines the methodology of the experiment, encompassing preprocessing, the application of computational models, and the visualization of the extracted models. Section 4 examines the extracted models by illustrating selected texts from the corpus to assess the efficacy of the proposed models. It discusses two kinds of outcomes from the experiment: one delineates the distribution of the concepts both in sub-models and primary models and the other explores the features of the models. This section also includes the disadvantages and challenges of the experiment. Section 5 delves into the metamodels using concept diagrams to develop a similar domain-specific non-standard corpus. Section 6 presents the conclusion of this article.

2. Creating computational models to formalize concepts in the British

3

4

5

India colonial corpus

Corpus building is an integral part of the machine learning process, especially when dealing with a large amount of text. Oberbichler and Pfanzer discuss the rewarding nature of identifying patterns for research questions in big data using quantitative methods; however, it oftentimes comes with its own intriguing challenges “to find and extract those parts in the massive data dumps that are relevant for the topic in question” [Oberbichler et al. 2021, ¶6]. As they say, identifying conceptual patterns within the British India colonial corpus is notably challenging owing to its intricate content infrastructure. This infrastructure exhibits variations across texts and incorporates a multitude of concepts and topics which hamper the application of formal models. Given that my research focuses on extracting information about water features in the corpus using computational methods, my inquiry revolves around understanding the discourses related to water and water-related features. Specifically, I aim to investigate the transformation of water bodies in the Coimbatore region. The goal of constructing this corpus is to present pertinent headings, sections of texts, and complete texts that address my research question. This will facilitate further quantitative and qualitative analysis.

6

Relevant methodologies for constructing and analyzing corpora based on concepts, topics, indices, and representativeness can be found in DH literature [Jähnichen 2017] [Englmeier et al. 2021] [Oberbichler et al. 2021] [Verheul et al. 2022]. Nevertheless, building and applying formal models to examine British India colonial corpus has never garnered attention within the realm of DH. The concepts embedded in this historical corpus not only hold significance for comprehending the past but also aid in establishing connections between the present and past. The proposed methodology is designed to extract both primary and secondary heterogeneous concepts. While the latter may not always be the primary focus of British India colonial manuscripts, it remains relevant for topic-oriented research.

7

Discussions concerning models are pivotal in the field of DH [McCarty 2004] [McCarty 2005] [Buzzetti 2002] [Beynon et al. 2006] [Flanders et al. 2015] [Ciula et al. 2018]. William McCarty defines a model as “a representation of something for purposes of study or a design for realizing something new” [McCarty 2004]. This definition draws upon Clifford Geertz’s analytical differentiation between a model (representation) of something, exemplified by a grammar describing the features of a language, and a model (design) for something, akin to an architectural plan providing design [McCarty 2004]. Typically, models inherently exhibit simplification and shed light on previously unknown aspects. However, models within the humanities often maintain only partial explicitness, commonly expressed informally through natural language [McCarty 2004] [Piotrowski 2019]. In contrast, the natural and engineering sciences extensively employ explicit and formal mathematical models [Epstein 2008] [Piotrowski 2019]. In the realm of computing, as Brian Smith, cited by McCarty, emphasizes, formal “models are fundamental,” running by manipulating representations, always formulated in terms of models [McCarty 2004]. Concerning formal models, the term “formal,” as elucidated by Piotrowski, signifies being “logically coherent + unambiguous + explicit” and in the domain of DH, a requisite degree of formalization is crucial to enable models to be processed and manipulated by computers, referred to as computational models [Piotrowski 2019]. Additionally, a metamodel, employing graphical representation and natural language to elucidate the criteria and parameters essential for the computational process, can inform computational model to improve the efficiency of the process. These metamodels are methodically transformed into computable implementations through varying levels of formality in modeling [Ciula et al. 2018]. While metamodeling remains relatively unexplored in DH, modeling, in general, is regarded as one of the fundamental research practices in DH. Nevertheless, the goal of a computational/formal model in DH is to mitigate complexity and uncertainty through the logical structure of the model.

8

In similar to the definition of model, the definition of a concept varies across disciplines, with philosophers viewing it as an inquiry into changes manifested through various events, linguists identifying concepts in onomasiological words and mapping conceptual changes over different periods, and historians exploring various concepts and their significant changes throughout periods in corpora using computational models [Brigandt 2010] [Linguistic DNA] [Verheul et al. 2022]. In this context, the concept, as defined for this specific study, can be inferred from the organization of words and their contextual relationships. Nonetheless, to identify, trace and inquire concepts, first understanding and identifying them in the corpus is imperative. In this case, the formal model plays a crucial role in formalizing and extracting concepts through its logical framework which interconnects the former and the latter. The concept-based formal model

9

endeavors to capture meaningful concepts by processing raw data, facilitating organization, correlation, and the establishment of a semantic network among concepts. This, in turn, enables more nuanced and context-aware analysis or decision-making. In this case, my goal is to formalize and extract the heterogeneous concepts hidden and layered in the British India colonial manuscripts using computational models.

To formalize the concepts, given the non-standard format of the corpus, the models can be divided into three kinds: sub-model, primary model, and larger model, as the concepts are layered at various levels in the texts, as demonstrated in the concept diagrams in Figures 1 and 2. The concepts embedded in the content of each heading are classified as sub-models and those present in the entire text are labeled as primary models (see Figure 1). The interconnection between these models can be used to develop the larger model, as the similarity network drawn in the diagram (see Figure 1). For a more nuanced comprehension of the model, let us study a hypothetical elucidation based on my research about the water in the colonial period. In Figure 2, number one enclosed in a red circle signifies the interconnection between the primary model “rivers and channels” and a sub-model “waterbodies and canal company” derived from text 3 and 1; number two delineates the correlation between a sub-model “water scarcity” and a sub-model “water infrastructure” identified in text 3 and 1; number three establishes a connection between a sub-model “public and investment” and and sub-model “dam construction” found in text 1 and 3; number 4 links a sub-model “water tax” and the primary model “waterbodies and canal company” observed in text 2 and 1, while number 5 designates the non-connected primary model of text 2 with any of the models.

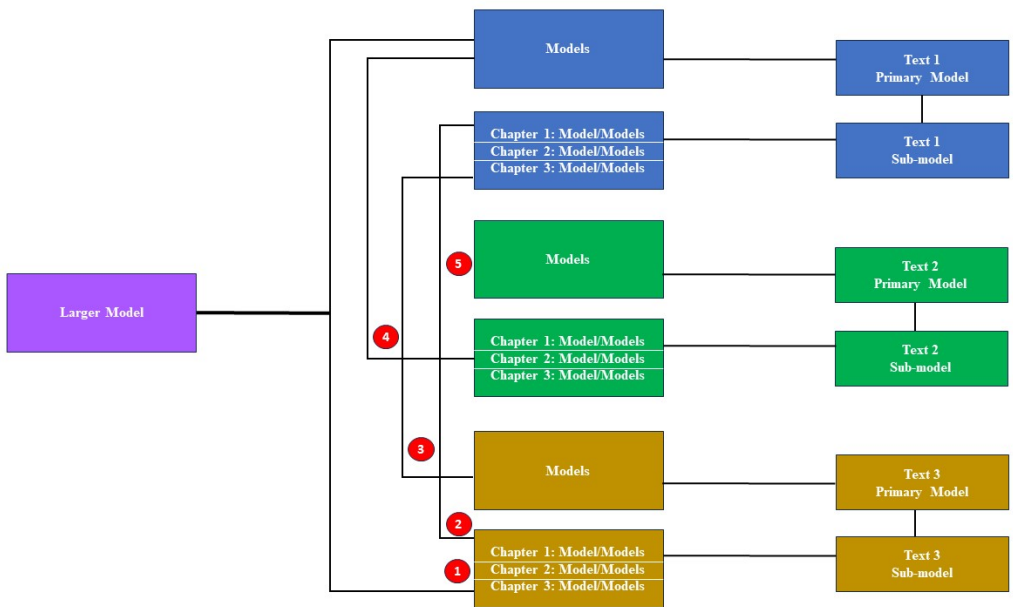


Figure 1. Concept-based model representing the features of both sub-model and primary model

