DHQ: Digital Humanities Quarterly
Preview
2023
Volume 17 Number 3

Project Quintessence: Examining Textual Dimensionality with a Dynamic Corpus Explorer

Samuel Pizelo <spizelo_at_ucdavis_dot_edu>, UC Davis https://orcid.org/0000-0002-4311-2980
Arthur Koehl <avkoehl_at_ucdavis_dot_edu>, UC Davis https://orcid.org/0000-0003-0476-2144
Chandni Nagda <cjnagda2_at_illinois_dot_edu>, University of Illinois at Urbana-Champaign
Carl Stahmer <cstahmer_at_ucdavis_dot_edu>, UC Davis https://orcid.org/0000-0002-5714-3497

Abstract

In this paper, we present a free and open-access web tool for exploring the EEBO-TCP early modern English corpus. Our tool combines several unsupervised computational techniques into a coherent exploratory framework that allows for textual analysis at a variety of scales. Through this tool, we hope to integrate close-reading and corpus-wide analysis with the wider scope that computational analysis affords. This integration, we argue, allows for an augmentation of both methods: contextualizing close reading practices within historically- and regionally-specific word usage and semantics, on the one hand, and concretizing thematic and statistical trends by locating them at the textual level, on the other. We articulate a design principle of *textual dimensionality* or approximating through visualization the abstract relationships between words in any text. We argue that *Project Quintessence* represents a method for researchers to navigate archives at a variety of scales by helping to visualize the many latent dimensions present in texts.

INTRODUCTION[1] [2]

While the use of computational methods in corpus-level analysis has become ubiquitous in recent years, the optimal integration of these tools into digital archiving and corpus presentation remains an open question. Massive archives such as JSTOR, Gale, and Google Books have experimented with computational tools for traversing and exploring underlying documents in their collections.[3] Meanwhile, progress in machine learning (ML) and natural language processing (NLP) continues to advance the state-of-the-art for the underlying computational tools and methods these archives employ. However, there is a critical disjuncture between the datasets these latter tools and methods are trained upon and the real-world needs and expectations of digital archives [Jo and Gebru 2020] [Birhane et al 2022]. The application of computational methods to archival research thus risks both misrepresenting the findings of models and techniques, on the one hand, and misconstruing the exigencies of the underlying archive on the other. These concerns have resulted in a trend toward out-of-the-box computational tools that are not well integrated with the underlying archive and often obscure rather than explain interesting phenomena. Computational tools in archival analysis are thus frequently designed, implemented, and presented as abstractions of the archive. As such, these tools rarely serve as more than a heuristic method of inquiry, failing to meet the higher expectations of either computational humanities scholarship or humanistic textual inquiry. 1

We designed *Project Quintessence* with these concerns in mind. By bringing together a team of scholars intimately acquainted both with computational analysis in NLP and with the texts and historical context of our archive, this project has attempted to build a corpus exploration tool from the ground up that fully integrates computational tools and methods with the specific archive in question. For our corpus, we chose the Early English Books Online Text Creation Partnership (EEBO-TCP) archive, a collection of 60,331 texts transcribed and tagged with TEI markup by the TCP 2 team. Each model we trained and technique we incorporated was designed to take advantage of this highly idiosyncratic archive. Rather than decoupling textual analysis from the texts themselves, we incorporated these tools into the very architecture of the site. The result is a tool that allows one to dynamically interact with the archive in a way that is not possible with any one method independently, and which we believe presents a new model for archival research and analysis. By presenting multiple, complementary tools side-by-side for use by domain experts, we allow for the tools to be checked against each other and also compared with the underlying archive. We believe that this framework provides a stable site for the continual improvement of knowledge about the archive and domain alike.

In this paper, we first contextualize *Project Quintessence* within a longer trajectory of digital archival research tools, and then describe how the various elements of the project cohere by employing our concept of *dimensionality*. We continue this discussion by outlining the project pipeline, including model training parameters and other considerations for each of the models and tools included. Finally, we explore some potential future directions for our project and for database visualization frameworks in general. We argue that an attention to *textual dimensionality* allows for an interdependent database framework that augments research at any scale of analysis and has the capacity to both to confirm the doxa of pre-existing research while also providing surprising results that enrich the body of knowledge in the field. We believe that our web tool accomplishes all of these charges. 3

Project Quintessence is built using well-established, highly nimble modeling techniques; primarily Latent Dirichlet Analysis (LDA) Topic Modeling and Word2Vec Word Embeddings. By leveraging dynamic visualization libraries such as Plotly, we were able to extract a high level of archival information from each of our models. Each dynamic visualization on the site can accommodate any number of research questions or approaches. In addition, by synthesizing approaches from multiple techniques on the same archive, we were able to harness the affordances of each method rather than attempting to answer all questions with a single approach. We describe below how these methods differ from more contemporary techniques, such as Structural Topic Models (STM), Dynamic Topic Models (DTM), Document Embeddings, Contextual Embeddings (CoVe), Transformer Embeddings (such as BERT), and Stacked Embeddings (such as Flair); most of which we used at different stages of our design process to supplement our findings and enrich our analysis. Overall, the trend toward large language models and massive amounts of masked perimeters has introduced elements of bias and structuring assumptions which do not hold true for all datasets, disguising crucial inferences the model makes based on its training data rather than the object being interpreted [Bender et al 2021] [Chun 2008]. By incorporating historical lexicographic tools like MorphAdorner into our text cleaning pipeline and generating many smaller language models on specific cross-sections of the archive, we were able to preserve germane distinctions within the archive while still providing a mode of ingress into its analysis. 4

RELATED WORK

The use of visualizations in representing digital archives has become more nuanced with the recent increase in critical attention to these methods. Most of the early literature on data visualization in the Digital Humanities has remarked on the late adoption of visualization techniques by researchers in the humanities, a trend that appears to be justified when examining naive approaches to what visualizations can and do tell us of the underlying data, and the extent to which that lens is mediated and controlled by its designers. However, as media studies and digital humanities scholarship have produced compelling critiques of the unmediated view of visualization, alongside compelling design objectives for researchers and engineers, dynamic visualizations are increasingly seen as a site for emancipatory politics and critical inquiry. 5

To that end, we have framed our project’s intervention as one of harnessing new technological advances in ways that opens rather than closes archives to countervailing narratives and critique of canon. In our design practices, we asked how digital methods might appropriately interact with digital textual archives. Michael Witmore has aptly described this relationship in terms of addressability, and describes archives of texts as digital objects that are, “massively addressable at different levels of scale” [Whitmore 2010]. We suggest that the oscillations of scale afforded by dynamic exploration tools could be best considered by thinking in terms of dimensionality. As described in David J. Staley’s influential account, the capacity for visualization techniques to augment the affordances of text can be understood of as an increase in dimensionality [Staley 2015]. Staley frames visualization as a way of breaking out of the implied linear one- 6 dimensionality of linguistic syntax. However, we can expand upon this account. As Peter Gärdenfors has suggested, treating semantic meaning as a conceptual space with abstract, dimensional relations allows for complex semantic representation and analysis [Gärdenfors 2014]. Word embeddings, for instance, are commonly employed to locate clusters of semantically similar words by assigning spatial coordinates to those words based on hundreds of dimensions of training data. Following Gärdenfors, we do not treat dimensionality rigidly as a geometrical principle, contrasted with hierarchical or ordinal relationships, but instead as a higher-level abstract relationality of terms. By allowing ourselves to depart from three-dimensional thinking, dynamic visualizations are freed from their metaphorical relationship with abstract, Cartesian space, and instead function as tools for making tangible the many relationships between words with startling complexity.

Over the course of our project, we have increasingly become informed by this dimensional thinking. Describing syntactic, semantic, and thematic relationships in texts as dimensional relationships insists on the reality of those relationships. By virtue of the relational character of meaning-making, individual uses of words are necessarily and inevitably put into relationship with other words. While any computational attempt at measuring or visualizing those relationships is necessarily an approximation of varying validity, these computational tools are nevertheless measuring real relationships. Thus, we argue that maximizing researcher access to *textual dimensionality* should be a primary objective for database design and visualization. This emphasis on dimensionality has been influenced by many earlier digital projects that we will discuss in three groupings; large scale digital archives, small-scale digital projects, and pre-existing work on the EEBO-TCP corpus. Our objective was to design *Project Quintessence* as a coherent visualization tool that incorporated lessons from each of these categories. 7

Large-Scale Digital Archives

The first strain of projects that influenced our approach are digital archives that are sufficiently large that they cannot be grouped thematically and are not housed within a single physical location. With the release of increasingly massive language models in recent years, such as OpenAI’s GPT-3 or Google’s PaLM, the trend in data analysis has undeniably moved toward the analysis of larger and larger quantities of data with less concern for text processing, archive curation, or corpus-dependent techniques [Gebru et al 2021]. This trend is most evident in tools provided by Google for the analysis of the Google Books archive, comprised of five million books published between 1800-2000, or over half a trillion tokens. Yet perversely, the astounding breadth of materials both risks misrepresenting the actual published content during that period, while also serving as a barrier to more nuanced NLP techniques [Pechenick et al 2015] [Schmidt et al 2021]. This trend is also evident to a lesser degree with recent archival exploration tools, such as the Digital Scholars Lab produced by Gale and JSTOR’s Text Analyzer beta tool. While both of these methods effectively incorporate computational methods in ways that enhance search functionality and research inquiry, in both cases modeling techniques are utilized with minimal control over parameters and do not change to fit the limitations and possibilities of the underlying archive. Thus, neither tool is intended for or achieves the rigor of computational humanities scholarship. 8

Small-Scale Digital Projects

While *Project Quintessence* houses an archive of over sixty-thousand texts — over two billion tokens — it shares more in common thematically with smaller digital archives that aim for a higher level of research efficacy than the large-scale archives. These projects — often supported by research, institutions, and scholars of the digital and computational humanities — tend to emphasize curatorial control, transcription accuracy, representative sampling or strong thematic coherence of archives, and descriptive visualization techniques. Some examples of this style that were instructive to our work were Andrew Goldstone’s “dfr-browser,” the JeSeMe Semantic Explorer project, and the English Broadside Ballad Archive (EBBA). Each of these projects emphasize different aspects of the dynamic exploratory archive model that we believe are all synthesized in *Quintessence*. 9

Pre-Existing Work on EEBO-TCP

10 The third strain of influence upon our work has been projects specific to the Early English Books Online (EEBO) and

EEBO Text Creation Partnership (EEBO-TCP) archives. As a digital archive, EEBO grew from the microfilm series of Early English Books that began in the 1930s, but has since grown to encompass over 146,000 texts [McCollough and Lesser 2019]. Of these 146,000 texts, 60,331 have been transcribed using TEI markup through the EEBO-TCP project. EEBO is currently housed within the ProQuest digital archives, which has maintained the general emphasis on facsimile reproductions of physical items in the archive, and a limited boolean and keyword search functionality.

