DHQ: Digital Humanities Quarterly
2020
Volume 14 Number 3

The Role of Critical Thinking in Humanities Infrastructure: The Pipeline Concept with a Study of HaToRI (Hansard Topic Relevance Identifier)

Ashley S. Lee <ashley_lee_at_brown_dot_edu>, Brown University
Poom Chiarawongse <t_dot_chia_at_brown_dot_edu>, Brown University
Jo Guldi <jguldi_at_mail_dot_smu_dot_edu>, Southern Methodist University
Andras Zsom <andras_zsom_at_brown_dot_edu>, Brown University

Abstract

This article proposes the concept of the pipeline as a category of tool that organizes a series of algorithms for users. The pipeline concept, adopted with limitations by the humanities, documents how a suite of algorithms produces a particular research result, with the goal of enabling interoperability, transparency, and iteration by future scholars who may switch out particular algorithms within the pipeline with different results. A pipeline-based application amplifies the concepts of interoperability and transparency for users by allowing the researcher to toggle on and off particular options, for example selecting and deselecting particular topics of interest from a program of visualizations based on a topic model of a large body of text. Pipelines support modular, interoperable, transparent, and documented processes of research that lend themselves to Prof. Guldi's Theory of Critical Search — the argument that critical thinking increasingly takes place at the design and research stage of digital processes. The article presents the case of how the pipeline concept influenced the development of *HaToRI* (Hansard Topic Relevance Identifier), an open-source pipeline-based tool for identifying a cohort of thematically-linked passages in the nineteenth-century debates of Britain's parliament. In our pipeline, a series of algorithms move through the steps of cleaning a corpus, organizing them into topics, and selecting particular topics that are used to extract a sub-corpus that matches the user’s interests. Users have the option of searching based on multiple topics rather than merely keywords or a single topic at a time, allowing iterative searches to build upon each other. As an example of the Critical Search process in action, we follow an inquiry based on matching parliamentary reports with material from the Hansard British Parliamentary Debates. Using the pipeline, the user is able to identify multiple common topics of interest, and from these topics, extract a sub-corpus specific to land use and rent in the 19th century British Empire.

Introduction

In the era of web-enabled research, an enormous scholarly infrastructure of apps, software, and research portals structures how many scholars go about their research. This scholarly infrastructure comprises a significant portion of the “sites” (as this issue of *Digital Humanities Quarterly* put it) that structure the modern meetings of minds. The design choices of that infrastructure have direct consequences on research. 1

The developer’s choices, vision, and values shape the experience of knowledge, the movement of ideas, and the room for critical reflection on the interpretations offered by any scholar. Choices made in developing infrastructure govern how easy it is for students to adapt or build on the work of their professors, for scholars to make new discoveries about the same corpus, or members of the public to make new discoveries in an entirely new corpus. 2

A world of interactive applications facilitating distant reading of texts already exists and many of these applications are organized in a way that allows a researcher to toggle endpoint options on and off, but choices upstream in the application development process are less available. Whether a piece of humanities infrastructure is pipeline-like plays a role in how far the results of one study can be extrapolated to another arena, for instance, how much effort is required to 3 switch the British parliamentary debates out with the Swedish parliamentary debates, given the same infrastructure. This is not to say that pipelines are a panacea to interoperability challenges in digital humanities — many algorithms are specific to language or corpora. While pipelines allow scholars to apply methods to new data sets with little coding effort, the interpretation of results is a human task that still requires rigor.

This article imports a concept (“the pipeline”) that has a specific meaning and use in the field of computer science. The pipeline becomes, in our argument, a metric for the kinds of online sites that digital humanists have built, which allows us to ask important questions about how our infrastructure facilitates transparency, exchange, and critical inspection of the results of research. Modularity, interoperability, and transparency are values in pipeline development and guide how we develop applications. 4

We argue that the pinnacle of pipeline-building is when insights developed in one sphere can percolate to other spheres of knowledge with minimum translation — and that how we develop and code our scholarly infrastructure directly determines the options for collaboration, extrapolation, and interdisciplinary insight. We then present a case study of a web application that exemplifies some of the values inherent to a pipeline-based architecture. In a completely modular way, the application, called “Hansard Topic Relevance Identifier (HaToRI)”, ingests the Hansard data set and presents the results of topic modeling and downstream analysis. 5

Hansard is an archived and digitized resource of British Parliamentary speech going back to 1803. It contains full transcriptions of all spoken words as well as speaker and debate metadata. It is made up of hundreds of millions of words and the digitization has very few errors. As such, it is a well-studied and valuable resource for historical and socio-political research. 6

Since its inception in 2003, Latent Dirichlet Allocation (LDA) has become a popular algorithm used to analyze large text corpora in Digital Humanities and beyond [Blei and Lafferty 2009]. LDA is a particular flavor of topic models, which more broadly are probabilistic models for discovering the abstract concepts that occur in a large body of documents. The text is represented hierarchically, where a corpus is a collection of documents and in turn, a document is a collection of words. Topics are distributions over a fixed vocabulary of words and documents are made up of a mixture of topics. For example, an article in a newspaper could be about biology, neuroscience, computational techniques, and biochemistry and another article could be about politics, corruption, and finance. There might be other articles about the biological topic and more articles still about the political topic. The goal of topic modeling is to figure out what a large corpus is about and the results can be used for other text analysis tasks, like honing in on just the biological articles or tracing the prevalence of the political topic over time (given the corpus contains dates). 7

The Pipeline Concept

The concept of the “pipeline” is already prevalent in the literature of computer science, where pipelines organize suites of algorithms [Butterfield et al. 2016]. In code, a pipeline describes a series of algorithms that handle repeated processes — for example, data cleaning, stemming, topic modeling — in a series, handling completion of the previous task before moving on to the subsequent task. The omnipresence of pipelines in coding today is a relatively recent feature of the conditions of the available computational power over the last two decades, when pipelines, originally having been developed for use in supercomputers, came to be a feature of all high-power processes. 8

The use of the term “pipeline” implied a series of steps, one leading on from the next, describing a purpose-built architecture for complex tasks, originally imagined after Henry Ford's assembly line [Tucker 2004]. For our purposes, the assembly line metaphor brings along with it the imagery of “interchangeable parts”: that is, the ideal pipeline describes a process, and a product, where most of the parts can be swapped in and out, allowing for rapid fixes, upgrading, and comparison. When a pipeline is modular and open-source, if users disagree with our particular choices of algorithms, they can try other ones and assess how the change in algorithm affects the analysis. 9

These qualities mark out a set of modern values appropriate to scholarly practice in a digital age, and they characterize coding practice around pipelines as a whole. The pipeline makes the process of going from raw text to topic modeling results modular, subject to dissent, and interchangeable with different corpora. It is worth looking at the three values in 10 greater depth.