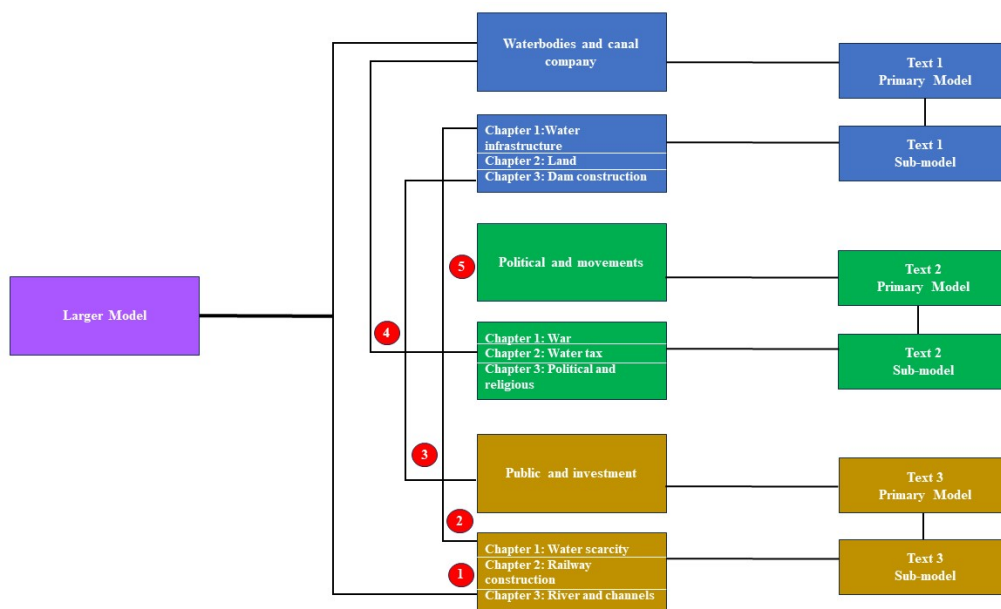


Figure 2. A hypothetical elucidation of concept-based model

During the British colonial period, the concept of water extended from the primary resource for drinking, farming, and ritual entity to a taxable resource and expandable colonial revenue. Such an attitude has resulted in focusing on “developing irrigation infrastructure through building larger dams” [Saravanan 2020, 37]. Ratna Reddy says “[t]he interest of the British Government in increasing public investment in irrigation was to gain higher revenue” [Reddy 1990, 1047]. The colonial government approved a private entity, the Madras Canal Company, to invest in irrigation projects. This Company started to focus on building bigger water bodies and dams. The British colonial government “used all possible ways to extract maximum revenue in the name of irrigation development” [Reddy 1990, 1057].

11

In such scenarios, the aforementioned hypothetical concepts elucidate concerns regarding water during the colonial period, excluding the primary model “political and movement” from text 2. Despite the relevance of the sub-model “water tax” from text 2 to other models, the logical framework of the formal model for the primary model might overlook the sub-model “water tax.” This oversight is due to the primary model being designed to extract the most crucial concepts from the entire text, and the aforementioned sub-model may not be estimated a pivotal concept. However, it remains pertinent to topic-driven research. For further analysis, one can employ additional machine learning models for a quantitative study of the selected sub-models and primary models or curate these models for a qualitative study. Alternatively, the corpus can undergo reorganization by filtering texts, informed by these models. In every aspect, these models contribute to our exploration, categorization, and understanding of the corpus.

12

3. Methodology

For this conceptual and theoretical framework, I chose an unsupervised approach using Affinity Propagation Algorithm (APA). The APA is widely used in many diverse fields for data mining such as medicine, chemistry and bioinformatics etc. Brendan J. Frey and Delbert Dueck introduced APA in 2007 [Frey et al. 2007]. It has a number of advantages compared to other clustering algorithms such as Kmeans. Kmeans requires a number of clusters to group words (or in other words data points) gleaned from its similarities. On the other hand, APA does not entail the number of clusters and it takes “input measures of similarity between pairs of data points and [r]eal-valued messages are exchanged between data points until a high-quality set of exemplars and corresponding clusters gradually emerges” [Frey et al. 2007, 972]. If the data point i is similar (s) to the data point k which is determined through negative Euclidean distance, then $s(i, k)$ is $s(i, k) = -||x_i - x_k||^2$ [Frey et al. 2007, 972]. and Levenshtein Distance is employed to calculate the distance between data points. It measures the distance between the data points based on their similarity and assigns a weightage to each word.

13

The communication between data points (i, k) happens through responsibility(r) and availability (a). The r on behalf of data point i will send a message to check the a of the k and looking for a s pattern. If k does not have similarity, then i will look for other data points. It is an iterative process till the data point i finds its group and vice versa [Frey et al. 2007, 972]. The other advantage feature of APA is “exemplar” which will be chosen as the “representative” of the cluster. For instance, if the data points (k, k) have larger values than other data points in the same cluster, it will be selected to represent the cluster. APA also chooses the number of clusters according to the density. This algorithm is applicable to the corpus like British India colonial as we cannot presume the number of clusters as the number could vary from text to text in the corpus. APA is efficient when it comes to small amounts of data. I applied APA to contents in each heading of text in the corpus for extracting sub models and also harnessed to each text separately for mining the primary models, but not to the whole corpus. 14

The corpus that I use for this article is part of an AHRC funded project “Digital Innovation in Water Scarcity” at Lancaster University in which I worked as a Research Associate. The primary aim of this project is to study the water transformation from early nineteenth century British India Coimbatore to present through diverse corpus of historical maps, texts and oral testimonies through various innovative DH methods. The British colonial India corpus is curated from various open access online platforms such as Internet Archive, Google books and National Library of Calcutta. These texts are out of copyright. This corpus is small and consists of 29 texts (see Figure 3). The corpus is cleaned as they were curated as PDFs and image files through third party with help of project fund. 15

In the preprocessing, each text in the corpus is divided into a number of parts such as titles, table of contents, preface, index and body of the content. These parts are marked using hyperlinks tags. These tags are used to extract table of contents and body of the content. The tagged “Table of contents” and “Body of the content” are cleaned to remove the punctuations and other symbols as they encumber mining the clusters. However, this preprocessing has a few challenges since not all historical texts are standardized using the table of contents. Three texts did not have table of contents but they have headings inside the text and five texts did not have neither table of contents nor headings inside the text. Since all headings inside texts are in uppercase and most of the table of contents are also in uppercase. The headings in uppercase facilitated extracting their content using first heading as start and second head as end. 16

No	Book Title	Author	Year
1	A Journey from Madras Through the Countries of Mysore, Canara, and Malabar vol 1	Francis Buchanan	1807
2	A Journey from Madras Through the Countries of Mysore, Canara, and Malabar vol 2	Francis Buchanan	1807
3	A journey from Madras through the countries of Mysore, Canara, and Malabar vol 3	Francis Buchanan	1807
4	Annual Report On The Administration Of The Madras Presidency 1862-63	No author name	1863
5	Census Report Of Madras Presidency	W.R. (Cornish) F.R.C.S., Surgeon	1871
6	Economic Conditions In The Madras Presidency	A.Sarada Raju	1941
7	Forestry Administration in the Madras Presidency	D.Brandis	1883
8	Handbook of the Madras Presidency	John Murray	1879
9	Illustrated Guide To The South Indian Railway	No author name	1926
10	Madras District Gazetteers Coimbatore District	Robert B. Buckley	1880
11	Madras District Gazetteers Trichinopoly	Hemingway	1907
12	Madras District Manuals Coimbatore	F. A. Nicholson	1898
13	Manual of the Administration of the Madras Presidency, in Illustration of the Records of Government & the Yearly Administration Reports	W. Ainslie, A. Smith, M. Christy	1816
14	Medical, Geographical, and Agriculture report of a committee appointed by the Madras government, To inquire into the cause of the Epidemic fever which prevailed in the provinces of Coimbatore, Madura, Dindigul, and Tinnivelly, during the years 1809, 1810, 1811	W. Ainslie and his team	1816
15	Memorandum of public works, calculated to obviate or mitigate famine, and notes of some	William Lewis C.I.E	1878
16	Report On The Administration Of The Madras (Presidency) During The Year 1879-80	No author name	1893
17	Report On The Administration Of The Madras Presidency During The Year 1868-69	H.Morgan	1870
18	Report On The Administration Of The Madras Presidency During The Year 1869-70	No author name	1893
19	Report On The Administration Of The Madras Presidency During The Year 1875-76	E.Keys	1877
20	Report on the improvement of Indian agriculture	K.N.Krishnaswami Ayyar	1933
21	The imperial gazetteer of India Volume 1	Hunter	1885
22	The irrigation works of India, and their financial results. Being a brief history and description of the irrigation works of India, and of the profits and losses which they have caused to the state	Robert B. Buckley	1880
23	The Madras Irrigation and Canal Company(Limited)	No author name	1857
24	The Madras presidency, with Mysore, Coorg and the associated states	Edgar Thurston	1913
25	The Madras road book	No author name	1839
26	The Railways of India: with an account of their rise, progress, and construction, written with the aid of the records of the India office	Edward Davidson	1868
27	General Sir Arthur Cotton his life and work	Elizabeth Hope and William Digby	1900
28	Memorandum On The Progress Of The Madras Presidency During The Last Forty Years Of British Administration	S.Srinivasa Raghavaiyengar, B.A., Dewan Bahadur, C.I.E	1893
29	Twelfth Annual Report Of The United States Geological Survey To The Secretary Of The Interior 1890-91	J. W. Powell	1891

Figure 3. List of texts in the corpus

However, the last heading, of course, would not have a subsequent heading for which I added “End of the Doc” at the end of the text. Then I extracted the contents inside the headings for twenty-four texts using regular expressions. For the remaining five texts, I had to extract the body of the text as there is no option to separate the content using table of contents. I used loop function to run the text one by one in Python. After preprocessing, I converted the extracted data into DataFrame using Pandas which later passed to APA. Kmeans^[2] algorithm in Sklearn is applied to predict the clusters through elbow method, which is applied to “test the consistency of the best number of clusters by comparing the difference of the sum of square error,” to compare the efficiency of means and APA to improve the efficiency [Umargono et al. 2020, 121]. The proficiency of the latter is much better than the former. I also decided to extract top fifteen words in each cluster after a few trials. Finally, the data in dataframe, prediction and the top features of the cluster passed into the APA’s similarity function to extract clusters and exemplars for the content in each heading and the entire text and both are stored into separate CSV files. Then I created three kinds of visualizations using Tableau and RAW online graph: representing the total number of clusters across the corpus; distinguishing between the sub-model and primary model through selected two texts and visualizing the models from the selected text for further investigation.