The greatest improvement on this current presentation of the archives has been undertaken in a joint project by Northwestern University and Washington University in St. Louis called Early Print. This archival project boasts both improvements to the EEBO-TCP transcriptions through inbuilt crowdsourcing functionality, along with a number of fascinating visualizations and experimentation with computational modeling to aid search and retrieval. Early Print represents an exciting development in EEBO-TCP digital projects, and has the most in common with *Project Quintessence* of the projects discussed in this section. Nonetheless, we feel that there remains a need for a generalizable framework that more fully integrates texts with their many contexts through computational tools. We also aim to better synthesize many pre-existing tools and visualizations into a single extensible research apparatus. As such, our emphasis is on providing a tool of sufficient rigor that its results can be relied upon for digital humanities research in the EEBO-TCP archive. 11

DIMENSIONALITY AT SCALE

Recent work in the digital humanities has rightly critiqued the over-reliance on spatial metaphors of analysis through terms like “close-reading” or “distant-reading” [Da 2019] [Bode 2018]. Project *Quintessence* fully incorporates these critiques into its design. As such, conventional notions of scale collapse entirely when using our web tool for textual analysis: at the same time as a researcher is engaging in a *close reading* of a text, they can inspect the semantic associations of any word in the document with other documents written by that author, in that decade, or printed at that print location. At the same time as one collects a personal archive of documents using our subsetting tool, one can also inspect the thematic topic distribution of that archive, and compare it to the distribution of topics in the corpus as a whole. These authorial, temporal, spatial, and thematic contexts allow for a renegotiation of how scale operates in textual analysis. Rather than having to choose between the closeness of a “human scale” or the distance of a “computational scale,” *Quintessence* instead encourages a multiscalar view of the corpus by emphasizing different units of analysis — word, document, or topic — each providing a unique approach to exploring the entire archive. 12

In summary, *Project Quintessence* de-emphasizes discussion of *scale* that predominate in digital archival contexts, favoring instead a focus on *dimensionality* . Prompted by the computational tools incorporated into our project, we describe token-, document-, and topic-level relationships in the archive as real and meaningful discursive relationships that can be explored through quantitative and inferential analysis despite not being reducible to the latter. The goal of computational modeling and data visualization can and should be the expansion of textual dimensionality in productive and meaningful ways. Moreover, we believe these visualizations should prioritize utility and interpretability for actual domain experts. That is why we chose tools and designed the site based on our own research priorities and those of our advisory team. In conversations with early users of *Project Quintessence*, we have observed that the site becomes more useful the better one understands the EEBO archive; the inverse trend of many out-of-the-box methods. 13

Elements of *Project Quintessence*

While a full catalog of all features of *Project Quintessence* exceeds the bounds of this paper, we outline here the two general modes of ingress into the EEBO-TCP archive that our site affords: a “Topic Exploration” view and a “Word Exploration” view, with both views connecting back to the individual documents themselves. These two views incorporate multiple interactive and dynamic visualizations using statistical models trained on the documents of the EEBO-TCP archive. These modes are meant to be used in conjunction when the site is used for research purposes. By synthesizing multiple visualizations users can better uncover patterns, identify what each model has learned from the archive, and better understand the features of the archive itself. 14

Topic Exploration

Since the emergence of topic modeling in the toolkits of Digital Humanities researchers a decade ago, it has been considered an archetypal *distant reading* method [Meeks and Weingart 2012]. The presentation of topic exploration in *Project Quintessence* recasts this association entirely. By integrating several different dynamic visualizations with database search, subsetting, and indexing functions, the “Topics” page allows a seamless toggling between the topics themselves, search keywords, author, print date and location metadata, and hand-tagged text keywords from the EEBO-TCP team. Rather than a static abstraction of the archive, the topic model on our site depicts abstract relationships between each of these categories as tangible visual dimensions. Connecting the topics learned with our topic model with document level metadata opens opportunities for users of the site to explore the EEBO-TCP corpus from the top down and the bottom up. In addition, the interactive and dynamic visualizations allow for greater ability to judge and understand the models’ output. Lastly, the visualizations connect with the other models and visualizations of the site, to allow users to leverage the strengths of each model, and critically evaluate each model’s limitation. While corpus visualizations often invite an impression of cohesion and totality, the organization of our site works against this tendency. The limitations of each of our models can be directly explored within the site’s framework. 15

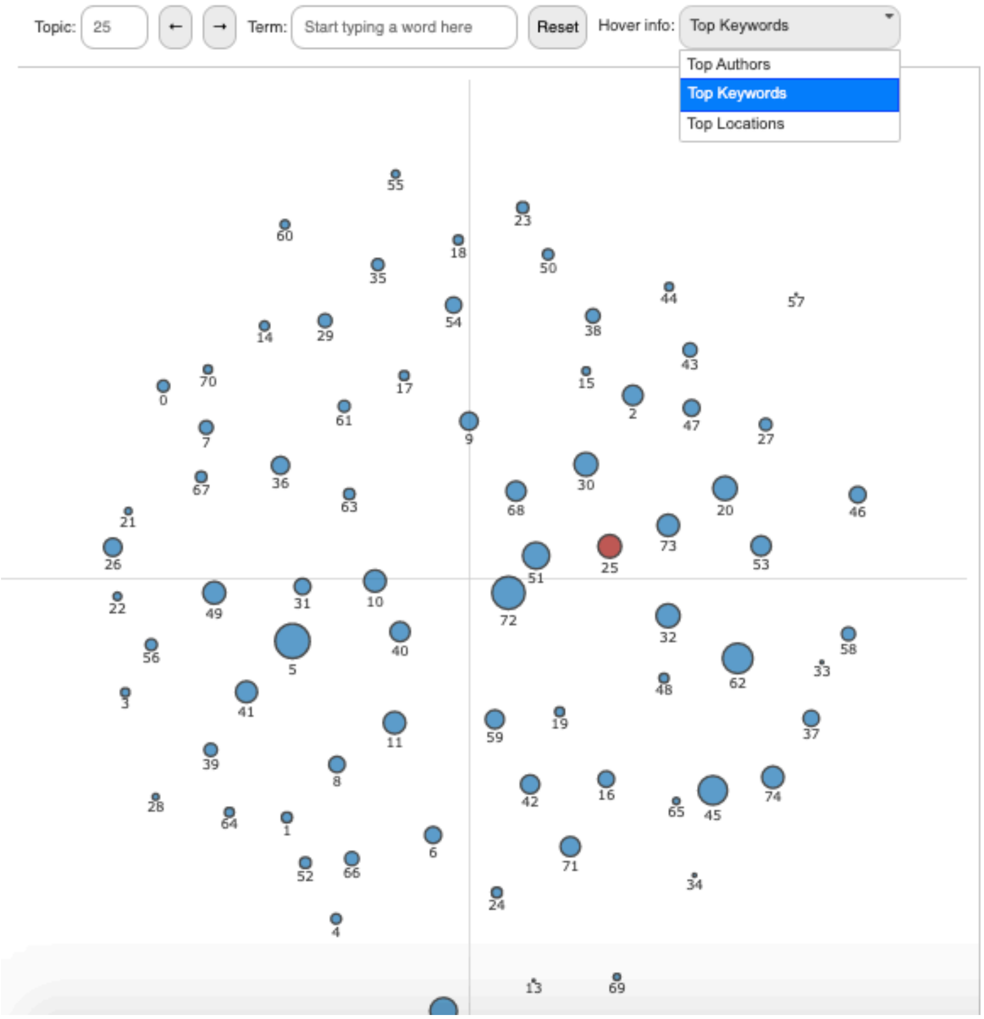
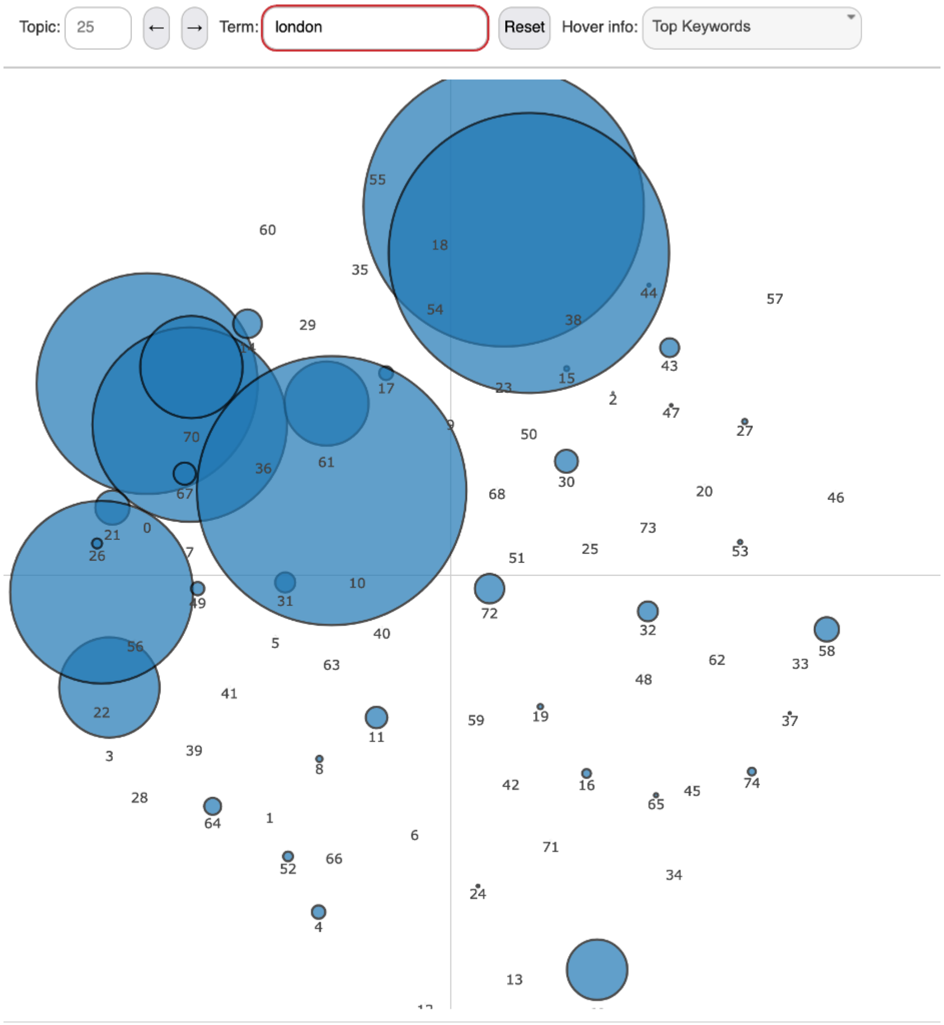


Figure 1. Text Topics view. Each bubble represents one of seventy-five topics and is numbered accordingly.

We trained a topic model with seventy-five topics on the EEBO-TCP corpus: the topics are displayed in Figure 1 above as nodes in a bubble plot. Each node in the bubble plot represents a single topic scaled according to the proportion of the archive that was assigned to that topic. Proportions were estimated using the method described in the LDAvis paper [Sievert and Shirley 2014].[4] The nodes are placed based on the coordinates of each topic mapped to two-dimensional space according to the calculated Jensen-Shannon (JS) divergence of topic term probabilities.[5] As a result, node proximity translates to similarity in topic composition. This means that nodes are roughly grouped based on topic 16 similarity. This feature alone allows for novel observations of the archive: religious topics are close to philosophical and poetical ones, and scientific texts are close to nautical guides and recipe books.