- By modular, we mean that an open-source pipeline insures that scholars can, at all steps in the pipeline not just the endpoint, swap out choices or settings for other ones. Modularity enables scholars to explore various results and arrive at richer, more rigorous results. It also makes it possible to do this at little cost to the researcher, in terms of time and resources needed to make and test the changes.
- By subject to dissent, we mean that scholars may take issue with any individual choice made in the construction of the pipeline. In order to intellectually engage the implications of their disagreement on the research process, scholars must be able to recreate alternative results that would have happened had one of the algorithms in the pipeline been switched out — for instance, had a different choice of stemmer or topic modeling algorithm been employed. In this process of openness to dissent, the pipeline becomes a tool for exploring best practices in digital research, where scholarly disagreements about tools and interpretation can be pinpointed for discussion and turned into instructive examples.
- By interchangeable, we mean that once an open-source pipeline is deemed useful for humanities and social science knowledge, other scholars may want to apply the insights from a research process to other bodies of text. Where no scholarly process exists for packaging tools into a pipeline, other scholars inspired by a research process must recreate the process from scratch. Where research is packaged as a pipeline, however, switching a few lines of code makes it possible for the scholar to apply one pipeline from the British Hansard debates to the Canadian or Australian Parliamentary debates, or any other text corpora. Interchangeability is a scholarly value because it enables the extension of insights from one domain of humanities or social science research into another domain.

Previous Waves of Critical Thinking about Humanities Infrastructure

A previous wave of efforts to bring structure to digital humanities research has foregrounded humanities infrastructure as a site of making. This literature has dealt with both the physical and organizational infrastructure of research [Svensson 2011], and data infrastructures [Edmond et al. 2015] [Brooke et al. 2015]. Brooke's GutenTag, for example, is a software that samples sub-corpora from the Gutenberg corpus based on publication dates, gender of the author, etc., and adds user-selected tags to the selection. Some, but not all of this research has suggested a formal set of values for analysis. Mattern argues that whether infrastructure is hard (roads, railways, bridges, data centers, fiber-optic cables, e-waste handlers, etc.) or soft (measurement standards, technical protocols, naming conventions, etc.), digital humanists among other interfacers with infrastructure should have some literacy around them, including the values that underlie them including transparency and modularity. Matterns also argues that infrastructure is not a revolutionary concept (though it has evolved since its inception in the 1920's), but allows people to "organize into communities and share resources amongst themselves." It is this self-organizing quality that necessitates better interfaces that reflect how communities want to and should interact with structures [Mattern 2016]. 11

Discussions on the biases in the process of big data research can be found in the literature of the past decade. A primary example is the idea that racist data leads to racist algorithms. Noble challenges the idea that scientific research is impartial, that it levels the playing field for "all forms of ideas, identities, and activities" [Noble 2018]. Negative biases are embedded in search engine results for "black girls," where suggested searches are radically different from those for "white girls." This results in biased search algorithms that privilege whiteness and oppress people of color. There is an increasing awareness of this issue - O'Neil et al. lay out situations where algorithms have the potential to amplify the biases and exacerbate the disparities present in society. Arenas such as going to college, online advertising, justice and law enforcement, getting a job and job performance, getting credit, and getting insurance can all be dangerously affected by algorithms' increasing role in the decision making process [d'Alessandro et al. 2016]. O'Neil calls for auditing in all steps of the data science development process [O'Neill 2016]. Additional research in the field of bias, discrimination, and oppression in algorithms and what to do about them are abundant [Eubanks 2015] [Introna and Nissenbaum 2006] [Nissenbaum 2010] [Vaidhyanathan 2012] [Vaidhyanathan 2018]. Kaplan, in his map of big data research in digital humanities, briefly raises the question of bias and notes that choices need to be made in the process of digitally translating from primary sources into high-level human-processible insights and that the inevitable biases that result from these choices apply [Kaplan 2015]. We support a solution, not by codifying any particular set of standards, 12 but by allowing these choices to be questioned in all steps in the pipeline, and changed directly by the questioner.

As a result, the pipeline values of modularity, subjectness to dissent, and interchangeability is apparent in many of the web portals available today — but not in all. Perhaps the work that most closely mirrors our viewpoint is Edmond's Data Soup, which describes an architecture and pipeline for text search and selection used in the [Edmond 2013] project with an emphasis on modularity of the processing algorithms and reproducibility. These works, however, deal mainly with the process of data preparation and selection, rather than the whole process of research, all the while treating it primarily as something to be done, rather than something to think about, discuss, and debate. Often times, in Digital Humanities, the description of the process is sidelined to the appendix or not discussed at all, which precludes any possibility of critical review of the pipeline. In fact, in a survey of close and distant reading techniques, [Janicke et al. 2015] found that many papers using sophisticated data analysis techniques do not even provide sufficient information about the preprocessing steps to be reproducible, let alone facilitate discussions on these processes. We argue that it is not enough for the data analysis tool to be sophisticated and user-friendly, but that all steps in the data analysis pipeline be truly transparent and modular. Mattern argues for a collective consciousness of "citizens/users" of infrastructure around both awareness of and critical listening and thinking around infrastructure, implicitly making an argument for these values [Mattern 2013]. 13