4. Discussion

Distribution of concepts in sub-model and primary model

Investigating the number of clusters obtained by models for both heading clusters (hereafter HC) and text clusters (hereafter TC) would help us discern the complexity and nature of the corpus before delving into studying the labels of the clusters. Although the concepts presented in the headings are pertinent to the text, often, the headings represent unique concepts that might not be traversed throughout the text. In this case, when the model identifies unique words and their similarities based on the distance, the output for the clusters of each heading could be greater in number, as they may be unique and dense only for their respective headings, unlike the entire text. Hence, Figure 4, depicting the

clusters of all the HC in red and TC in blue, reveals that the former outnumber the latter. However, to discern this difference, we will discuss how the concepts distributed across the headings and text vary from one another through selected texts from the corpus.

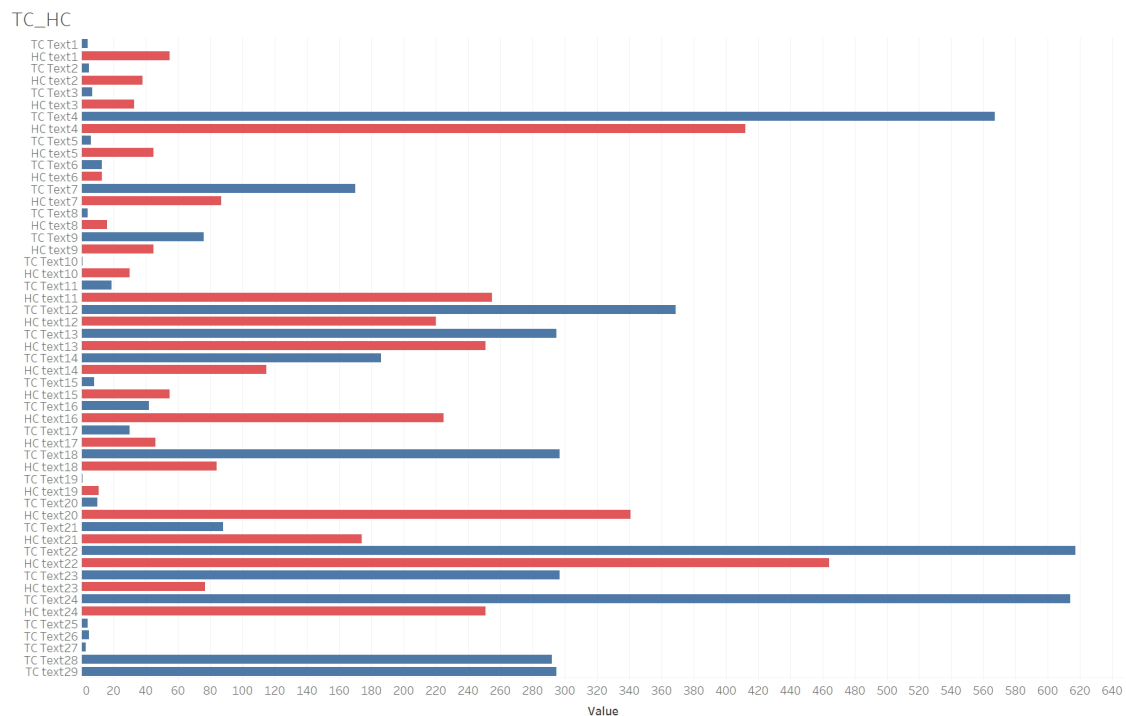


Figure 4. The clusters of all the heading clusters in red and text clusters in blue

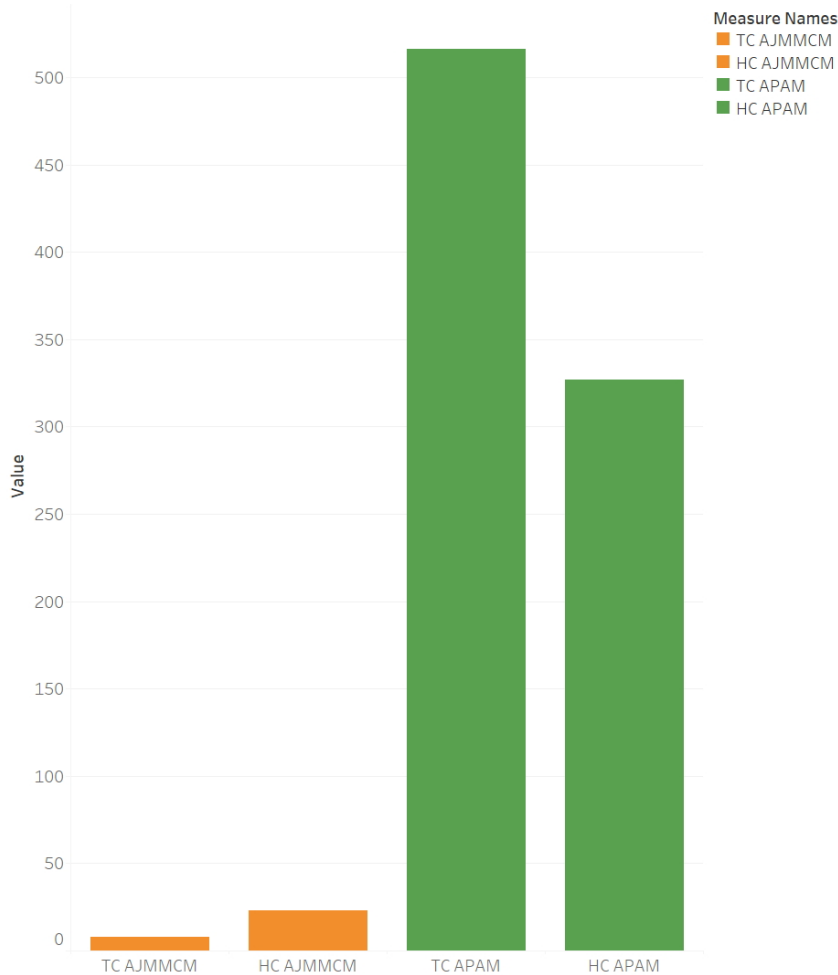


Figure 5. The total of heading clusters and text clusters of the texts AJMMCM and APAM

Figure 5 shows HC and TC of Francis Buchanan's *A Journey from Madras Through the Countries of Mysore, Canara, and Malabar* (1807) (hereafter AJMMCM) *Volume 1* and *Annual Report On The Administration Of The Madras Presidency 1862–63*^[3] (hereafter APAM). We can observe that the TC of the latter are more in number than HC of the former and vice versa. Bernard Cohn states,

18

For many British officials, India was a vast collection of numbers. This mentality began in the early seventeenth century with the arrival of British merchants who compiled and transmitted lists of products, prices, customs and duties, weights and measures, and the values of various coins. [Cohn 1996, 8]

As he rightly says, such numbers were used to govern and control the colonized people and land as Nicholas B. Dirks, in his Foreword to Cohan's *Colonialism And Its Forms Of Knowledge The British In India*, says "[c]olonial knowledge both enabled conquest and was produced by it; in certain important ways, knowledge was what colonialism was all about" [Cohn 1996, ix]. As Dirks states, the British officials were so keen in surveying and documenting everything about Indian society. In the selected text AJMMCM, the Scottish surgeon and botanist Buchanan surveyed the recently annexed kingdom of Mysore, Canara, and Malabar in southern India at the beginning of the nineteenth century. The survey covers the physical and human geography of the region, commerce, detailing agriculture, arts, culture, indigenous religions, society, customs and natural history. In this Volume 1, he emphasizes the agricultural aspects, including irrigation systems, variety of crops and their cultivation details, the condition of the soil and many more. AJMMCM is divided into six long chapters with specific sub-headings. On the other hand, APAM, containing thirty-six headings, offers diverse information and details on legislative, judicial, criminal justice, and also topics related to forest

19

conservancy, plantations, and irrigation. The aim of AJMMCM is to survey the features of the recently annexed southern regions, and the concept and theme of the text are consistent through its lengthy descriptive narrative. Hence, APA found a few crucial concepts to cluster for the primary model. But they clustered the unique heterogeneous concepts in each heading that might not be overlapped with other parts text. Conversely, APAM has numerous concepts and information but is presented concisely in analytical narrative. Therefore, APA could not find many clusters in the headings but, *de facto*, grouped many diverse concepts that appeared throughout the text.