17 To aid visual exploration of the topic model, we have included hover-text for each topic, which can be modified by the user in the top-right dropdown to display the associated metadata that best fits their research question: “Top Keywords,” “Top Authors,” or “Top Locations” of text publication.



Topic: 25 Term: london Hover info: Top Keywords

Figure 2. Text Topics view. Topic bubbles are scaled according to importance of search term for topic.

18 In addition to displaying the topic proportions for the archive, the scatter plot can be rescaled to reflect the topic proportions for each word (in this case the number of times the word has been assigned to each topic – see Figure 02). The display above reflects how the scale of each node is dynamically shifted to reflect the importance of each topic for the searched word.

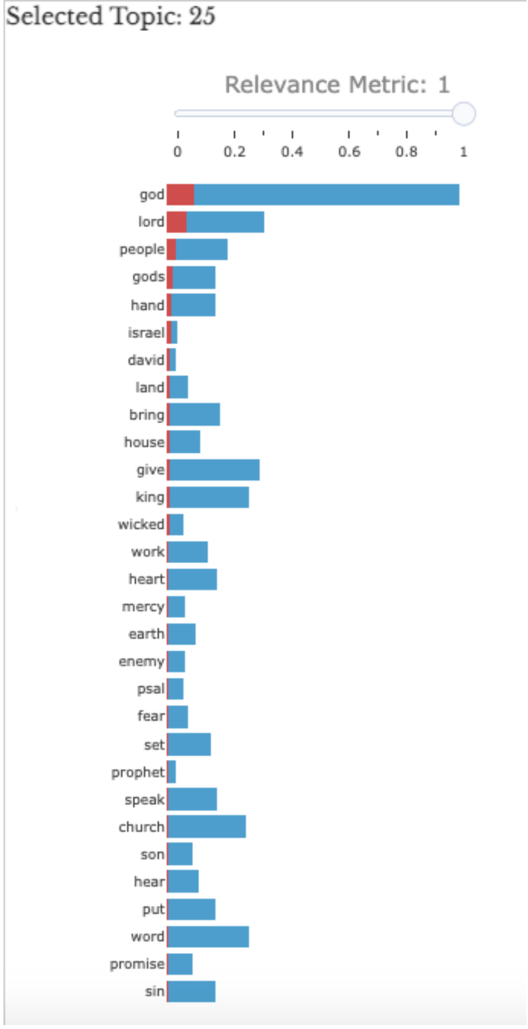
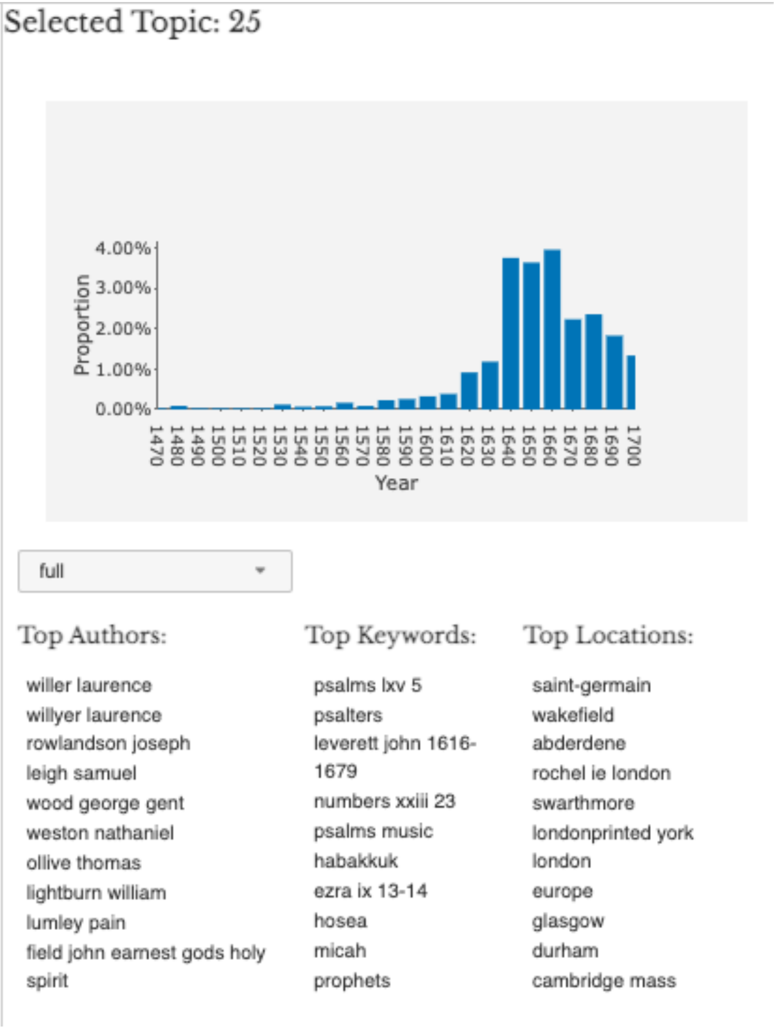


Figure 3. Topic terms histogram. Words are ranked by importance to topic. Adjusting the relevance metric reweights the scores for each term, the lower the value for the metric the more words that are unique to that topic are boosted according to the formula found in Sievert and Shirley 2014.

19

To the right of the topic visualization proper is another companion visualization (Figure 03), which depicts the thirty most important words for that topic. The red bar in this stacked bar plot represents the prevalence of the word within the selected topic, while the blue bar represents the total occurrences of the word throughout the entire archive. To better understand the particularities of a specific topic, we have implemented a “Relevance Metric,” also modeled on the LDAvis package, which allows the user to preference words that are uniquely important to the selected topic (e.g., “promise”) rather than words that tend to be overrepresented across the archive and are thus important to many topics (e.g., “god”) [Sievert and Shirley 2014].


Selected Topic: 25

Top Authors:

willer laurence
willyer laurence
rowlandson joseph
leigh samuel
wood george gent
weston nathaniel
ollive thomas
lightburn william
lumley pain
field john earnest gods holy spirit

Top Keywords:

psalms lxv 5
psalters
leverett john 1616-1679
numbers xxiii 23
psalms music
habakkuk
ezra ix 13-14
hosea
micah
prophets

Top Locations:

saint-germain
wakefield
abderdene
rochel ie london
swarthmore
londonprinted york
london
europe
glasgow
durham
cambridge mass

Figure 4. Topic info tab. Histogram displays topic proportion of entire corpus per decade in percentages.

20 But again, the design of the topics page insists on a contextual understanding of each topic. The “Info” tab for each topic will display the proportion of the archive captured by that topic in each decade and the top authors, text keywords, and print locations associated with that topic (Figure 04). In addition, the latter metadata categories can be explored for each decade of the archive. This feature might allow a researcher to answer questions such as; “which print locations were printing herbals in the 1620s?”, or “which authors most heavily employed New Testament biblical language around 1700?” These increases in the textual dimensionality of the EEBO-TCP archive help researchers forge new relationships with the texts themselves.

Topics Over Time

21 The EEBO-TCP archive contains texts from 1473-1700. We provide users two ways to explore the evolution of the topics covered in the archive over this time range. A well-known limitation of LDA Gibbs topic modeling is that the model does not account for the date of publication when analyzing topic composition [Blei and Lafferty 2007] [Roberts, Stewart, and Tingley 2019]. This imprecision sometimes results in topics that group together words despite linguistic shifts in the meaning of words and changes in local context of use at the sentence- or paragraph-level of a text. This effect is made worse by the uneven distribution of documents across time within EEBO-TCP, which contains significantly more documents starting in 1640 than the previous years. As a result, an abstracted static display of the topic proportions of our archive would collapse much of the potential information about the themes present in the underlying texts, especially from documents published before 1640. However, by maintaining a relationship between multiple dynamic visualizations and the texts themselves, and by presenting topic modeling as one of several modes of obtaining archival information, *Quintessence* strikes a balance between using language models as research tools while

foregrounding the limitations of such models.

22 In the same ‘Info’ tab described in Figure 4, we provide a visualization of the selected topic’s proportion within the archive over the full range, shown above. Thus, for each decade, from 1470-1700, we compute each topic’s proportion of the subset of EEBO-TCP that was published within that decade.

23 To further aid in the exploration of these trends, we link the topic proportions back with the original bubble plot that visualizes both topic similarity and topic proportions. On the “topics over time” page, we provide an animation of the computed topic proportions for each decade. By playing the animation a user can see the bright spots on the quadrants of the bubble plot grow and shrink as the associated topics change in proportion. This animated plot can also be paused and resumed, and any decade can be manually selected.