Mattern references "path dependency," a concept coined by Edwards et al., where past decisions limit future choices. While this is true of software, open source software gets around this dependency because the code is shared freely, with any infinite number of branches or forks possible from the original base code. If a scholar or user dislikes a past decision, they can change it so that future choices are not limited by a previous developer's choices. Thus, the software not only is transparent and modular and subject to dissent, but allows scholars and users to go beyond dissent and towards action and agency to diverge from the path. 14

It is advantageous to the community of researchers to discuss ideas about infrastructural choices built into a particular pipeline. If we do collectively adopt the convention of publishing articles wherein we describe the pipelines we have designed — and perhaps other articles wherein we critique recent conventions in building infrastructure — then the community of knowledge will benefit by collectively learning. The author can describe the work that went into designing a pipeline that maximizes critical thinking. Places in the research project where the best algorithm is uncertain — the places that lend themselves to critical review — can be described and highlighted, and the community of researchers can grow in its awareness of these uncertainties and the biases they create on the research project. Herein lies a major opportunity for collective critique and argumentation around the ideas that structure our infrastructure. 15

The Pipeline-Based Web App

In most web applications in the digital humanities, some pipeline exists in the background, but some of its functions are obscured from the user. The visualizations displayed are based on the results of cleaning and topic modeling run behind the scenes and then uploaded to the web app for the researcher to interact with. Simple navigation by date or keyword allows the user to expand or explore particular aspects of the results. Toggle options exist, but they are limited to later parts in the pipeline. For example, in the VoyantTools web application, we can upload any text we want and toggle visualization settings, but there are no options for how the text is cleaned and tokenized and furthermore, the way these normalization steps are done is obscured from the user [Sinclair and Rockwell 2016]. InPhO Topic Viewer, one of the HathiTrust Research Center algorithms, is a more modular pipeline than VoyantTools because it provides some options in the text pre-processing and normalization parts of the pipeline. However, the only two options are the choice of tokenizer (of which there are only two useful selections for English texts) and whether or not to decode unicode characters [Murdock et al. 2017]. 16

In web applications inspired by the pipeline concept, the designers seek to highlight and expand particular choices made in all steps of the design of the pipeline, giving the user options for how the data is handled and presented. In an ideal pipeline-based app, the user would have control over every process in the pipeline, from choosing different stemming and cleaning options to visualizations, allowing the community to upload different possible algorithms. 17

By giving users the option of reviewing the choices behind any given analysis and visualization, the design principle of the pipeline-based app reinforces the possibility of scholarly dissent in analyzing and interpreting documents. Two 18 scholars may have different instincts when it comes to how to clean the data, which topics are relevant to a query, or how to visualize the data, and these different instincts may lead them in different ways. To encourage a healthy climate of dissent and investigation, practitioners in the digital humanities need to incorporate design opportunities for dissent and counter-inspection of the evidence into the full pipeline of their software.

In a sense, then, valuing scholarly dissent means that interface designers must aspire to web-apps that perform like *code* in terms of their flexibility and interoperability, making every choice of algorithm transparent and interchangeable. The ideal pipeline-based application would, in a sense, lead the user through the experience of coding, where the coder typically chooses various packages that are assembled into a process of cleaning, analyzing, and visualization. One example of a pipeline-based tool is Jean-Philippe Cointet's Cortext,[1] a platform that allows users to upload textual documents and queue them for cleaning and various transformations including named entity recognition, topic modeling, and vector analysis [Rule et al. 2015]. 19

Pure pipeline-based tools have both advantages and disadvantages. The advantage, like the advantage of the pipeline itself, is the proliferation of the values of modularity, openness to dissent, and interoperability. Pipeline-based tools like Cortext recreate the experience of interactive, iterative coding with data for non-coders and mixed classrooms or readerships where not all who want to experience the data come from an equal background with respect to code. 20

The major disadvantage of existing pipeline-based tools are, generally speaking, features of first-wave design. For instance, some web interfaces designed by computer specialists often lack an eye for design principles that would make their use transparent to users, which results in interfaces that are cumbersome to use, for instance in the case of Zotero. Zotero is an application for curating, storing, and organizing journal articles, but it is possible to build text mining and visualization on top of it. With Zotero, a user can add or remove journal articles from their hand-curated collection, topic model their corpus, and view some results. Changing the inputs is trivial, but changing the parameters in modeling and visualization are not. In Zotero, the emphasis is on interoperability, not user-friendliness, with the result that some users may be put off an artificially abstract interface. Other web interfaces command limited access to computational resources, which results in the case of Cortext in long waiting times while users queue for a server. These are issues that could be remedied by later design and improvement. However, for designers to rebuild existing infrastructure to align with new principles, they would need access to grants that emphasize extending access to computational resources and usability. 21

Tool-builders who aspire to pipeline-level transparency and easy user interface may nonetheless adopt the principles of openness to dissent and interchangeability in certain *parts* of the research process, with the result of creating web interfaces that are more flexible, generative, and useful for scholarly debate than the first generation of web applications in the digital humanities. This article presents one such iteration, HaToRI, which aspires to the pipeline concept. It realizes this promise by transparency — that is, by documenting the background pipeline of its code for users — and by limited modularity, that is, allowing the user to choose different seed topics and tweak the z-value of our sorting algorithm. Transparency and open-source software give users options: if they want to change other parts of the pipeline, like for example lemmatizing rather than stemming, the code should be run again but with that parameter changed. 22

Introducing HaToRI, a Pipeline for modeling the British Parliamentary Debates

HaToRI is the Hansard instance of the "ToRI" (Topic Relevance Identifier) tool. Its creation was motivated by the research questions asked of the Hansard British Parliamentary Debates data set by our scholarly collaborator, but it was built with a generic text analysis pipeline in mind. ToRI exemplifies the values of the pipeline concept in various ways. At the highest level, it is applicable to any text corpus with trivial customizations. For example, if a user likes all of the visualizations and features available in the HaToRI app but wants to create an instance of ToRI analyzing Twitter data, we can spin up a "TweetToRI" web app with slight modifications to the underlying ToRI code pipeline. 23

The code pipeline consists of three steps: we prepare and identify the topics of a corpus, a humanist selects topic(s) of interest, and we post-process the results. Post-processing consists of ranking the documents based on how prevalent 24 the topic(s) of interest are, and using such a ranking for in-depth studies. We prepare a set of visualizations to illustrate how the topics cluster based on similarity, how their prevalence changed over time, etc. The choices that we made in each of the three steps were informed by the historical research aims and by the content and format of the Hansard data, but may not be the right choices for all data. They can be adjusted with ease for other data or research aims. For this reason, we believe that the pipeline driven web tool can greatly aid humanists in their research.