Nevertheless, this helps us fathom out how the concepts are distributed in each heading and the entire text, which can vary. Corpas and Seghiri rightly point out that “[t]he number of tokens and/or documents a specialized corpus should contain may vary in relation to the languages, domains, and textual genres involved, as well as to the objectives set for a specific analysis (i.e., a corpus should provide enough evidence for the researchers’ purposes and aims)” [Corpas Pastor and Seghiri 2010, 135]. It also brings attention to the selection of the text for building this corpus. As I mentioned in the introduction, my aim is to mine the details of water in British India colonial documents, particularly the documents, texts, reports, and surveys of The Madras Presidency. I aggregated texts that might have any potential data about water. The above-mentioned two texts, for instance, although vary in terms of their rationale and aim, have much data about water.

20

Studying the models

To comprehend the semantics of the mined clusters, I explored the labels of clusters and its exemplar feature of APA. According to Frey and Dueck, “[a] common approach is to utilize data to learn a set of centers such that the sum of squared errors between data points and their nearest centers is small. When the centers are chosen from actual data points, they are referred to as ‘exemplars’” [Frey et al. 2007, 972]. Exemplars serve as representatives of their respective clusters and also help us to build the semantic model of clusters. However, a comprehensive exploration of the entire labels of clusters and exemplars extends beyond the scope of this article; therefore, I will closely study AJMMCM. The potential sub-models derived from the six chapters of AJMMCM encompass fanams, farmer, irrigation, drugs, fades, fair, iron, turban, canara, cloth, cubits, oil, prey, weavers, crop, zemindars, extent, and july (see Figure 6). These exemplars and their clusters succinctly encapsulate the distinct concepts of AJMMCM. For instance, the exemplars fanams, fades, cloth, turbans, and weaver and their significant clustered terms such as families, brahmans, devangas, villages, natives, strata, customs, cotton, silver^[4] etc. from Chapters 1, 3 and 4 signify Buchanan’s survey of social and cultural milieu of southern India. These models hold relevance for research inquiries concerning socio-cultural settings in southern India. Similarly, the exemplars irrigation, july, crop, harulu, farmers, cultivation, extent, country, and zemindars and their clustered terms such as ragy, casts, seed, rice, corn, buffalo, water, field, straw, plough, soil, barugu, weights, grain, sugar, bees, tobacco^[5] etc. from all six chapters convey Buchanan’s detailed study of the agricultural system in the southern regions. These models offer valuable insights for research related to environmental, agrarian, and economic history.

21

It is crucial to acknowledge that these exemplars should not be entirely relied upon, as they do not serve as either the topic or title of the clusters; instead, they merely function as representatives. They provide only a glimpse into the clusters. To comprehend the model, one must delve into the terms of the clusters. Moreover, not all exemplars are truly useful and provide an immediate sense of the clusters. For instance, in the previously mentioned exemplars, terms such as extent, fades, and july did not contribute any meaningful sense to construct the concepts. However, a meticulous examination of the terms of these exemplars, including customs, measures, plough, sows, sesame, palm gardens, cultivation, soil, bushes, jola, barugu, etc., once again signifies the extended discussion on the agrarian culture of the regions. For example, in the quotes below, Buchanan explains the crop of Jola, its kinds, and cultivation.

22

Of these crops Jola (*Holcus sorghum*) is the greatest. There are two kinds of it, the white and the red which are sometimes kept separate, and sometimes sown mixed. The red is the most common. Immediately after cutting the Vaisaka, crop: of, rice, plough four times in the course of twenty days. [Buchanan 1807, 283]

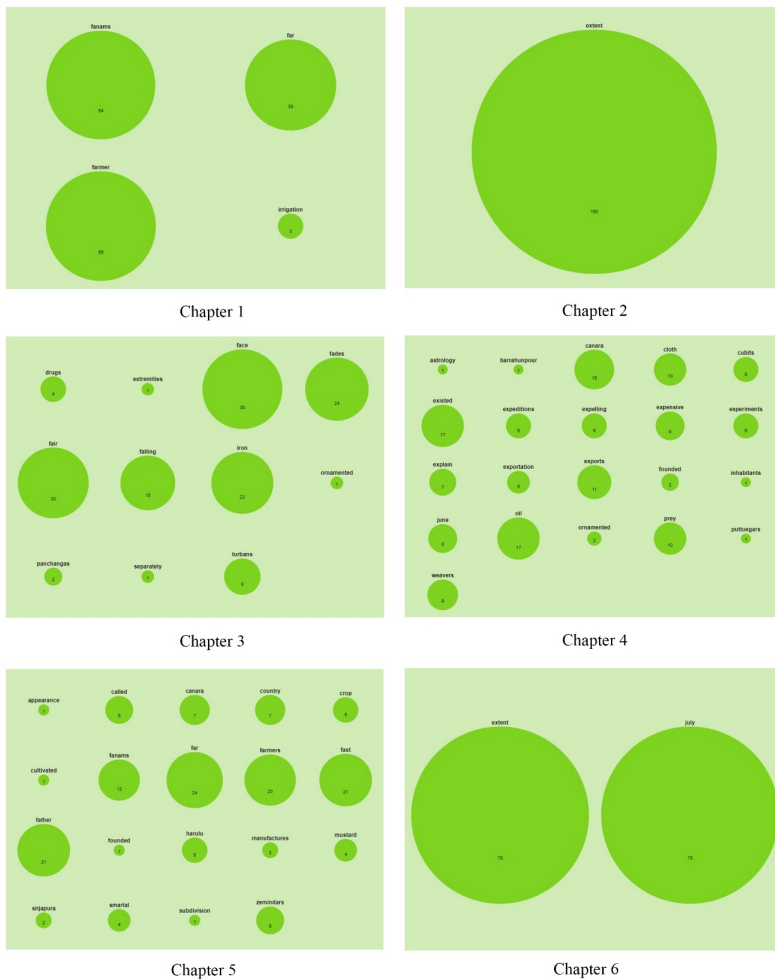


Figure 6. Sub-models of the chapters of the text AJMMCM

There are nine clusters in the primary model and the important exemplars are barugu, extract, fair, famine, harica^[6], water etc. (see Figure 7). These representees and their terms such as rice, fanam, sugarcane, cultivation, irrigation, jola, ragy, july, bushes, plough, seed, trade, dry, land etc. indeed convey the agricultural facets which is the primary concept of the text. Although the close study of the terms presented in the model can be associated with the pivotal concept of the text, the nuanced heterogenous concepts extracted in the sub-model have been disregarded in the primary model.