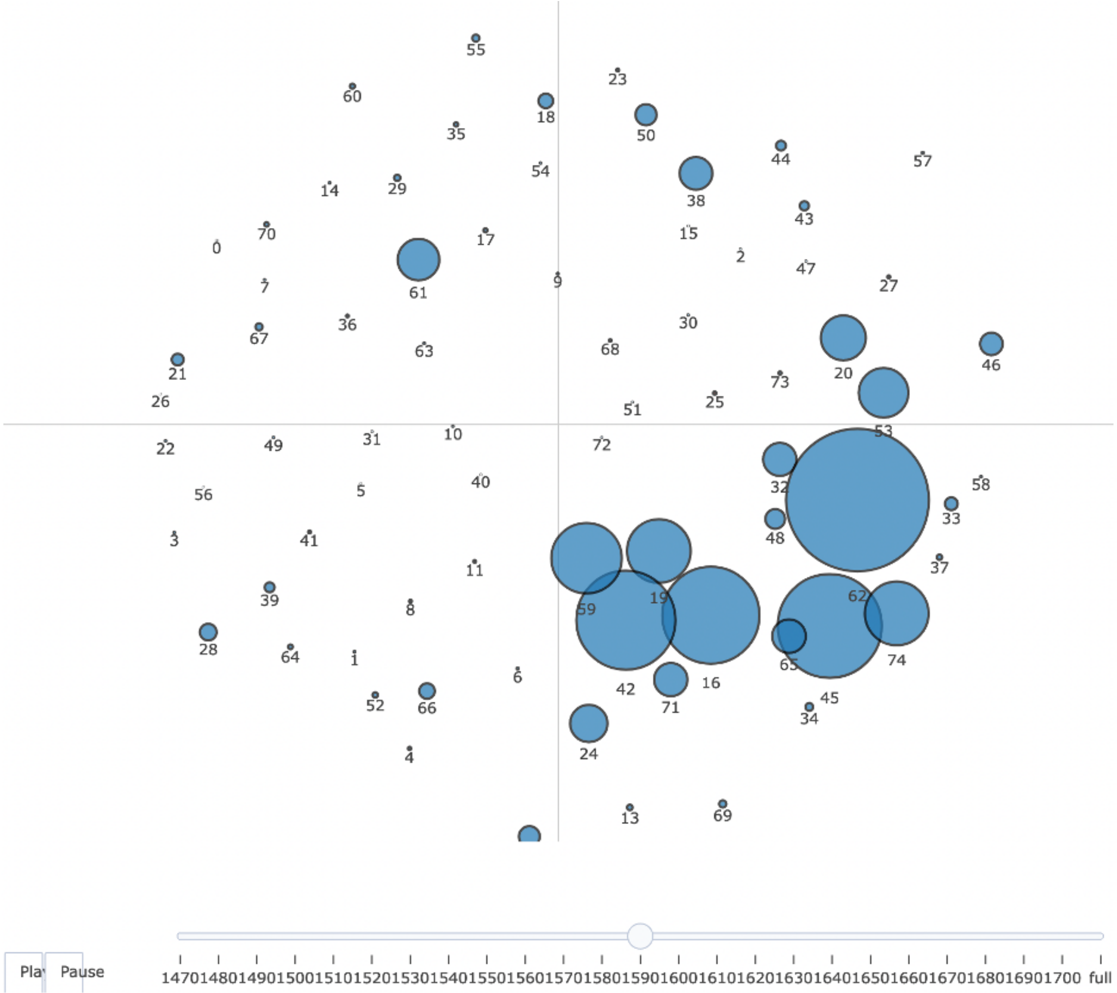


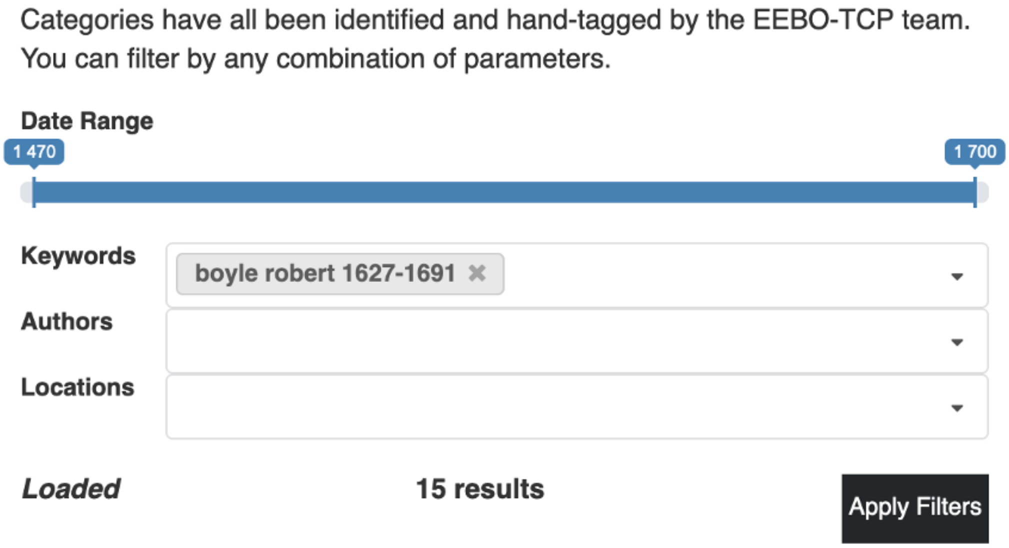
Figure 5. Animated view of topic composition proportion by decade. Each bubble represents one topic as labeled.

24 The animated view of the topic model depicted in Figure 5 helps to illustrate in a tangible way changing representations in the archive over the time frame analyzed. While a static topic model has great utility for exploring unexpected textual relationships across the archive, this presentation of the texts risks abstracting somewhat-arbitrary topic assignments from historical changes and contingencies in the texts themselves. We believe this time series animation helps remind the user of these elements of the archive.

Topic Document Exploration

25 Due to the dynamic design of the topic model visualization, we were able to provide researchers an additional capability: our subset function allows one to create an individualized research archive out of any combination of texts; filtered by

date range, keyword, author, or location.



Categories have all been identified and hand-tagged by the EEBO-TCP team. You can filter by any combination of parameters.

Date Range

1 470 – 1 700

Keywords boyle robert 1627-1691 ✖

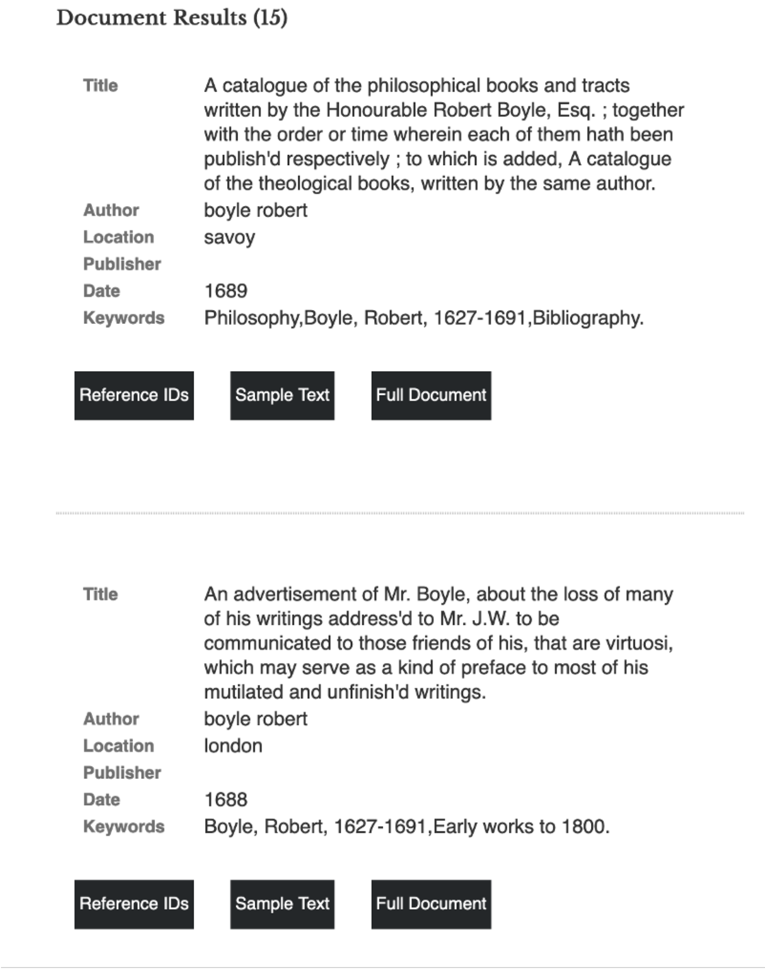
Authors

Locations

Loaded 15 results

Figure 6. Topic subset settings. Date range slider is in decades. Keywords, authors, and locations taken from EEBO-TCP metadata tags.

Once the filter for this research archive has been applied to the model, the topic visualization will display the topic proportions for whatever collection of texts are included in the selected subset (Figure 06). This tool allows researchers to examine the prevalence of religious language in the texts of Robert Boyle, for instance, or assess at a glance the topic similarities between William Shakespeare, Christopher Marlowe and Ben Jonson. 26



Document Results (15)

Title A catalogue of the philosophical books and tracts written by the Honourable Robert Boyle, Esq. ; together with the order or time wherein each of them hath been publish'd respectively ; to which is added, A catalogue of the theological books, written by the same author.
Author boyle robert
Location savoy
Publisher
Date 1689
Keywords Philosophy,Boyle, Robert, 1627-1691,Bibliography.

Title An advertisement of Mr. Boyle, about the loss of many of his writings address'd to Mr. J.W. to be communicated to those friends of his, that are virtuosi, which may serve as a kind of preface to most of his mutilated and unfinish'd writings.
Author boyle robert
Location london
Publisher
Date 1688
Keywords Boyle, Robert, 1627-1691,Early works to 1800.

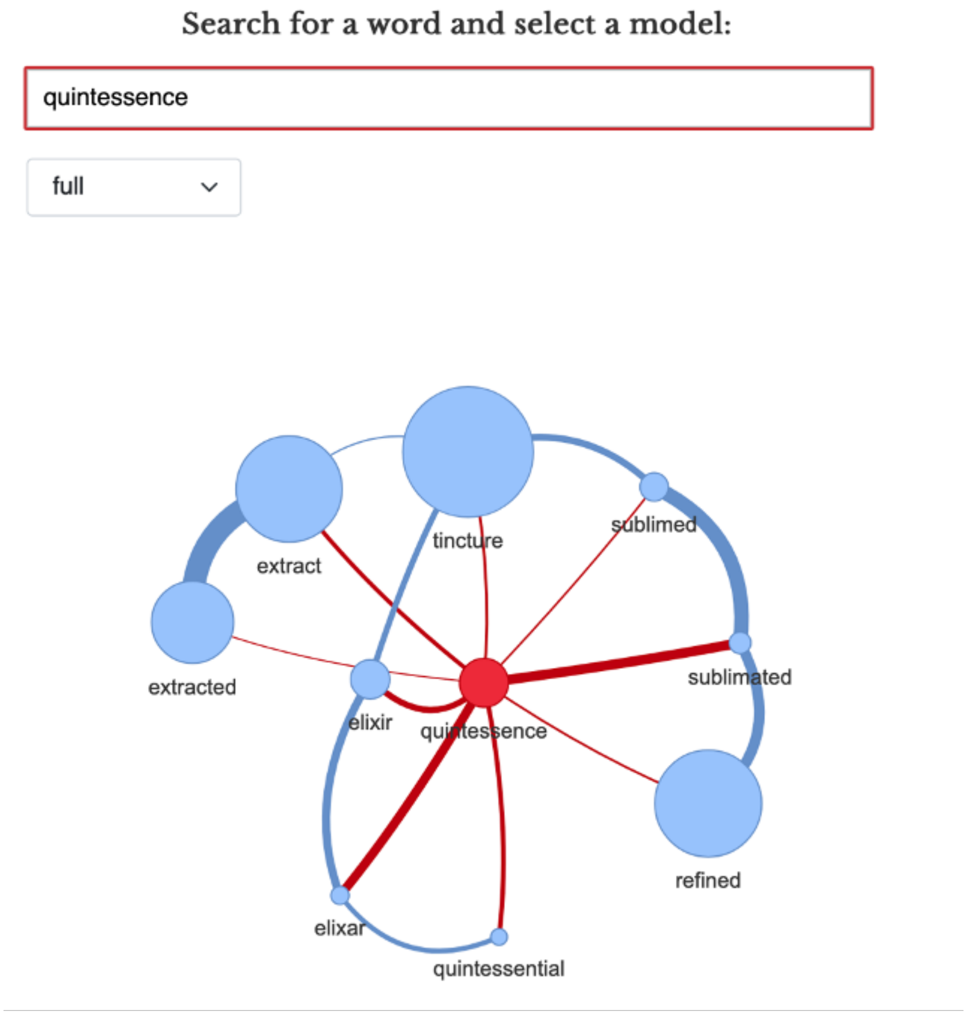
Figure 7. Document subset view (continued). All filtered documents appear in a list with associated metadata.