25 The ToRI web app takes inspiration from many other useful Natural Language Processing (NLP) tools, most notably the Topic Explorer [Goldstone and Underwood 2014], but is distinguishable in its highly customizable end-to-end pipeline and novel document ranking algorithm. Many tools, like VoyantTools and OldBaileyVoices.org, are limited to basic keyword-based text analysis and corpus exploration [Sinclair and Rockwell 2016]. And while VoyantTools can accommodate various corpora, it does not accept a tab-separated value format that is ideal for corpora like Hansard that contain structured fields. Other tools explore and extract semantic sub-corpora by ranking documents [Tangherlini and Leonard 2013] [Goldstone and Underwood 2014], but use similarity to a single document or topic as the sorting value. Lengthy documents, in particular, can contain too much noise and a single topic does not capture the nuance that multiple topics can, resulting in a sub-corpus that potentially contains many false positives and leaves out harder to detect true positives. Meandre is a workflow tool that is modular and customizable to many corpora and offers a drag-and-drop interface for non-technical users to perform many NLP tasks including data cleaning, topic modeling, sentiment analysis, etc. [Llora et al. 2008]. While it is user-friendly and feature-rich, it does not offer sub-corpus extraction using multiple topics.

26 The advantages of our pipeline over other corpus exploration tools are that they are self-contained and complete, modular, and open source. These qualities make the code generalizable to any corpus, in any format and highly adaptable to incorporate other algorithms or NLP decisions. Specifically, if a user does not agree with our choice of the stemming method (Snowball Stemmer), they can add a different algorithm (e.g., the Porter Stemmer) by changing two lines of code and then assess how the change influences the results.

27 Creating instances of the web app with other corpora requires the addition of a single, custom pre-processing script that converts the corpus to our generic data format. The generic format is a tab-separated file with three fields: *Document ID*, *Text*, and *Year*. The *Document ID* can be any identifier such as a sequence of integers or document titles if available and it needs to uniquely identify the documents. The *Text* field contains the full, natural text of the documents, prior to cleaning and stemming. Then, the rest of the pipeline can run without any modifications. The pipeline performs the complete set of tasks needed to transform natural language to machine-readable vectors, a noisy collection of words to a hierarchically sorted set of topics, and an impossibly large volume of text to a manageable sub-corpus of semantically relevant documents.

28 Additionally, the code is modular so the work required to make changes to the pipeline is trivial. The decisions we made at all points in the pipeline — such as our custom stopword list, our spell-checking dictionary, our decision to stem rather than lemmatize, topic modeling method and implementation, number of topics, document ranking algorithm, and z-value — can be customized to different corpora or user interests. The code is designed this way to make space for flexibility in decision-making; algorithms best suited for one corpus or domain may not be optimal for another corpus or domain. Furthermore, some of these decisions are built into the endpoint of the pipeline, the web app user interface, with no code changes required. For example, the z-value is a tune-able parameter that the user interacts with in a value entry bar on the detailed topic view pages of the web app. Future work will incorporate more user input at the endpoint. [2]

Comparing Applications: How Pipelines Broaden the Scholarly Implications of Research

29 The Hansard parliamentary debates are important information mines for scholars of British history and attempts to digitize Hansard date back to at least the 1970s. Open source versions of the nineteenth-century Hansard corpus have been freely available in a digitized version for at least a decade, which has made them the subject of a great deal of infrastructure design already. At least two applications allow users access to the debates today — the Hansard Corpus site, designed by a team of linguists at Glasgow and Brigham Young, and the Millbank Hansard site, commissioned by

parliament itself of a private developer. The two existing applications have already enabled a flood of new research about language use in parliament [Blaxill 2013] [Alexander and Struan 2017].

However, the first generation of applications made little use of the pipeline concept. In many cases, little documentation was available about how the corpus had been cleaned. Users were allowed to search by keyword, year, and (in some cases) speaker or other named entity, but few tools offered other ways of navigating the corpus or matching a user's interests. Hansard, thus, offers an ideal test-case for building a web application that models the advantages of modularity, openness to dissent, and interoperability, due, in part, to the heavy scholarly interest in the subject matter; the material covered by Hansard is the subject of interest in history, literature, linguistics, and geography. Moreover, an interoperable pipeline or pipeline-based app for working with the British Hansards could be easily extended to the Australian, Irish, and Canadian Hansards, the debates of the EU parliament, or any other digital corpus that represents the debates of a democratic body. 30

Interoperability: How to Pipe HaToRI to Other Corpora

Unlike applications that readily ingest text data in the web browser without ever having to touch the underlying code, HaToRI requires some programming knowledge to adapt it to other corpora. The advantage of this approach is that the corpus can have associated structure and metadata beyond a single document of plain text. Additionally, minor code changes can be made in the process of spinning up another instance of ToRI, making the intermediary steps between data ingest and data visualization completely modular. To spin up an instance of ToRI for another corpus, clone the Inquiry for Philologic Analysis repository from Github and make modifications to the code in the *src* folder. 31

The pipeline is set up in `main.py`. Each step in the pipeline can be toggled on or off by setting the switches at the top of the script. For example, to run the first step of the pipeline on the test data given, the user can toggle the `to_tsv` switch to "True", and the sample 10 Hansard data will be converted to a tabular, tab-separated value (TSV) file. To ingest another corpus, the user must first convert the raw text to a TSV file. The TSV must represent a corpus as one row per document, and each document must have a unique identifier associated with it. The identifier is the first column in the TSV file and the text of the documents should be the second column. Metadata is optional - for HaToRI the speaker name and debate year were added as metadata by appending them as columns three and four. This step is unique to each data set - for Hansard we wrote a code (`raw_corpus2tsv.py`) to parse the data we needed from XML to TSV. For other corpora, the code will need to be custom written for the format it originates in. 32