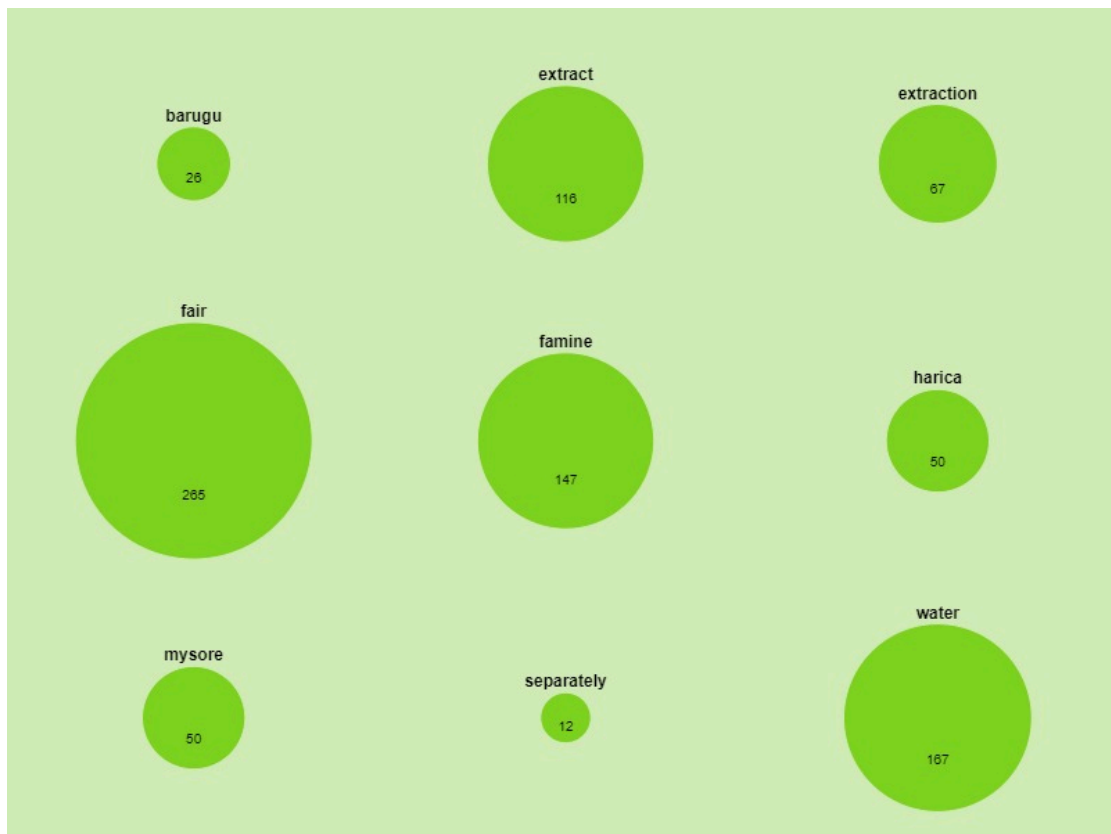


Figure 7. Primary models of the text AJMMCM

For example, Buchanan demonstrated a keen interest in surveying autochthonous resources, as manifested in exemplars such as drugs, oil, and iron. The text features subheadings specifically addressing these resources. The term drug occurs 11 times, oil 111 times, and iron 68 times in the text. Nevertheless, in the clustered terms of the primary model, oil appears 5 times, and iron appears once. Remarkably, the term drug does not appear at all due to its dense paucity.

24

Drugs. A kind of drug merchants at Bangalore, called Gandhaki, trade to a considerable extent. Some of them are Baniijgaru, and others are Ladaru, a kind of Mussulmans. They procure the medicinal plants of the country by means of a set of people called Pacanat Jogalu, who have huts in the woods, and, for leave to collect the drugs, pay a small rent to the Gaudas of the villages. They bring the drugs hither in small caravans of tea or twelve oxen, and sell them to the Gandhaki, who retail them. None of them are exported. [Buchanan 1807, 204]

In the above excerpt, Buchanan elucidates the procurement process of drugs by Gandhaki, drug merchants in Bengaluru, from local suppliers Pacanat Jogalu, who gather them in the woods. Additionally, he provides a detailed explanation of the manufacturing, trade, and application of various oils, including coconut oil, sesame oil, castor oil, bassia oil, and hoingay oil. Likewise, Buchanan delves into the examination of natural minerals. In a subsequent passage, he narrates how a specific local community acquires materials for iron manufacturing and he dedicates a substantial portion in Chapter 3 to elucidate the comprehensive iron production process.

25

Iron forges. About two miles from Naiekan Eray, a torrent, in the rainy season, brings down from the hills a quantity of iron ore in the form of black sand, which in the dry season is smelted. The operation is performed by Malawanlu, the Telinga name for the cast called Parriar by the natives of Madras. Each forge pays a certain quantity of iron for permission to carry on the work. [Buchanan 1807]

Owing to the extensive discussions on these resources within the Chapters, APA has selected drugs, oil, and iron clusters based on their density in the corresponding headings in the sub-model. These diverse concepts were

26

disregarded in the primary model. On the contrary, as detailed in the initial subsection of this section, certain texts exhibit more clusters in the primary model than in the sub-model, or in the case of the APAM, wherein crucial concepts like “settlement” have been omitted from its sub-model. Nevertheless, when scrutinizing the British colonial India corpus, the integration of these models demonstrates greater efficacy in formalizing the heterogeneous concepts.

The model, in general, is designed to identify overarching patterns in the data. However, it is imperative to incorporate the sub-model in the exploration. Theoretical DH should address these crucial considerations in the characteristics of formal models for examining a complex corpus. Nevertheless, building computational models to extract primary and sub-models was quite challenging as big data models, in general, are significantly utilized to discern trends within the data to generate novel insights [Bhattacharyya 2017]. The data models might neglect non-trends that still constitute part of the data. Sayan Bhattacharya conducted an experiment utilizing the Bookworm tool, designed to visualize language usage trends within millions of digitized texts in HathiTrust. He contends that the model, crafted to explore and visualize language usage trends, has a limitation in identifying “words from less hegemonic languages” [Bhattacharyya 2017, 34]. He illustrates his argument by showcasing underreported transliterated words (English) from Global South languages and delves into the causes behind such limitations^[7]. Indeed, the issue stems from tools like HathiTrust Bookworm relying on an index that, for performance reasons, excludes entries for low-frequency words. This disproportionately impacts the representation of low-frequency words in larger collections.

This is applicable when studying corpora like British colonial India, as less trendy concepts are overshadowed by the trend concepts within the text in both models. Unlike digital tools, which do not permit alterations to their frameworks, computational models can be manipulated to formalize these less-trendy concepts in the corpus. Hence, the amalgamation of primary and sub-models proves advantageous in studying and formalizing the heterogeneous concepts within the British colonial India corpus as demonstrated using the selected text AJMMCM. I can reorganize the texts in the corpus based on the formalized concepts and apply formal models for further investigation.

Disadvantages, challenges and future work

Numerous issues were encountered during the experiment, including problems with text format, non-standard text, parameters for cleaning texts, and limitations in the selected algorithm. The first issue arose from tagging the table of contents. Some texts lacked a table of contents but had headings inside the text, while others had neither. Separate algorithms were designed for each case. The second issue was the exclusion of stemming^[8] and lemmatization despite removing stop words. These processes could impact clustering, especially sub-models derived from a few paragraphs or pages. Lemmatization might reduce counts and, additionally, the inconsistent content distribution across headings, with some having only one or two paragraphs, led to an increase in outliers^[9]. Spelling variation in Indian names and place names posed another challenge. For instance, the river name “Noyal” had various spellings like “Noyil,” “Noel,” and “Noyl” affecting frequency and clustering patterns.

Another challenge is the inclusion of footnotes and references running throughout most texts. There are many terms and exemplars derived from these citations. APA accumulated many outliers, such as “separately,” which did not contribute explicitly to clustering concepts, but indicated numerous tables attached separately with the content. Subsequently, I excluded tables, prioritizing the narrative over statistics in The Madras Presidency reports. Involving grain details, surveyors included various statistics — crops, revenue, a census of houses, and population categorized by religion, castes, and more. These details are crucial for event-based research questions and should be formalized in future work. Another limitation is in the chosen algorithm, APA, with constraints like “high time complexity” for larger datasets. On the FAQ page for Affinity Propagation, Frey and his team addressed dataset size concerns, assuring APA’s reliability for small datasets. For instance, they answered a question: “Is affinity propagation only good at finding a large number of quite small clusters?” Their answer is:

It depends on what you mean by “large” and “small”. For example, it beats other methods at finding 100 clusters in 17,770 Netflix movies. While “100” may seem like a large number of clusters, it is not unreasonable to think that there may be 100 distinct types of movies. Also, on average there are 178 points per cluster, which is not “quite small”. However, if you’re looking for just a few clusters (eg, 1 to 5), you’d probably be better off using a simple method [Affinity

In this case, APA was suitable for sub-models and should also work for primary models since I mined the latter per text, which is not indeed a large dataset. However, it did not select potential exemplars for all primary models due to inconsistency in dissemination of the concepts.

32

5. Metamodel for concept-based corpus building

In the realm of computer science and related fields, a metamodel has surfaced, serving the purpose of “facilitating conceptual modeling, defining constructs of conceptual modeling languages, specifying constraints on the use of constructs, and encoding the similarities of different models” [Jeusfeld 2009, 1728]. Jeusfeld characterizes a metamodel as encompassing several models that include models, parameters, features, challenges, and limitations. Metamodels are designed “to build explicit” models, which are meticulously delineated to enable a thorough understanding of their contents [Epstein 2008]. Their significance in theoretical DH lies in facilitating a computational approach to studies in the humanities. Piotrowski underscores the importance of metamodels in theoretical DH, asserting that “the theoretical digital humanities create and study the metamodels whose concrete application to research questions in the disciplines of the humanities is the object of the applied DH” [Piotrowski 2022a, 3].

33

In the absence of an existing metamodel for the non-standard corpus to formalize the concepts within theoretical DH, I have formulated one based on processes, results, experiments, and experiences and represented it through concept diagrams. This metamodel is poised to significantly contribute to similar corpus-building research, serving as a valuable resource to improve future endeavors by addressing errors noted during experimentation and redesigning for enhanced comprehension.