27 But due to the persistent nature of the subsetting function, one can also scroll past the model itself to see each text included in the selected research archive (Figure 07). This list displays all available metadata, a sample fragment of the document, as well as a link to the full document in our database. The simultaneous access to dimensional representations of discursive relationships in the archive as well as the underlying documents from which those relationships were derived allows for *Project Quintessence* to reorient common assumptions about depth and proximity in Digital Humanities scholarship. As we will return to at the end of this article, future directions for this project will emphasize making accessible these discursive relationships in the document reader itself, thus emphasizing dimensionality in close-reading practices in addition to the visualizations themselves.

Word Exploration

28 Topic modeling provides an efficient way of summarizing an unreadable amount of text. However, on its own, a topic model does not enable nuanced exploration of language evolution in the EEBO-TCP. This is due in part to the fact that many of the characteristics of the EEBO-TCP archive come in direct conflict with some common constraints of the model’s statistical assumptions – namely, variance of document length, variance in time of publication, a reliance on a “bag-of-words” representation of documents (that is, removing metadata and token sequence when examining textual relationships), and an oversimplification of terms as static entities (ignoring polysemy and semantic shift). To ameliorate this shortcoming in an otherwise useful modeling method, we incorporate word embedding models, which are trained using a sequential, *skip-gram* learning process, where the local context of the word is the primary information ingested by the model.

Another limitation of topic modeling that word embeddings avoids is the relative arbitrariness of topic assignments. 29 While we relied on a series of well-known metrics to determine an optimal number of topics for our model, the design of an LDA topic model necessitates prior determination of certain hyperparameters that are imposed upon the text from outside, rather than inferred from the texts themselves. We used the *ldatuning* package in R by Nikita Murzintcev, which combines metrics designed to prioritize low topic entropy and high topic diversity.[6] Ultimately, the metrics we prioritize will result in several candidate topic counts that will present similarly useful topic distributions. Rather than choosing an exceedingly large topic count that renders visual model analysis exceedingly difficult, we chose 75 topics, which struck a balance between metric optimization and readability. We help to correct for many of these limitations with our dynamic visualizations. However, we also supplement topic models with word level exploration. This includes visualizations of word frequency, views of words in context, and longitudinal word2vec models of the archive. Through the use of these visualizations, researchers can explore the semantic evolution of terms, distinctions in language usage across documents and metadata, and the broader thematic contexts of topics learned from LDA Gibbs.



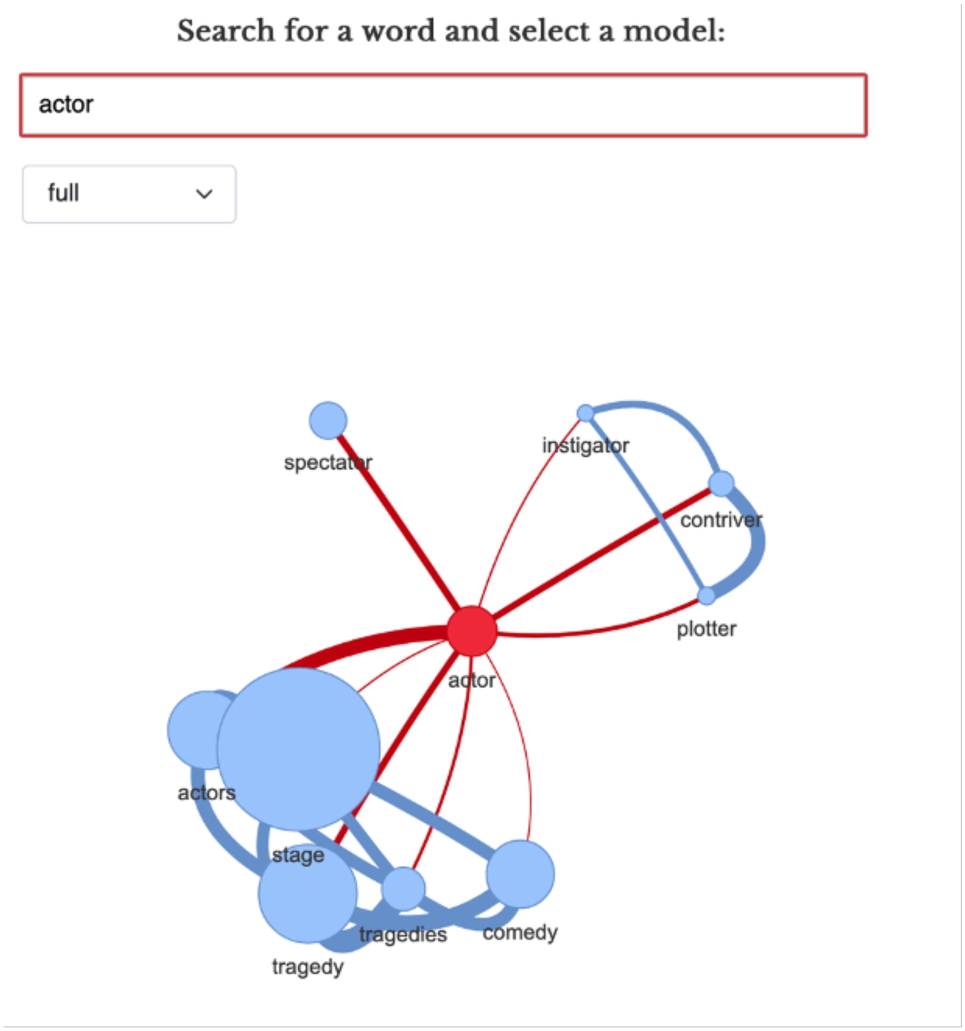
Search for a word and select a model:

quintessence

full

Figure 8. Word Meanings view. Search word appears at center of network plot with nine most similar words. Bubble size scaled for word frequency in selected model.

By training hundreds of word embeddings models on subsets of our corpus grouped by author, print location, and 30 decade of publication, we have obtained data on contextual distinctions in usage patterns for approximately four hundred thousand unique tokens. Gathering contextual word use data on this high number of words would be intractable with topic modeling for most computational humanities labs due to the enormous amount of computational resources necessary to dedicate to this process. Because word embeddings output dimensional coordinates for each word in the model, they also allow for a measurement of the similarity between words in the learned embedded space. The network plot of the search term “quintessence” depicted in Figure 8 allows for a comparison of the term usage across every text in the archive with the usage of the top ten most similar words [Katricheva et al 2020]. The size of each node reflects total occurrences of the word in question, and the width of each edge reflects the similarity between the two connected nodes.



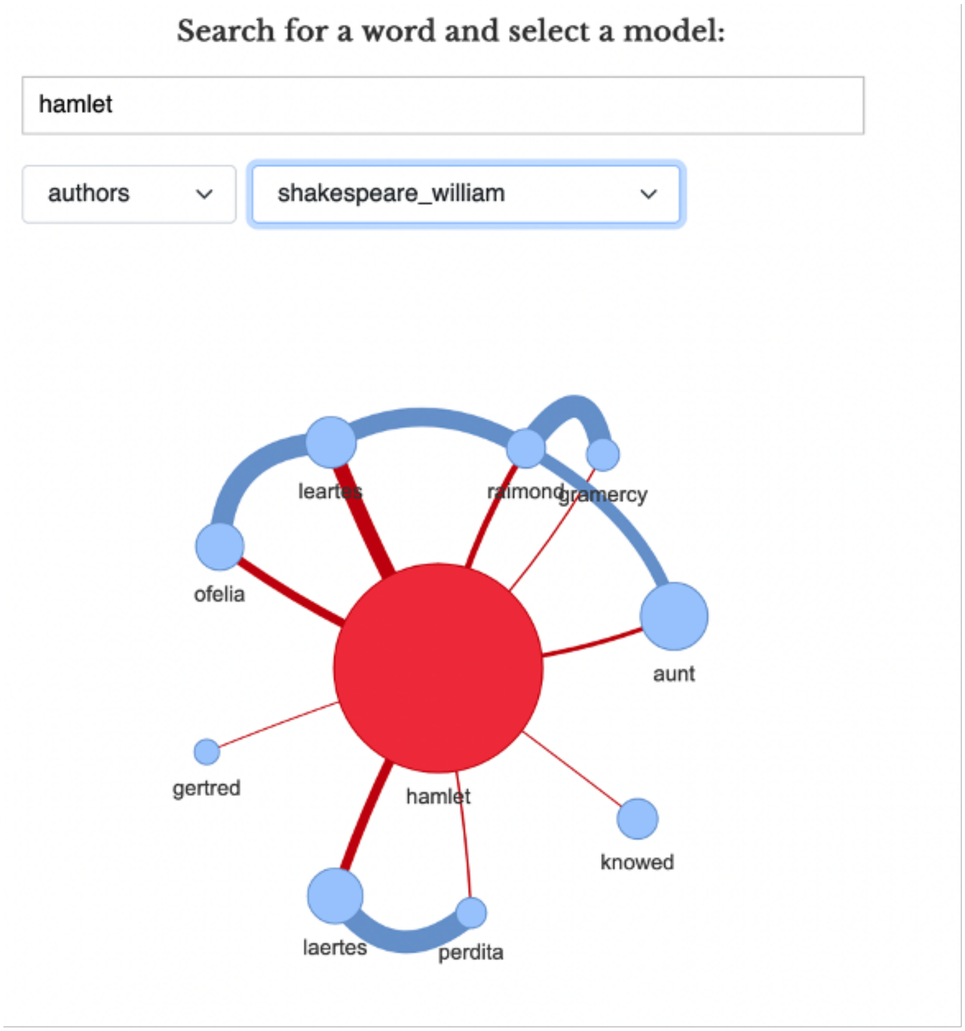
Search for a word and select a model:

actor

full

Figure 9. Word Meanings view. Word clustering suggests polysemy of search term.

31 Due to their contextual specificity, word embeddings also emphasize another linguistic feature missing in most topic modeling representations: the allowance for polysemy. The network graph of the word “actor” in Figure 9 illustrates this feature. By clustering nearest words to the word being examined, this network representation allows one to see when a word has obvious distinctions in use. “Actor” usually denotes a dramaturge in our archive, but also has a conspiratorial valence (e.g., “instigator,” “plotter”). Finally, the presence of the word “spectator” demonstrates the high level of interchangeability between antonymous words.