The next step, data cleaning, is a single script (`preprocess.py`) that can be run as is, changed with minor code edits to accommodate different scholarly choices or domain-specific corpora, or replaced by a custom preprocessing script entirely. Tokenization is by word, but can be changed to ngrams (n-length word phrases). Non-alphanumeric characters are removed, but perhaps punctuation and special characters are important, such as with a Twitter data set. All text is lower-cased, though with texts with many abbreviations like medical notes, capitalization could be preserved. Spell-checking can be turned from on to off, and word truncating can be changed from stemming to lemmatization, or turned off completely. Common as well as custom stopwords are removed, and either or both stopwords lists can be turned off or modified. Once cleaning choices have been made and run, the user must download the MALLET program and run topic modeling with the desired numbers of topics, k. The output of the topic models are then ready to be loaded into the website. 33

There are two steps to creating the website: uploading the corpus and putting the website online. The user should clone the *hatori* repository from Github and change out the data in the *serv/data/* folder with their own corpus. Then, once the website works locally, they should host the website using one of a variety of web hosting services. We suggest doing this through Github Pages[3] because of its integration with Github. 34

Methods: the Pipeline Outlined

The remainder of this paper describes a full NLP pipeline for sub-corpus extraction using topic-based search and tests the hypothesis that topics are better than keywords for discovering a thematically coherent sub-corpus that includes 35 under-studied documents. We will describe the pipeline’s machinery in detail, which includes custom data extraction and pre-processing, conversion of the text corpus to a numeric and machine-readable data structure, topic modeling the corpus, document ranking and sub-corpus extraction, and visualization in a web app. We give an example scholarly use case to show how features of the web app work, and how the pipeline enables reproducible and interoperable digital humanities research and openness to dissent as a way of collaboration.

In characterizing the pipeline, we give a transparent description of the entire research project from raw text file (in this case; in others, it might be scanned documents) to visualizations. In emulating the value of transparency, this pipeline description highlights the opportunities for critical thinking along the way and how scholars can engage them. Indeed, we believe that this kind of description, which both illustrates the facts about the corpus and highlights the places for interpretive work by scholar-users, should become the standard for scholarly publications that document new scholarly infrastructures in the humanities. 36

Pre-processing

Here, we describe the process by which we turned the raw Hansard parliamentary debates into machine-readable material for downstream steps in the pipeline. The process described contains choices specific to the Hansard dataset, but it is modular and easily interchanged with a different suite of pre-processing steps. 37

We download Hansard from the web as a series of XML files[4] and parse full text and metadata from the XML into a tabular, tab-separated value (TSV) format using the Python programming language. The input data could be any collection of documents (news articles, tweets, political speech, etc.) in any text format (cannot be images or scans of documents), so long as it is converted into a TSV structured such that each row is a document. In the Hansard TSV file, each row is a debate and the columns are *debate ID*, *full text*, and *metadata*. We additionally append four written reports crucial to the historical analysis to our corpus. The result is a corpus made up of 45,585 speakers who uttered 294,203,233 words, 1,033,536 speech acts, and 111,689 debates. 38

Figure 1 shows what parliamentary speech and the written reports look like over time. The number of debates per year correlates closely with the number of speakers per year, and they both increase over the course of the century with a notable spike in the last quarter century. The average length of our documents shows a weak inverse correlation. The number of speech acts and the number of unigrams per debate are both somewhat reduced in the last quarter of the century. In other words, as the number of debates and number of speakers increase, the speech acts tend to get less verbose over time. 39

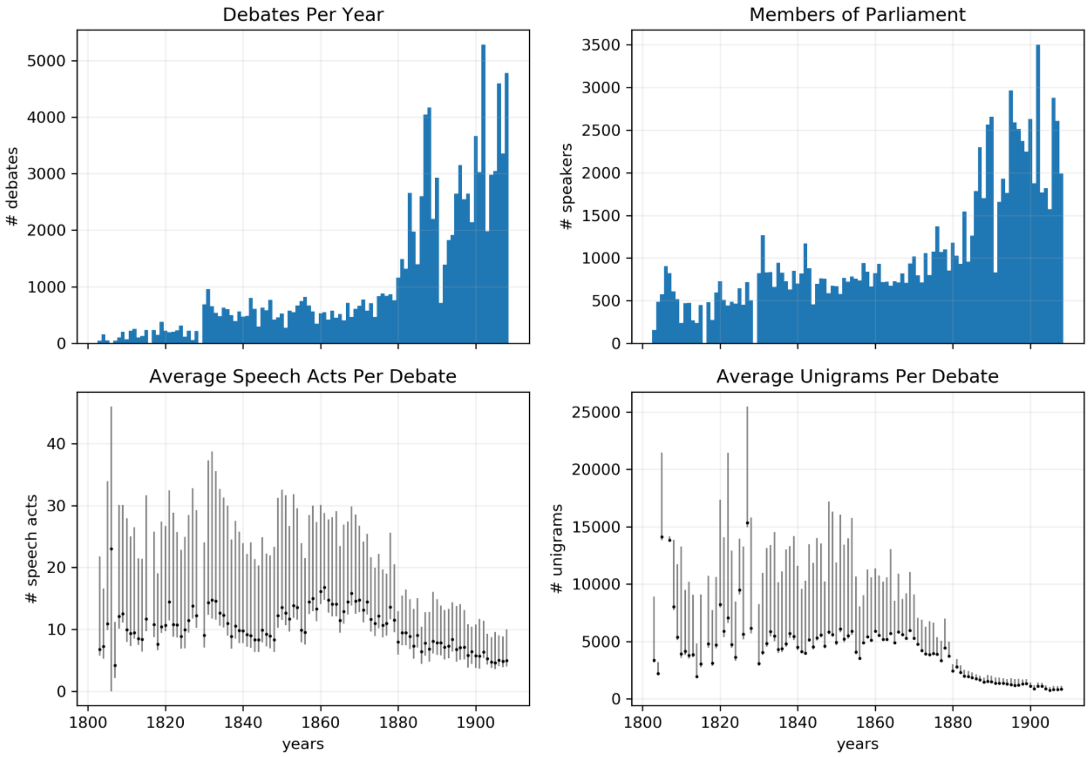


Figure 1. The number of debates and the number of speakers stayed relatively constant until the last quarter of the century when there was a sharp increase in both. While there are more debates and more speakers, the speakers were less verbose during the last quarter century.