34

Model 1 provides a detailed depiction of the corpus selection and pre-processing (see Figure 8). Establishing a corpus involves the construction of a model. Consequently, modelers must address inquiries such as: What constitutes the original? In what ways does the model serve as a reduction of it? And for whom and for what purpose is the model being created? [Piotrowski 2022b, 90]. Curating texts for a corpus pose significant challenges, including considerations about the collection of texts, copyright status, and existing machine-readable formats for materials. Despite this, the curated corpus is primarily guided by a general understanding of the research field. Consequently, eliminating the most and/or less relevant materials from the corpus is neither straightforward nor transparent. Subsequently, preprocessing becomes a crucial step in Model 1, marking the initial phase of any quantitative study. The texts should not only be in machine-readable formats (such as plain texts or PDF files) but also formalized using Text Encoding Initiative or hyperlink tags, particularly essential for non-standard corpus. A notable challenge in digital humanities methods for historical corpus is the presence of spelling variations in the text [Gregory 2014]. Acknowledging the impact of spelling issues on the models is paramount. Being cognizant about the influence of spelling issues in the models is pivotal. The nature of the research inquiry will determine whether the texts in the corpus necessitate deep cleaning, involving not only the removal of stop words but also stemming and lemmatization of words, significantly influencing the results in the subsequent model.

35

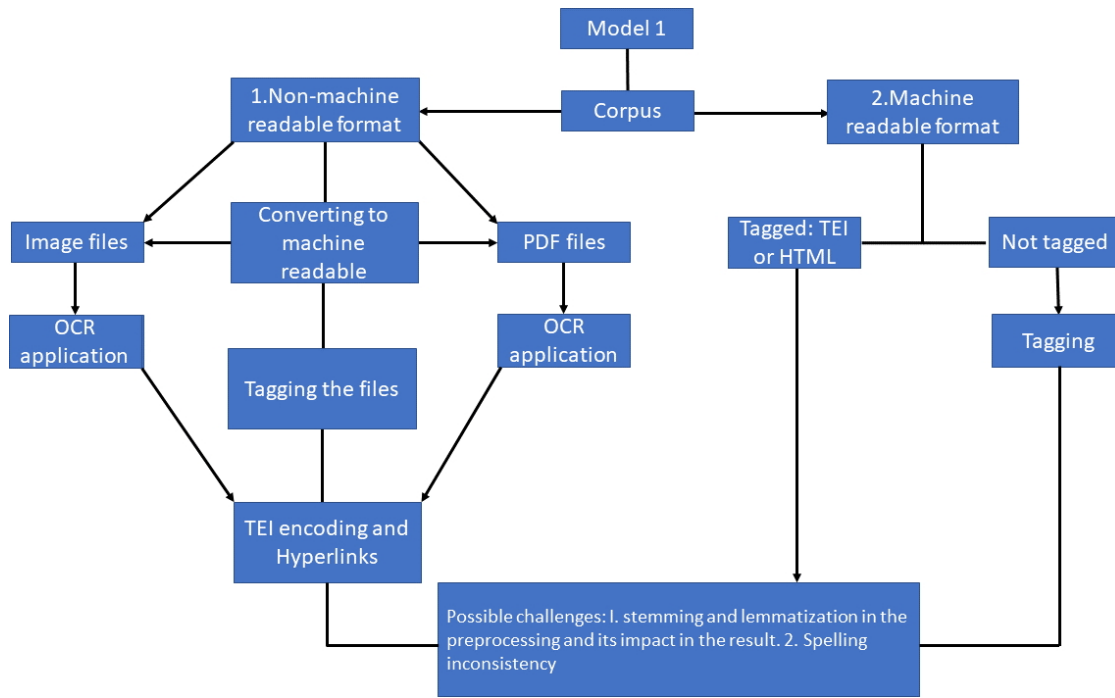


Figure 8. Model 1

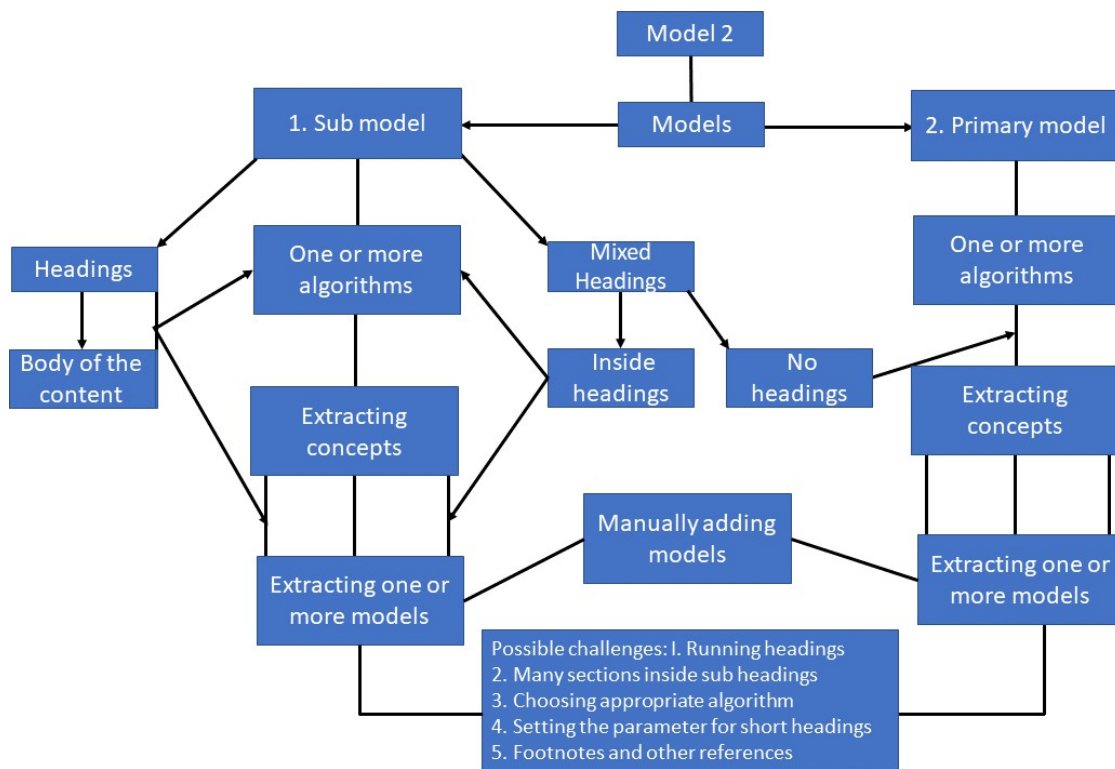


Figure 9. Model 2

Model 2 delineates the parameters and criteria governing the construction of the corpus through concept-based models (see Figure 9). The determination of whether to ground sub-models and/or primary models is contingent upon event- or topic-based research inquiries. Additionally, the choice between a single algorithm or a combination of algorithms for extracting concepts and their models significantly influences the selection of relevant texts for the corpus. The decision is also influenced by the nature of the research, as mining sub-models may be appropriate for some topic-driven research, while extracting both sub-models and primary models may be essential for others. In cases where the corpus

is intricate and comprises various forms and genres, such as documents, surveys, and reports, building the corpus based on both sub-models and primary models is deemed more fruitful. However, the selection of appropriate algorithms is crucial, as it determines the outcome of the quantitative study.

Moreover, the process of choosing models to represent concepts is pivotal and can be achieved either manually or through an algorithm. For example, if one opts for APA due to its exemplar feature, which is significant for the corpus used in this article, relying on a single exemplar from the cluster may sometimes be misleading. This bias arises as the concept is determined based on the preferred representative, leading to potential distortion. To mitigate this bias, selecting multiple exemplars based on the density of the cluster will, to a certain extent, evade this issue in concept-based models. However, this model presents other potential challenges, such as running headings on each page and an increase in the number of outliers due to fewer words in the content of the headings. Setting parameters to deriving concepts and models for headings with only a few sentences can address this issue; however, it may also result in the neglect of crucial data. Decisions regarding such trade-offs can be made through a process of trial and error.