Search for a word and select a model:

hamlet

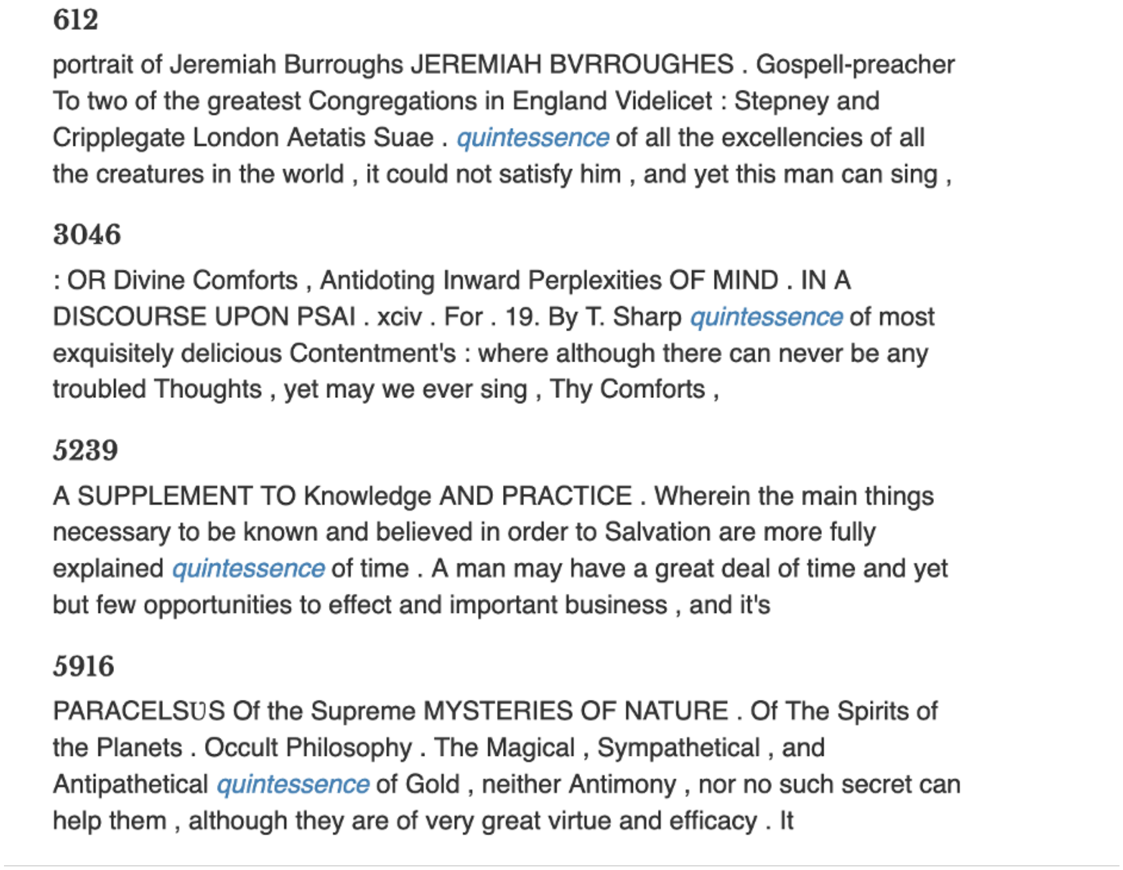
authors

shakespeare_william

Figure 10. Word Meanings view. Network graph of term “hamlet” in the William Shakespeare model.

32 Another feature of Quintessence, due to the high number of smaller embeddings models we have trained, is the ability to examine the oeuvres of specific authors to examine the semantic relationships of their idiosyncratic word usage. In Figure 10, we can see that William Shakespeare uses the word “Hamlet” with respect to certain other proper names and character titles, indicating the term takes on a unique meaning in the Shakespeare model.

Word in Context



612

portrait of Jeremiah Burroughs JEREMIAH BVRROUGHES . Gospell-preacher To two of the greatest Congregations in England Videlicet : Stepney and Cripplegate London Aetatis Suae . *quintessence* of all the excellencies of all the creatures in the world , it could not satisfy him , and yet this man can sing ,

3046

: OR Divine Comforts , Antidoting Inward Perplexities OF MIND . IN A DISCOURSE UPON PSAI . xciv . For . 19. By T. Sharp *quintessence* of most exquisitely delicious Contentment's : where although there can never be any troubled Thoughts , yet may we ever sing , Thy Comforts ,

5239

A SUPPLEMENT TO Knowledge AND PRACTICE . Wherein the main things necessary to be known and believed in order to Salvation are more fully explained *quintessence* of time . A man may have a great deal of time and yet but few opportunities to effect and important business , and it's

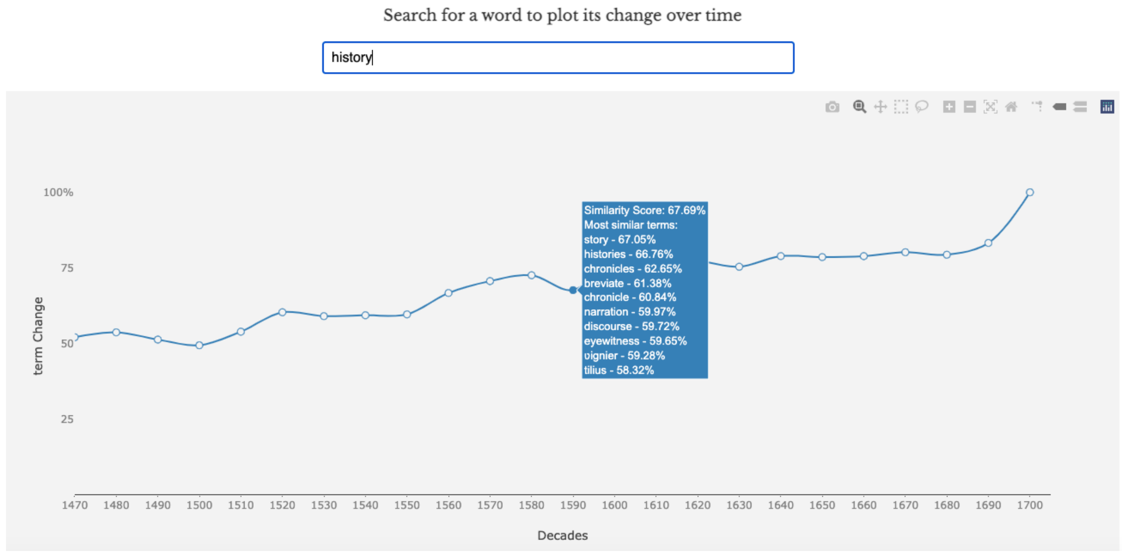
5916

PARACELSUS Of the Supreme MYSTERIES OF NATURE . Of The Spirits of the Planets . Occult Philosophy . The Magical , Sympathetical , and Antipathetical *quintessence* of Gold , neither Antimony , nor no such secret can help them , although they are of very great virtue and efficacy . It

Figure 11. Word Meanings view. Keyword in context for search term. Documents labeled by document number.

33 Exploration of individual words in the corpus is also facilitated by the Word in Context tab, which gives contextual word usage data for every instance of the chosen word in the corpus (Figure 11). These results are filtered by *tf-idf* scores (term frequency-inverse document frequency) to ensure the top results are highly relevant to an exploration of the word’s usage. By displaying this textual context alongside our word embeddings model outputs, our site maintains a relationship between machine learning abstractions and close reading techniques.

Word Semantic Shift



Search for a word to plot its change over time

history

Similarity Score: 67.69%
Most similar terms:
story - 67.05%
histories - 66.76%
chronicles - 62.65%
breviate - 61.38%
chronicle - 60.84%
narration - 59.97%
discourse - 59.72%
eyewitness - 59.65%
vignier - 59.28%
tilius - 58.32%

Figure 12. Word Meanings view. Search term represented for each decade according to percentage similarity with term usage in 1700 model. Precise similarity depicted as “Similarity Score” on mouseover, along with most similar terms.

Another function word embeddings allow for is the alignment of different models to gauge the relative position of words in their respective subset. Such a method estimates the comparative usage and semantics of the same word in two different subsets of an archive. This has classically been implemented only for time-slices in historical archives, and is thus termed “diachronic embedding” [Hamilton et al 2016]. *Project Quintessence* has implemented an experimental beta visualization that allows one to explore an aligned embedding space comprising the twenty-four decades in the EEBO-TCP archive. The visualization depicted in Figure 12 can be uniquely generated for over sixty thousand words in our archive, and represents an alignment between each decade and 1700 — the last decade captured in the archive. The higher the “Similarity Score,” the more similar the word is to its usage in 1700. Sudden peaks or valleys in the plot usually signal a semantic shift in that word’s usage during that time period. The node for each decade can be hovered over to display the Similarity Score along with the top ten most similar words in that decade. 34

While recent empirical analyses of such embedding alignment methods have cast some doubt upon their reliability [Hellrich et al 2018] [Hellrich 2019]), and thus a final version will need to overcome certain pitfalls, we have observed that this tool, when combined with document retrieval and keyword in context functionality, provides a unique representation of the evolution of language across historical time. In addition, the textual metadata of the EEBO-TCP will allow us to use similar alignment methods to compare word usage similarity between authors, in different regions, and more. We are not aware of any other attempts to implement such longitudinal linguistic analysis. 35

In summary, the three primary modes of analysis afforded by Project *Quintessence* are word-specific analysis, topic-level patterns, and document-based inquiry. Rather than understanding these as operating at three distinct scales of analysis, the dynamic design of our site allows an interpolation of these methods into an integrated dimensional analysis. Word usage and semantics are available during a close reading; individual documents can be examined while investigating archive-wide thematic trends; and the gradual semantic and contextual shifts of a word sit alongside the actual uses of that word in the texts underlying our computational analyses. We believe that this rich conjunction of visualizations and tools provides a new model for digital archive presentation and scholarly analysis in the digital humanities. 36

Project Pipeline

While idiosyncrasies found in the EEBO-TCP corpus are the norm for large textual archives, they present unique challenges and opportunities for digital humanities research. This is in no small part due to the reliance on uniform, carefully manicured datasets for the development of nearly every method of computational analysis. Increasing attention has been paid to the reproducibility of language models tuned to a few industry-standard datasets that bear little 37 resemblance to those encountered in archival research [Jo and Gebru 2020]. While models improve on benchmark performance year-by-year, it becomes increasingly difficult to describe the nature of those improvements or to predict the performance of competing models in real-world contexts.