As is standard in Natural Language Processing, we took steps to pre-process the corpus to prepare the text for downstream analysis. These steps include tokenization, removing non-alphanumeric characters, lowercasing, spell-checking, replacing both common stopwords and custom stopwords with a substitute word, and stemming. We walk through each step in detail during the rest of this subsection. 40

- We lowercase the corpus and then tokenize the documents by splitting the text on whitespace (e.g. spaces, tabs, returns). There are a variety of tokenizers to choose from (e.g. Penn Treebank, Punkt, Multi-Word Expression, Tweet, Regular Expressions, etc.) and we chose whitespace tokenization based on the structure of Hansard. This reduces our unit of text from paragraphs to words. Our corpus, Hansard, consists of 198,338 unique words.
- We remove non-alphanumeric characters from each word, so they are free from punctuation, white space, and symbols. This reduces the vocabulary to 186,060 words. Other text corpora may require the removal of different characters, like emoji from tweets or html tags from data scraped from the web. Changes to the removal list are trivial to implement in the code pipeline.
- We spellcheck the words using a dictionary of British English words provided by the Python library, *enchant*. If the words in our vocabulary are in the dictionary, we keep the word; if they are not, we replace them with a substitute word that does not appear natively in our corpus. This reduces the vocabulary to 49,789 words. We also replace common stopwords (e.g. articles and conjunctions like the, and, or, it, etc. that have no semantic meaning) and custom stopwords (proper nouns) with a substitute word. Replacing stopwords reduces our vocabulary by less than one thousand words. We can, with minimal effort, swap out the British English dictionary for other language dictionaries or change the stopwords used.
- We stem the words using Snowball Stemmer in the *NLTK* Python library. Stemming further decreases the size of the vocabulary by reducing inflected and derived words to their stem (base or root form) to avoid duplicate counting words that may have slightly different endings but the same semantic meaning. “Rent”, “rents”, “rented”, and “renting” would all be reduced to the stem, “rent”, and “property” and “properties” would both be reduced to the stem, “properti”. We call the stemmed words unigrams because stems like

> “properti” are not words. There are a number of different stemmers and lemmatizers to choose from, should the Snowball Stemmer be called into question.

By the end of the process, the original 294 million words are reduced to 97 million. Almost 200 million words are replaced by the substitute word because they are common stopwords (for example: “a”, “an”, “the”, “I”, “you”) or misspelled words. While the reduction seems significant, substitute words would contribute very little to the topic modeling results and thus minimal information is lost during the process. The final step in pre-processing is to create a document-term matrix, which is a numerical representation of the corpus that can be read and analyzed by a machine. Each document is represented by a vector of word counts. Different weights and normalizations can be applied to the document-term matrix, given the needs of the data or research. 41

Pre-processing Step	Number of unigrams	Number of non-substitute unigrams
All unigrams	198,338	196,642
Alphanumeric characters only	186,060	184,447
Correctly spelled unigrams	49,789	49,190
Stems	21,828	21,444

Table 1. We illustrate how the various pre-processing steps reduces the number of unique unigrams in Hansard by a factor of 10. We get the greatest reduction in unique unigrams by spell-checking and stemming words.

Topic Modeling

The goal of applying NLP methods to Hansard is to answer historical questions about the shifting concepts of land ownership and property in the nineteenth-century British empire. Traditional historical scholarship involving close reading and analysis is not suited to this task because the text corpus is too large to be read by a single historian or even a team of researchers. Instead, we offload the task to a computer. 42

Topic models are probabilistic models for discovering the abstract topics that occur in a large corpus of documents through a hierarchical analysis of the text [Blei and Lafferty 2009]. The idea is that a corpus is made up of topics, topics are distributions over a fixed vocabulary of terms, and documents are mixtures of topics in different proportions. The vocabulary of terms can be words or pre-processed tokens like unigrams or n-grams. Topic modeling can answer questions that would otherwise be intractable with a large corpus: (1.) What is a large, prohibitively long corpus of documents about? (2.) Can we extract a human-readable sub-corpus of semantically-related documents? A random process is assumed to have produced the observed data, which are the words that make up the documents in the corpus. Given the observed data, the posterior distribution of the hidden variables (word distributions for each topic, topic proportions for each document, topic assignment per word for each document) determines the hidden topical structure of the corpus which tells us what the corpus is about, answering the first question we seek to solve with topic modeling; the posterior estimates can be applied to tasks such as document browsing and information retrieval, answering our second question. In fact, topic modeling has been used as the basis for “trawling” for a smaller corpus [Tangherlini and Leonard 2013]. 43

The evaluation of topic modeling results is subjective, requiring manual inspection by subject matter experts and cannot replace the traditional textual analysis. The benefit, however, is that the subject matter experts need only to read through a handful of terms per topic rather than a significant part of the corpus. The drawback is that we lose vital information encoded in natural language when we numericize it. For example, the bag-of-words representation of a corpus loses context entirely by treating documents as unordered collections of words [Harris 1954]; n-grams cannot preserve word relationships past their immediate neighbors [Manning and Schutze 1999]; and while vector models like word2vec and doc2vec preserve far-apart word relationships, they are not suitable for identifying a relevant sub-corpus within a large corpus [Mikolov et al. 2013] [Le and Mikolov 2014]. 44

We use a particular flavor of topic modeling called Latent Dirichlet Allocation (LDA) [Blei et al. 2003]. The model does 45 not have a prior notion about the existence of the topics; it is given a hyperparameter, k, which describes how many topics are associated with the corpus, and discovers the k topics from the observed data, the words in the documents. The algorithm is implemented in a number of programming languages and is simple to use with few tunable parameters, but requires hand labeling of topics by subject matter experts. We use a Java implementation of LDA called MAchine Learning for LanguagE Toolkit (MALLET) [McCallum 2002]. In developing the pipeline, we made topic models with varying k values (50, 100, 200, 500, 1000) and different topic modeling methods like Non-negative Matrix Factorization [Berry and Browne 2005] and Dynamic Topic Models [Blei and Lafferty 2006] available to the users.