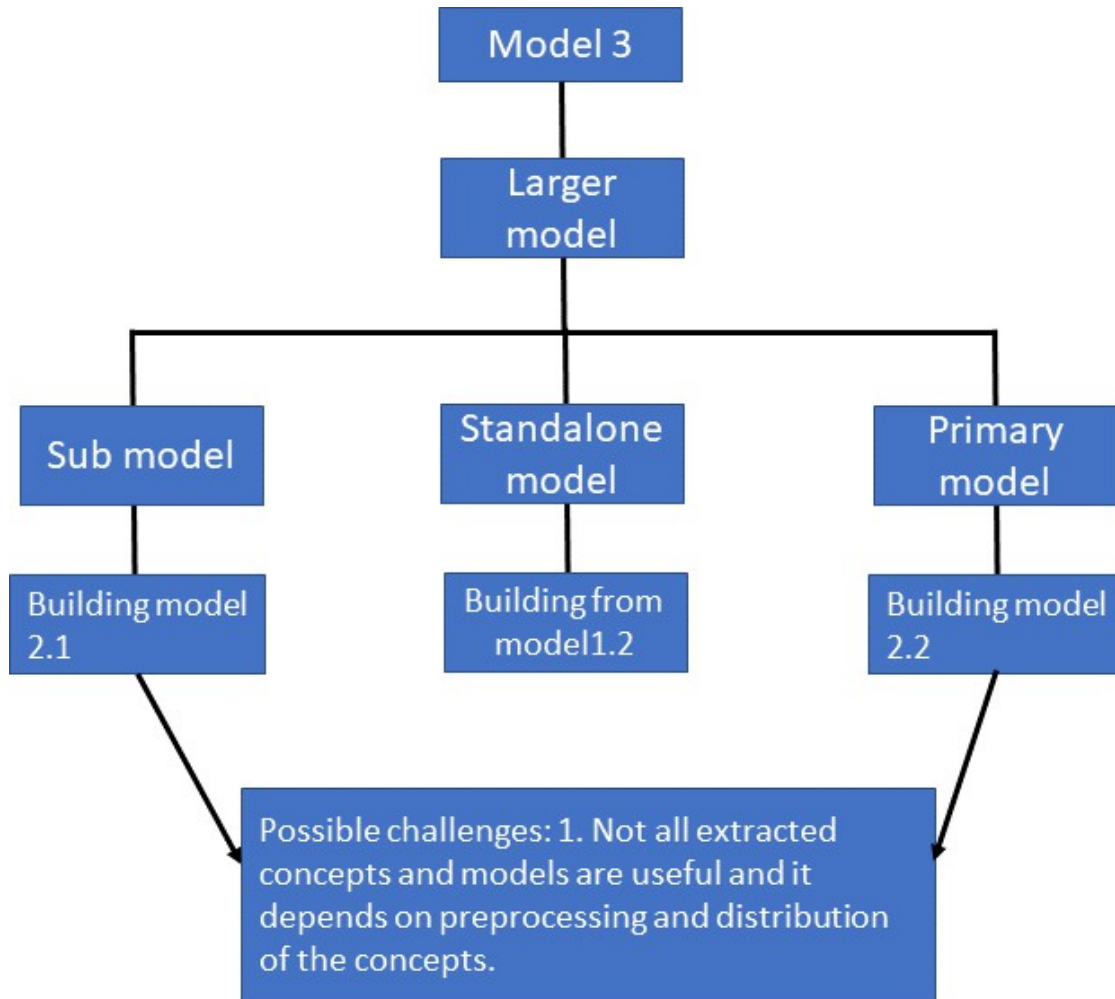


Figure 10. Model 3

Model 3 outlines the parameters for developing a larger model for the corpus (see Figure 10). Larger models can be generated from three types of datasets: sub-model, primary model, and standalone model. It can be created through a identifying the semantic network between sub-models and primary models. Such semantic network can also be employed to reorganize the texts in the corpus based on the priority and to develop a sub-corpus. Comparing the three models would be worthwhile to explore the dispersal of concepts and their models throughout the corpus.

6. Conclusion

The British colonial India corpus, which has not received attention in the field of DH, poses considerable challenges and drawbacks due to its intricate format and diverse concepts. This research seeks to confront these challenges by introducing concept-based formal models designed to formalize the heterogeneous concepts present in corpus, leveraging computational methods. In summary, this study illustrates the potential and challenges inherent in the development and application of formal models for the specified corpus. Derived from the investigation, several promising findings emerge:

1. Quantitative studies in DH typically prioritize the identification of trend patterns within a corpus. However, the computational approach proposed in this work assigns equal importance to both trend and non-trend patterns in the text, as the latter also significantly contributes to shaping the overall trend pattern. This stands in contrast to traditional approaches to studying historical concepts, which often rely on limited texts through qualitative studies or use computational methods to identify broader conceptual patterns. The recommended model enables a comprehensive study of concepts, even within extensive datasets.
2. Additionally, the concept-based model emerges as a convincing and promising framework, not only illustrating how concepts are distributed across the text and its headings but also capable of formalizing nuanced concepts to construct a more comprehensive model. The construction of the concept involves studying exemplars and their associated terms, underscoring the significance of domain knowledge in the decision-making process. Nevertheless, this model introduces a novel computational approach to trace concepts and provides insight into the curation of colonial knowledge.
3. The application of the theoretical framework, as outlined in Section 2, is highly effective in exploring the chosen corpus. This model proposes to categorize the text into two types: the text itself and the units of the text based on its peritext. The development of formal models can be accomplished using this categorization. However, as mentioned in Section 3, some challenges arose in these divisions, as a few texts could not be segmented into units due to the absence of peritext. Nevertheless, these challenges were addressed by manipulating the formal models to extract units based on categorization within the text (headings).
4. The analysis in Section 4 reveals that the number of models between sub-models and primary models may vary based on the distribution of concepts within the unit and the text. The examination of the extracted models from the chosen text underscores the significance of the sub-model in deriving nuanced concepts from the text, which are overlooked in the primary model and vice versa. However, in the process of constructing and studying the corpus for topic-specific research, both models are indispensable and can mutually enhance exploration of the concepts within the text, aligning with the primary objective of this proposed model.
5. Lastly, Section 5 delineates the metamodels, a facet not prioritized in theoretical DH. Derived from the experiment, it provides insights into the general properties, possibilities, and challenges involved in studying a similar corpus.

The findings of this article propose several directions for future research:

1. The current corpus is relatively modest in size. In subsequent work, an expansion of the corpus by curating additional texts is planned and the texts will be categorized based on its forms and genres, including documents, surveys, and reports. Moreover, the adoption of TEI guidelines to generate tags for the texts will be considered. Systematic categorization and tagging systems aim to enhance the formalization process.
2. While the APA computational algorithm demonstrates efficacy for the proposed model, there are limitations in the exemplar features and clustering patterns, notably a surge in outliers, and not all exemplars convey meaningful insights. Future investigations will explore alternative advanced computational models or combinations thereof to achieve more efficient pattern clustering. Additionally, efforts will be made to fine-tune parameters for extracting more than one exemplar per cluster based on content density.
3. The next research focus involves formalizing the spelling of Indian names. As highlighted in the final subsection of Section 4, a considerable number of Indian names did not appear in the cluster due to

inconsistencies in their spellings. Ongoing efforts are dedicated to addressing this issue, with a specific emphasis on formalizing concepts based on Indian names [Shanmugapriya 2023].

Acknowledgements

I express my gratitude to Mohanapriya, the software trainer, for her assistance in preparing the codes for the entire project. Many thanks to Bhavani Raman, a historian at the University of Toronto Scarborough, for providing valuable insights into the format and non-standard nature of the British colonial India corpus. I am also thankful to the AHRC for funding the “Digital Innovation in Water Scarcity Coimbatore India” project, through which the corpus was collected and cleaned using project funds. Additionally, I appreciate Deborah Sutton, who served as the Principal Investigator of this project at Lancaster University.

41

Notes

[1] The Madras Presidency was one of the subdivisions of British India. It covered most part of the southern states. It was divided into five districts after the independence, namely Erode, Coimbatore, Karur and Tirupur.

[2] I used kmeans to explore through prediction to compare and verify the efficiency of APA.

[3] The author’s name was not mentioned in the document.

[4] The Brahmans and Devangas are cast in South India.

[5] Harulu refers to Ricinus, and Ragi and Barugu are primary millets in South Indian regions.

[6] Harica is also a kind of millet in South India.

[7] Besides the limitation of the tool, Bhattacharya also discusses other reasons such as spelling inconsistencies and optical character recognition issues in digitized texts from the Global South impacted the frequency of the transliterated words [Bhattacharyya 2017, 36].

[8] Stemming and lemma are text normalization methods used in Natural Language Processing. The former is applied to remove the affixes of the word, for example, the stem of “reading” and “reads” is “read.” The latter is used to find the root words for instance, the lemma of the word “went” is “go.”

[9] If the cluster has only one word, it is known as outlier as it has a “minimal membership proportion” [Evans et al. 2015, 2].