However, these discussions on archive construction rarely broach the question of data asymmetry, enormous vocabulary sizes, or multilingual texts. All of these aspects are present in the EEBO-TCP corpus. The largest document in the corpus contains nearly 3.1 million words, whereas the smallest document is only 5 words long. There are over 2.5 million unique words in the corpus. After aggressive spelling variant reduction, spell-correction, and lemmatization, that number still only shrinks to 1.5 million. Even after filtering out texts tagged in primary languages other than English–of which fourteen are represented–the corpus still contains frequent block quotes, references, and terms in Latin, Greek, French, and occasionally other languages. For these reasons, EEBO-TCP becomes an exemplary corpus for computational research, not for its uniformity, but for precisely the opposite reason. 38

Development of a coherent toolset for *Project Quintessence* thus necessitated an iterative, empirical process. Each word stemmer and lemmatizer had to be compared statistically and hand-checked for overall accuracy. The same held true for clustering algorithms, word embedding models, topic models, and network analysis. This arena of labor, so often ignored by out-of-the-box toolkits and invisible in massive search interfaces, necessitated the care and expertise of scholars educated in early modern literature and culture. Thus, our project pipeline incorporated at every stage a negotiation between conventional literary knowledges and the production of computational tools. 39

The first stage of our pipeline involved text cleaning. Due to the crowdsourced nature of the EEBO-TCP project, we encountered inconsistencies in character encodings as well as TEI tagging schemata. Thus, we had to encode all texts into UTF-8 and include several automated character alterations to ensure each encoding was handled properly. As we wanted to maintain reproducibility of research findings for the archive itself, we adopted a minimalist approach to textual and tag alterations. We intend to make available an open-source version of our text cleaning and modeling source code to maximize transparency of results. Tags that were either too widely used or too seldomly used were winnowed to preserve functionality. In order to model linguistic change over time and obtain decade-specific language models, we also had to normalize dates by converting Roman numerals and reducing date ranges to a single four-digit year.[7] Word spelling corrections and standardizations, along with word stems, lemmata, and part-of-speech information, were obtained by using a locally-run build of MorphAdorner–a bespoke NLP software tool trained on an early modern archive. 40

The second stage of the pipeline was dedicated to database architecture and API development. While we relied on relational database configurations in early versions of the project, the performance gains and nimbleness of MongoDB ultimately outweighed the intuitive structure of SQL and similar configurations. Because each visualization and search function on our website accesses different elements of the text database and models, we found database calls impossible to construct efficiently enough to maintain speed and performance. This performance was aided after our implementation of a RESTful API written in PHP to serve as intermediary for database calls. In addition to these obvious end-user functionality gains, our shift to MongoDB (a popular document-oriented database program) resulted in a drastic reduction of compute-time, as new versions of the database could be iterated without a complete reindexing. This flexibility gives us the opportunity to grow our archive over time without the need to start from scratch and debug queries with each alteration. 41

The third stage of the pipeline was dedicated to modeling and model optimization. While we experimented with many different topic models and word embedding techniques in preliminary stages, we decided to use Latent Dirichlet Allocation (LDA) topic modeling with posterior Gibbs sampling for our base topic models. Our topic models were computed using the Gensim wrapper to MALLET, a Java-based machine learning toolkit dedicated to NLP modeling. We optimized the total topics in the model according to a balance of metrics provided by the LDATuning package in R. We have found that what is sacrificed in performance and possible insight is made up for in reproducibility, field familiarity, efficiency and flexibility. As our corpus grew to over two billion words, the compute-time for LDA modeling quickly grew unfeasibly large with several modeling techniques.[8] 42

A different issue was faced with our choice of word embedding model. While BERT, ELMo, CoVE, and other "state-of- 43 the-art” techniques boast much higher performance on carefully-tagged and uniform models in “resource-rich” contemporary languages (e.g. English, French, German), it is uncertain how this performance translates to different historical and multilingual contexts. We currently lack a sufficiently curated early modern database to support training a similar transformer or tag-dependent model from scratch. Thus, the potential for historical bias to influence model outcomes is great, and our ability to detect such distortions is highly limited. Our choice to use Word2Vec for our embeddings models was also influenced by the highly nimble nature of the technique, which allows for reliable word vectors in corpora as low as two million words. This allowed us to train smaller embedding models for subsets of our corpus based on EEBO-TCP metadata tags, including; Author, Location, and Publication Decade.

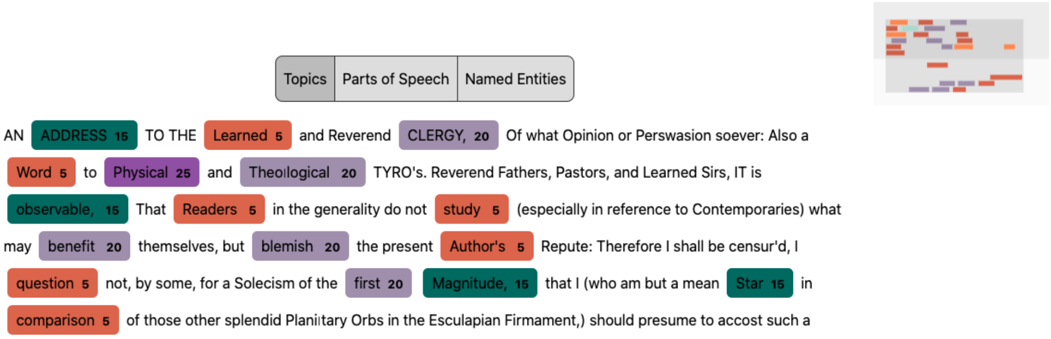
To preprocess the corpus for our Word2Vec models, we tokenized documents into sentences using the NLTK Sentence Tokenizer. We trained embeddings models using 250 dimensions, with a skipgram window of 15. For all author, location, and decade models, we trained models on any subset greater than two million words in size. For our diachronic embeddings visualization, we used procrustes alignment on the word vector for each decade model it appeared in, beginning with the last decade in our archive (1700) and working backward to leverage the richer spatial orientation of a high-vocabulary model.[9] Because larger language models are generally intended for word prediction and translation, they are built by design to leverage pre-existing model data to predict a single, correct meaning for each word usage. While it is perhaps possible to counteract this tendency with a nuanced implementation, we determined that the loss in accuracy of idiosyncratic word usage data was far greater than any potential gains in a naive implementation of such a method. 44

With the website construction, emphasis was placed on creating virtual environments to most effectively toggle between scales of the database. The website itself was built with PHP and JavaScript, with a sparse UI and accessible color palettes. We worked on UI/UX in consultation with Kimmy Hescock at the UC Davis Library to present users with a learnable visual grammar that combined the many different visualizations into a single text-exploration environment. With this in mind, we found the open-source dynamic visualization library, plotly.js, particularly useful.[10] In choosing specific visualizations from the many possibilities, we wanted to minimize redundancy or contradictions in the data presented, maximizing instead the recombinant affordances of using the many tools in conjunction to glean research insights from the database. Thus, we also worked periodically with test users researching early modern literature to better understand what sorts of research questions for which a scholar might approach this archive, and to imagine ways of accommodating the exploration of those questions. 45

Due to our prioritization of accuracy, reproducibility, and flexibility, we have built a framework for interactive language modeling that persists beyond any given implementation. While our methodology insists on a close relationship between area experts and data scientists, we nevertheless have produced a framework that allows for continual additions to and alterations of our archive. This allows both internal and external users and researchers to participate in the construction of more representative and inclusive historical archives, and safeguards our project from ossifying a dated snapshot of textual history. Looking forward, we also anticipate that this framework can be used similarly on other textual archives to similar effect. While archives should not be treated as interchangeable by data scientists and engineers, the agility of our framework nevertheless facilitates its tuning on archives beyond the EEBO-TCP. 46

CONCLUSION AND NEXT STEPS

We hope in this paper and through the *Quintessence* site to have contributed to the design of future archival research frameworks through an emphasis on dimensionality and dynamic exploration that interfaces with texts. As with any project at this scale, limitations on resources and time have necessarily influenced the scope of the site itself. Nevertheless, continued advancements in textual analysis tools present many exciting opportunities for future enhancements. 47



AN ADDRESS 15 TO THE Learned 5 and Reverend CLERGY, 20 Of what Opinion or Perswasion soever: Also a
Word 5 to Physical 25 and Theological 20 TYRO's. Reverend Fathers, Pastors, and Learned Sirs, IT is
observable, 15 That Readers 5 in the generality do not study 5 (especially in reference to Contemporaries) what
may benefit 20 themselves, but blemish 20 the present Author's 5 Repute: Therefore I shall be censur'd, I
question 5 not, by some, for a Solecism of the first 20 Magnitude, 15 that I (who am but a mean Star 15 in
comparison 5 of those other splendid Planitary Orbs in the Esculapian Firmament,) should presume to accost such a

Figure 13. Prototype of Document Viewer. Clickable tags are generated for word topic assignment, part-of-speech, or named entity assignment.

48 Our “Document Viewer” prototype (Figure 13) perhaps best represents the future direction of our project. We are designing this tool to be fully integrated into other models and visualizations on our site as a dynamic overlay to any text that is accessed from our database. Through an automatically-generated element tagging method, topic, part-of-speech, and named entity information will be available as one is examining the text itself. The Document Viewer figures this relational data for tokens in the corpus as latent dimensions within the texts themselves that can be studied and visualized. As with every other tool on our site, one can seamlessly navigate between this textual view and the models discussed above, which we believe mirrors the recursive nature of the archival research process.

49 As mentioned above, a project to produce a ground-up bidirectional encoding or contextual embedding model for the EEBO-TCP would enable a trove of new tools and methods. Such a project might help to integrate top-down topic modeling approaches with bottom-up embeddings, like in the Neural Topic Modeling approaches of Wang et al. and Gupta et al., or phrase- and document-focused comparison tools like phrase vectorization, *Gensim’s Doc2Vec*, or Kusner et al.’s *Word Mover’s Distance* [Wang et al 2018] [Gupta, Chaudhary, and Schütze 2021] [Kusner 2015]. These approaches could radically alter researcher interactions with texts if vector data were built into document indexing through a program like Apache Lucene. A researcher could plausibly search or access dynamic maps of phrases, documents, topics and more based on semantic similarity. A vector-based approach has also been fruitfully paired with image and sound tagging, and our analysis of the latter could be enhanced by reinserting these media into their textual contexts.

50 The normalized tagging necessary to produce such an embedding model could be used for network analysis of named entities, enabling search and categorization by metadata categories such as author, publisher, and location, and even diegetic textual categories like character, historical figure, location, or chronology. A synthesis between diachronic embeddings and diachronic topic modeling, such as Blei et al.’s Dynamic Topic Model (DTM) or Roberts et al.’s Structural Topic Model (STM) could help enrich our understanding of the complexity of discursive change beyond a simple word-based or topic-based approach. We have also experimented with the use of Gerrish et al.’s Document Influence Model (DIM) to map the persistence and change of topics over time linked to the publication of specific texts. This modeling technique would only be enhanced by integrating embeddings with topic modeling, and could assist with recovery projects and questions of canonicity and biases in scholarly attention.

51 While the above by no means exhausts the potential future directions of the *Quintessence* framework, we hope it paints a picture of the rich potential for future archival research that harnesses textual dimensionality. As we extend the *Quintessence* framework beyond the EEBO-TCP archive to inform database design on other archives and in other domains, we hope to reverse the trend toward massive and opaque automated textual modeling and encourage critical and scholarly apprehension of an ever-growing archive of digitized historical material.

Notes

[1] The first two authors contributed equally to this research. The authors would like to thank the UC Davis DataLab and UC Davis Library for supporting this project. In particular, we would like to acknowledge the invaluable feedback and assistance of Tyler Shoemaker, Kimmy

Hescock, Hugo Mailhot, Jane Carlen, Wesley Brooks, Anupam Basu, and Duncan Temple Lang. Finally, we would like to thank our anonymous reviewers for their useful recommendations and the DHQ editorial team for their work.

[2] Our code repositories can be found at the following links: For the web app, https://github.com/datalab-dev/quintessence_web_app; and for all NLP analysis, https://github.com/datalab-dev/quintessence_analysis.