Extracting a Sub-corpus

Within *HaToRI*, the process of extracting a subcorpus starts with LDA. LDA not only tells us the topics that are discussed in a corpus, but also which ones each document contains. Knowing what documents are about allows us to narrow our focus to a smaller sub-corpus from the whole of Hansard, reading the debates and speech acts that are about our topic of interest — land and property. Sub-corpus extraction using topic modeling allows us to find a richer set of documents than keyword search alone because the topics are discovered from the data, rather than imposing domain-specific and subjective knowledge on the text. Keyword searches and term frequencies work well when the topic of interest can be unambiguously identified by one word or a short phrase. Documents in a corpus can be clustered based on similarity to one another and subject matter experts need to analyze each cluster to decide if the clusters make semantic sense and if any are about their topic of interest. This can only be done if the corpus is not too large and the experts can at the very least scan through each document. Sometimes it is already known that one or a few documents in a corpus are about the topic of interest and the question is whether there are other so far undiscovered relevant documents in the corpus? The goal then is to find documents which are similar to the seed documents. A search based on seed documents works if most words and terms in the seeds are about the topic of interest and not other general topics. This is not often the case. Our proposed method for sub-corpus extraction is a topic-model-based document ranking process where users choose multiple topics of interest, rank documents by how closely they match the topics of interest, explore the results, and iterate by toggling on or off particular topics. We harness a code pipeline to build a highly customizable web-app that makes available to the ordinary user a process of Critical Search characterized by iterative interaction with large corpora [Guldi 2018]. 46

We use three steps in topic modeling: import, modeling, and post-processing. The first two steps are fairly straightforward parts of the MALLET program. In modeling, we tune the number of topics, keeping other hyperparameters (e.g. alpha and beta (priors), number of iterations, sampling method, etc.) constant. We created topic models with the number of topics equal to 100, 200, 500, and 1000. In post-processing, we asked a group of humanists to read the ten most heavily weighted words in each topic, give the topic labels based on those words, and qualitatively evaluate which k number of topics is ideal for Hansard (see the Results section for more details). 47

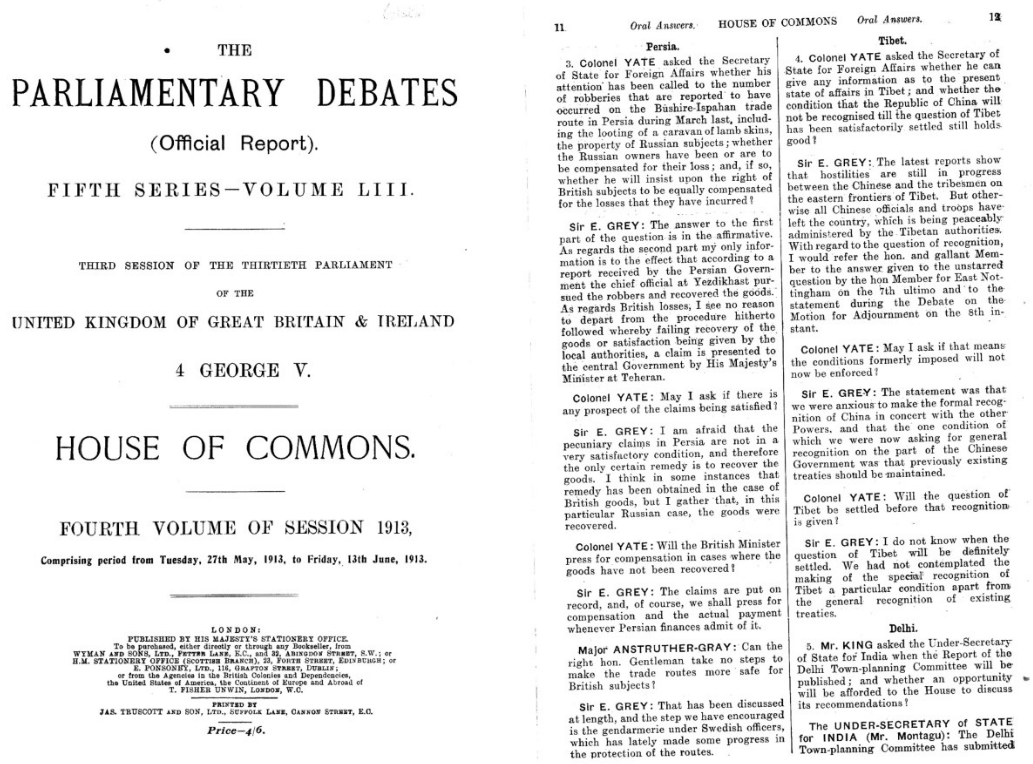
Sub-corpus extraction is done by ranking the documents in a corpus using their relevance to one or more topics of interest. The simplest ranking algorithm counts the relative frequency of topic words, or word occurrences assigned to at least one topic of interest, and ranks the documents with the highest frequencies as most relevant. However, this algorithm can incorrectly down-rank longer documents. We improve on this simple ranking algorithm by using the lower bound of Wilson score confidence intervals for a Bernoulli parameter, z, as the sorting value [Agresti and Coull 1998]. The parameter z is tuneable and a larger z sporadically introduces longer documents into the top-ranked sub-corpus; in our web app, the user can visually explore how increasing z changes the rankings by reading the document titles and linking to the full text of the documents. Different values of z may be optimal for sub-corpus extraction depending on the topics of interest and the user's goals. 48