Works Cited

- Affinity Propagation FAQ 2009** “Affinity Propagation FAQ.” (2009) *Probabilistic and Statistical Inference Group University of Toronto*. Available at: <http://genes.toronto.edu/affinitypropagation/faq.html> (Accessed: 5 October 2022).
- Beynon et al. 2006** Beynon, M., Russ, S. and McCarty, W. (2006) “Human Computing — Modelling with Meaning,” *Literary and Linguistic Computing*, 21(2), pp. 141–157. Available at: <https://doi.org/10.1093/lc/fql015>.
- Bhattacharyya 2017** Bhattacharyya, S. (2017) “Words in a world of scaling-up:: Epistemic normativity and text as data,” *Sanglap: Journal of Literary and Cultural Inquiry*, 4(1), pp. 31–42. Available at: <https://sanglap-journal.in/index.php/sanglap/article/view/86> (Accessed: 27 November 2023).
- Brigandt 2010** Brigandt, I. (2010) “The epistemic goal of a concept: accounting for the rationality of semantic change and variation,” *Synthese*, 177(1), pp. 19–40. Available at: <https://www.jstor.org/stable/40985618> (Accessed: 20 November 2022).
- Buchanan 1807** Buchanan, F. (1807) *A Journey From Madras Through The Countries Of Mysore, Canara, And Malabar Volume 1*. London: The Directors Of The East India Company.
- Buzzetti 2002** Buzzetti, D. (2002) “Digital Representation and the Text Model,” *New Literary History*, 33(1), pp. 61–88. Available at: <https://www.jstor.org/stable/20057710> (Accessed: 5 November 2023).
- Ciula et al. 2018** Ciula, A. et al. (2018) “Models and Modelling between Digital and Humanities: Remarks from a Multidisciplinary Perspective,” *Historical Social Research*, 43(4), pp. 343–361. Available at: <https://doi.org/10.12759/hsr.43.2018.4.343-361>.

- Cohn 1996** Cohn, B.S. (1996) *Colonialism & Its Forms of Knowledge – the British in India*. Princeton, NJ: Princeton University Press.
- Corpas Pastor and Seghiri 2010** Corpas Pastor, G. and Seghiri, M. (2010) *Size matters: A quantitative approach to corpus representativeness*. León: Universidad de León, Área de Publicaciones, 2010. Available at: <https://buleria.unileon.es/handle/10612/4752> (Accessed: 7 November 2022).
- Edney 1997** Edney, M. H. (1997) *Mapping an Empire: The Geographical Construction of British India 1765–1843*. Chicago: University of Chicago Press.
- Ehrlich 2023** Ehrlich, J. (2023) *The East India Company and the Politics of Knowledge*. Cambridge University Press. Available at: <https://doi.org/10.1017/9781009367967>.
- Englmeier et al. 2021** Englmeier, T. et al. (2021) “Using an Advanced Text Index Structure for Corpus Exploration in Digital Humanities,” 15(1). Available at: <https://www.digitalhumanities.org/dhq/vol/15/1/000526/000526.html> (Accessed: 8 November 2022).
- Epstein 2008** Epstein, J.M. (2008) “Why Model?,” *Journal of Artificial Societies and Social Simulation*, 11(4). Available at: <https://www.jasss.org/11/4/12.html> (Accessed: 8 October 2022).
- Evans et al. 2015** Evans, K., Love, T. and Thurston, S.W. (2015) “Outlier Identification in Model-Based Cluster Analysis,” *Journal of Classification*, 32(1), pp. 63–84. Available at: <https://doi.org/10.1007/s00357-015-9171-5>.
- Flanders et al. 2015** Flanders, J. and Jannidis, F. (2015) *Knowledge Organization and Data Modeling in the Humanities*. Available at: <http://www.wwp.northeastern.edu/outreach/conference/kodm2012/index.html> (Accessed: 8 November 2023).
- Frey et al. 2007** Frey, B.J. and Dueck, D. (2007) “Clustering by Passing Messages Between Data Points,” *Science*, 315(5814), pp. 972–976. Available at: <https://doi.org/10.1126/science.1136800>.
- Gregory 2014** Gregory, I. (2014) “Challenges and Opportunities for Digital History,” *Frontiers in Digital Humanities*, 1. Available at: <https://www.frontiersin.org/articles/10.3389/fdigh.2014.00001> (Accessed: 8 November 2022).
- Jeusfeld 2009** Jeusfeld, M.A. (2009) “Metamodel,” in L. LIU and M.T. ÖZSU (eds) *Encyclopedia of Database Systems*. Boston, MA: Springer US, pp. 1727–1730. Available at: https://doi.org/10.1007/978-0-387-39940-9_898.
- Jähnichen 2017** Jähnichen, P. et al. (2017) “Exploratory Search Through Visual Analysis of Topic Models,” *Digital Humanities Quarterly*, 011(2). Available at: <https://www.digitalhumanities.org/dhq/vol/11/2/000296/000296.html> (Accessed: 8 November 2023).
- Linguistic DNA** Linguistic DNA. (n.d.) “Approaching concepts,” *Linguistic DNA Modelling concepts and semantic change*. Available at: <https://www.linguisticdna.org/approaching-concepts/> (Accessed: 2 October 2022).
- McCarty 2004** McCarty, W. (2004) “Modeling: A Study in Words and Meanings,” in S. Schreibman, R. Siemens, and J. Unsworth (eds) *A Companion to Digital Humanities*. Oxford: Blackwell, pp. 254–270. Available at: <https://doi.org/10.1002/9780470999875.ch19>.
- McCarty 2005** McCarty, W. (2005) *Humanities Computing* | SpringerLink. London and New York: Palgrave.
- Oberbichler et al. 2021** Oberbichler, S. and Pfanzer, E. (2021) “Topic-specific corpus building: A step towards a representative newspaper corpus on the topic of return migration using text mining methods,” *Journal of Digital History*, 1(1), pp. 74–98. Available at: <https://journalofdigitalhistory.org/en/article/4yxHGiqXYRbX> (Accessed: 2 November 2022).
- Piotrowski 2019** Piotrowski, M. (2019) “Accepting and Modeling Uncertainty,” *Zeitschrift für digitale Geisteswissenschaften* [Preprint]. Available at: https://zfdg.de/sb004_006#fn32 (Accessed: 6 November 2023).
- Piotrowski 2022a** Piotrowski, M. (2022) *Epistemological Issues in Digital Humanities*, Zenodo. Available at: <https://doi.org/10.5281/zenodo.6498979> (Accessed: 1 November 2022).
- Piotrowski 2022b** Piotrowski, M. (2022) *Some Reflections on Historiographical Uncertainty and Computational Modeling*, Zenodo. Available at: <https://zenodo.org/records/6672504> (Accessed: 10 November 2022).
- Piotrowski et al. 2020** Piotrowski, M. and Fafinski, M. (2020) “Nothing New Under the Sun? Computational Humanities and the Methodology of History,” in *CEUR Workshop Proceedings. CHR2020: Workshop on Computational Humanities Research*, Amsterdam, The Netherlands, pp. 171–181. Available at: <https://ceur-ws.org/Vol-2723/> (Accessed: 6 November 2022).
- Reddy 1990** Reddy, V.R. (1990) “Irrigation in Colonial India: A Study of Madras Presidency during 1860–1900,” *Economic*

and *Political Weekly*, 25(18/19), pp. 1047–1054. Available at: <https://www.jstor.org/stable/4396266> (Accessed: 5 November 2023).

Saravanan 2020 Saravanan, V. (2020) *Water and the Environmental History of Modern India*. London: Bloomsbury Academic.

Shanmugapriya 2023 Shanmugapriya, T. (2023) “From Uncertainty to Action: Recalibrating Digital and Spatial Humanities Methods and Tools for Non-standard Historical Data from Global South,” in *GeoHumanities '23: Proceedings of the 7th ACM SIGSPATIAL International Workshop on Geospatial Humanities. 7th ACM SIGSPATIAL International Workshop on Geospatial Humanities*, Hamburg, Germany: Association for Computing Machinery Library, pp. 60–62. Available at: <https://dl.acm.org/doi/10.1145/3615887.3627762> (Accessed: 13 November 2023).

Tomasi 2018 Tomasi, F. (2018) “Modelling in the Digital Humanities: Conceptual Data Models and Knowledge Organization in the Cultural Heritage Domain,” *Historical Social Research / Historische Sozialforschung. Supplement*, (31), pp. 170–179. Available at: <https://www.jstor.org/stable/26533637> (Accessed: 1 November 2023).

Umargono et al. 2020 Umargono, E., Suseno, J. and Gunawan, S.K. (2020) “K-Means Clustering Optimization Using the Elbow Method and Early Centroid Determination Based on Mean and Median Formula,” in *Proceedings of the 2nd International Seminar on Science and Technology (ISSTEC 2019). 2nd International Seminar on Science and Technology*, Atlantis Press, pp. 121–129.

Verheul et al. 2022 Verheul, J. et al. (2022) “Using word vector models to trace conceptual change over time and space in historical newspapers, 1840–1914,” *Digital Humanities Quarterly*, 016(2). Available at: <https://www.digitalhumanities.org/dhq/vol/16/2/000550/000550.html> (Accessed: 10 November 2022).



This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.