[3] Three examples of this style of tool are: JSTOR Labs Text Analyzer beta https://www.jstor.org/analyze/, Gale's Digital Scholars Lab https://gale.com/intl/primary-sources/digital-scholar-lab, and Google Books N-Gram Viewer, https://books.google.com/ngrams/.

[4] Our proportion formula is:

$$\frac{\text{(document length)} \times \text{(document topicdistribution)}}{\text{sum of all document lengths}}$$

This formula for estimating topic proportion is used across the topic model visualizations, including when subsetting the archive and for the topic proportions over time plots, both of which are described below.

[5] Due to the collapse of the JS divergence into two dimensions, a certain level of imprecision is expected in the visual representation of the distance between topics. See [Lin 1991].

[6] For the metrics used, see [Arun et al 2010] [Cao et al 2009] [Deveaud, San Juan, and Bellot 2014] [Griffiths and Steyvers 2004].

[7] When dates were included as ranges, the earlier year was used. If a date was not included, or the uncertainty threshold was too great (e.g. the century was known but the year itself was not), the text was excluded from any temporal analysis, but remained in other models.

[8] Because LDA is a widely-used technique, it is also the most likely source of operability and efficiency gains in the topic modeling space. As an example, see the recent GPU implementation of LDA called CuLDA_CGS [Xie et al 2019].

[9] This methodology follows that of Hamilton, Leskovec and Juravsky [Hamilton et al 2016].

[10] In addition to plotly.js, we also used Vec2Graph in the design of the word embeddings visualizations. [Katricheva et al 2020].

Works Cited

Arun et al 2010 Arun, Rajkumar, Venkatasubramaniyan Suresh, C. E. Veni Madhavan, and Narasimha Murthy. (2010) "On finding the natural number of topics with latent dirichlet allocation: Some observations", in *Pacific-Asia conference on knowledge discovery and data mining*. Berlin, Heidelberg: Springer, pp. 391–402.

Bender et al 2021 Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. (2021) "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?🦜", in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623.

Birhane et al 2022 Birhane, Abeba, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. (2022) "The values encoded in machine learning research", in *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 173–184.

Blei and Lafferty 2007 Blei, David M., and John D. Lafferty. (2007) "A Correlated Topic Model of Science", *The Annals of Applied Statistics*, 1(1), pp. 17–35. Available at: https://doi.org/10.1214/07-AOAS114.

Blei, Ng, and Jordan 2003 Blei, David M., Andrew Y. Ng, and Michael I. Jordan. (2003) "Latent dirichlet allocation", *Journal of machine Learning research*, 3(Jan), pp. 993–1022.

Blei, Ng, and Jordan 2012 Blei, David M., Andrew Y. Ng, and Michael I. Jordan. (2012) "Topic Modeling and Digital Humanities", *Journal of Digital Humanities*, 2(1).

Bode 2018 Bode, Katherine. (2018) *A World of Fiction: Digital Collections and the Future of Literary History*. Ann Arbor: University of Michigan Press.

Burns 2013 Burns, Philip R. (2013) *Morphadorner v2: A Java Library for the Morphological Adornment of English Language Texts*. Evanston, IL: Northwestern University.

Cao et al 2009 Cao, Juan, Tian Xia, Jintao Li, Yongdong Zhang, and Sheng Tang. (2009) "A density-based method for adaptive LDA model selection", *Neurocomputing*, 72(7–9), pp. 1775–1781.

Chun 2008 Chun, Wendy Hui Kyong. (2021) Discriminating data: Correlation, neighborhoods, and the new politics of recognition. MIT Press.

Crowther et al 2008 Crowther, Stefania, et al. (2008) “New Scholarship, New Pedagogies: Views from the ‘EEBO Generation’”, *Early Modern Literary Studies*, 14(2), pp. 3–1.

Da 2019 Da, Nan Z. (2019) “The computational case against computational literary studies”, *Critical inquiry*, 45(3), pp. 601–639.

Deveaud, San Juan, and Bellot 2014 Deveaud, Romain, Eric San Juan, and Patrice Bellot. (2014) “Accurate and effective latent concept modeling for ad hoc information retrieval”, Document numérique, 17(1), pp. 61–84.

Froehlich, Whitt, and Hope 2012 Froehlich, Heather, Richard J. Whitt, and Jonathan Hope. (2012) *EEBO-TCP as a Tool for Integrating Teaching and Research*. Available at: https://ora.ox.ac.uk/objects/uuid:3a7c3b8f-cdac-4bdd-8b6c-33caa6f04179.

Gärdenfors 2014 Gärdenfors, Peter. (2014) *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. Massachusetts: MIT press.

Garg et al 2018 Garg, Nikhil, et al. (2018) “Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes”, *Proceedings of the National Academy of Sciences*, 115(16), pp. 3635–44. Available at: https://doi.org/10.1073/pnas.1720347115.

Gebru et al 2021 Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. (2021) “Datasheets for datasets”, *Communications of the ACM*, 64(12), pp. 86–92.

Gerow et al 2018 Gerow, Aaron, et al. (2018) “Measuring Discursive Influence across Scholarship”, *Proceedings of the National Academy of Sciences*, 115(13), pp. 3308–13. Available at: https://doi.org/DOI.org.

Gerrish and Blei 2009 Gerrish, Sean, and David Blei. (2009) *Modeling Influence in Text Corpora*. Available at: https://cdn.tc-library.org/Rhizr/Files/XtymaxJvoXXM89ffG/files/document_influence_model_1.pdf.

Griffiths and Steyvers 2004 Griffiths, Thomas L., and Mark Steyvers. (2004) “Finding scientific topics”, *Proceedings of the National academy of Sciences*, 101(suppl_1), pp. 5228–5235.

Gupta, Chaudhary, and Schütze 2021 Gupta, Pankaj, Yatin Chaudhary, and Hinrich Schütze. (2021) “Multi-source Neural Topic Modeling in Multi-view Embedding Spaces”.

Hamilton et al 2016 Hamilton, William L., et al. (2016) “Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change”, in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers. Association for Computational Linguistics, pp. 1489–501. Available at: https://doi.org/10.18653/v1/P16-1141.

Hellrich 2019 Hellrich, Johannes. (2019) *Word embeddings: reliability & semantic change*. IOS Press.

Hellrich et al 2018 Hellrich, Johannes, et al. (2018) “JeSemE: A Website for Exploring Diachronic Changes in Word Meaning and Emotion”. Available at: http://arxiv.org/abs/1807.04148.

Jo and Gebru 2020 Jo, Eun Seo, and Timnit Gebru. (2020) “Lessons from archives: Strategies for collecting sociocultural data in machine learning”, in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 306–316.

Katricheva et al 2020 Katricheva, N., Yaskevich, A., Lisitsina, A., Zhordaniya, T., Kutuzov, A., Kuzmenko, E. (2020) “Vec2graph: A Python Library for Visualizing Word Embeddings as Graphs”, *In*, 1086. Available at: https://doi.org/10.1007/978-3-030-39575-9_20.

Kim et al 2014 Kim, Yoon, et al. (2014) “Temporal Analysis of Language through Neural Language Models”, in Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science. Association for Computational Linguistics, pp. 61–65. Available at: https://doi.org/10.3115/v1/W14-2517.

Kulkarni et al 2015 Kulkarni, Vivek, et al. (2015) “Statistically Significant Detection of Linguistic Change”, in *Proceedings of the 24th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee*, pp. 625–35. Available at: https://doi.org/10.1145/2736277.2741627.

Kusner 2015 Kusner, M. J., Sun, Y., Kolkin, N. I., and Weinberger, K. Q. (2015) “From Word Embeddings to Document Distances”, in *Proceedings of the 32nd International Conference on Machine Learning, ICML*, pp. 957-966.

Lesser 2019 Lesser, Zachary. (2019) “Xeroxing the Renaissance: The Material Text of Early Modern Studies”, *Shakespeare Quarterly*, 70(1), pp. 3–31. Available at: https://doi.org/10.1093/sq/quz001.

Lin 1991 Lin, Jianhua. (1991) “Divergence Measures Based on the Shannon Entropy”, *IEEE Transactions on Information Theory*, 37(1), pp. 145–151.

McCollough and Lesser 2019 McCollough, Aaron, and Zachary Lesser. (2019) *Xeroxing the Renaissance: The Material Text of Early Modern Studies*. Folger Shakespeare Library.

Meeks and Weingart 2012 Meeks, Elijah, and Scott B. Weingart. (2012) “The digital humanities contribution to topic modeling”, *Journal of Digital Humanities*, 2(1), pp. 1–6.

Mikolov et al 2013 Mikolov, Tomas, et al. (2013) “Efficient Estimation of Word Representations in Vector Space”. Available at: http://arxiv.org/abs/1301.3781.

Mueller 2012 Mueller, Martin. (2012) *Towards a Book of English: A Linguistically Annotated Corpus of the EEBO-TCP Texts*.

Nikita 2016 Nikita, Murzintcev. (2016) “Ldatuning: Tuning of the Latent Dirichlet Allocation Models Parameters”. R Package Version 0.2-0, URL https://CRAN, R-project. org/package= ldatuning.

Pechenick et al 2015 Pechenick, Eitan Adam, Christopher M. Danforth, and Peter Sheridan Dodds. (2015) “Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution”, *PloS one*, 10(10).

Roberts, Stewart, and Tingley 2019 Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. (2019) “Stm: An R package for structural topic models”, *Journal of Statistical Software*, 91, pp. 1–40.

Schmidt et al 2021 Schmidt et al. (2021) “Uncontrolled Corpus Composition Drives an Apparent Surge in Cognitive Distortions”, *PNAS*, 118(45), pp. 1–2.

Shoemaker 2020 Shoemaker, Tyler. (2020) *Literalism: Reading Machines Reading*. UC Santa Barbara. (Dissertation).

Sievert and Shirley 2014 Sievert, Carson, and Kenneth Shirley. (2014) “LDAvis: A Method for Visualizing and Interpreting Topics”, in Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces. Association for Computational Linguistics, pp. 63–70. Available at: https://doi.org/10.3115/v1/W14-3110.

Staley 2015 Staley, David J. (2015) *Computers, visualization, and history: How new technology will transform our understanding of the past*. Routledge.

Wang et al 2018 Wang, Wenlin, et al. (2018) “Topic Compositional Neural Language Model”, in *International Conference on Artificial Intelligence and Statistics*. PMLR.

Welzenbach 2012 Welzenbach, Rebecca. (2012) “Transcribed by Hand, Owned by Libraries, Made for Everyone: EEBO-TCP in 2012”.

Whitmore 2010 Whitmore, Michael. (2010) “Text: A Massively Addressable Object”. Available at: http://winedarksea.org/?p=926.

Xie et al 2019 Xie, Xiaolong, Yun Liang, Xiuhong Li, and Wei Tan. (2019) “CuLDA: solving large-scale LDA Problems on GPUs”, in *Proceedings of the 28th International Symposium on High-Performance Parallel and Distributed Computing*, pp. 195–205.