We argue that topic-based sub-corpus extraction is a highly effective way to identify a relevant sub-corpus and highlight this point in the context of exploring past land use and eviction in the British Parliamentary debates. 49

Use Case: Searching for Property in Hansard

Parliamentary speech has been carefully archived since 1803 and is available as a digitized resource called the 50

Parliamentary Debates; it contains full transcriptions of all of the spoken words in the British Parliament from 1803 to 1908 as well as metadata like speaker name, speaker constituency, date of speech, and debate titles.[5] The collection is commonly referred to as Hansard after Thomas Curson Hansard, a London publisher and first official printer to the parliament. We use the terms Parliamentary Debates and Hansard interchangeably throughout this paper. Hansard provides a rich corpus for text analysis using Natural Language Processing methods because its digitization is high quality (low errors in optical character recognition) and it is comprehensive, spanning over two centuries and hundreds of millions of words of speech.



THE

PARLIAMENTARY DEBATES

(Official Report).

FIFTH SERIES—VOLUME LIII.

THIRD SESSION OF THE THIRTIETH PARLIAMENT
OF THE
UNITED KINGDOM OF GREAT BRITAIN & IRELAND

4 GEORGE V.

HOUSE OF COMMONS.

FOURTH VOLUME OF SESSION 1913,

Comprising period from Tuesday, 27th May, 1913, to Friday, 13th June, 1913.

LONDON:
PUBLISHED BY HIS MAJESTY'S STATIONERY OFFICE.
To be purchased, either directly or through any Bookseller, from
WYMAN AND SONS, LTD., FETTER LANE, E.C., and 32, ABINGDON STREET, S.W.; or
H.M. STATIONERY OFFICE (SCOTTISH BRANCH), 23, FORTH STREET, EDINBURGH; or
E. PONSONBY, LTD., 116, GRAFTON STREET, DUBLIN;
or from the Agencies in the British Colonies and Dependencies,
the United States of America, the Continent of Europe and Abroad of
T. FISHER UNWIN, LONDON, W.C.

PRINTED BY
JAS. TRUSCOTT AND SON, LTD., SUFFOLK LANE, CANNON STREET, E.C.

Price—4/6.

11 *Oral Answers.* HOUSE OF COMMONS *Oral Answers.* 12

Persia.

3. Colonel YATE asked the Secretary of State for Foreign Affairs whether his attention has been called to the number of robberies that are reported to have occurred on the Bûshire-Ispahan trade route in Persia during March last, including the looting of a caravan of lamb skins, the property of Russian subjects; whether the Russian owners have been or are to be compensated for their loss; and, if so, whether he will insist upon the right of British subjects to be equally compensated for the losses that they have incurred?

Sir E. GREY: The answer to the first part of the question is in the affirmative. As regards the second part my only information is to the effect that according to a report received by the Persian Government the chief official at Yezdikhast pursued the robbers and recovered the goods. As regards British losses, I see no reason to depart from the procedure hitherto followed whereby failing recovery of the goods or satisfaction being given by the local authorities, a claim is presented to the central Government by His Majesty's Minister at Teheran.

Colonel YATE: May I ask if there is any prospect of the claims being satisfied?

Sir E. GREY: I am afraid that the pecuniary claims in Persia are not in a very satisfactory condition, and therefore the only certain remedy is to recover the goods. I think in some instances that remedy has been obtained in the case of British goods, but I gather that, in this particular Russian case, the goods were recovered.

Colonel YATE: Will the British Minister press for compensation in cases where the goods have not been recovered?

Sir E. GREY: The claims are put on record, and, of course, we shall press for compensation and the actual payment whenever Persian finances admit of it.

Major ANSTRUTHER-GRAY: Can the right hon. Gentleman take no steps to make the trade routes more safe for British subjects?

Sir E. GREY: That has been discussed at length, and the step we have encouraged is the gendarmerie under Swedish officers, which has lately made some progress in the protection of the routes.

Tibet.

4. Colonel YATE asked the Secretary of State for Foreign Affairs whether he can give any information as to the present state of affairs in Tibet; and whether the condition that the Republic of China will not be recognised till the question of Tibet has been satisfactorily settled still holds good?

Sir E. GREY: The latest reports show that hostilities are still in progress between the Chinese and the tribesmen on the eastern frontiers of Tibet. But otherwise all Chinese officials and troops have left the country, which is being peaceably administered by the Tibetan authorities. With regard to the question of recognition, I would refer the hon. and gallant Member to the answer given to the unstarred question by the hon Member for East Nottingham on the 7th ultimo and to the statement during the Debate on the Motion for Adjournment on the 8th instant.

Colonel YATE: May I ask if that means the conditions formerly imposed will not now be enforced?

Sir E. GREY: The statement was that we were anxious to make the formal recognition of China in concert with the other Powers, and that the one condition of which we were now asking for general recognition on the part of the Chinese Government was that previously existing treaties should be maintained.

Colonel YATE: Will the question of Tibet be settled before that recognition is given?

Sir E. GREY: I do not know when the question of Tibet will be definitely settled. We had not contemplated the making of the special recognition of Tibet a particular condition apart from the general recognition of existing treaties.

Delhi.

5. Mr. KING asked the Under-Secretary of State for India when the Report of the Delhi Town-planning Committee will be published; and whether an opportunity will be afforded to the House to discuss its recommendations?

The UNDER-SECRETARY of STATE for INDIA (Mr. Montagu): The Delhi Town-planning Committee has submitted

Figure 2. Scanned PDF images of the Parliamentary Debates. The left image shows some of the metadata captured in the records, such as title, volume number, date, and house of parliament. The right image shows some speech acts, separated by debate topic and by speaker.

51 Because of the volume and breadth of topics in Hansard, it is a valuable resource for historical and socio-political analyses by students and scholars, alike. And while there is great value in being publicly available in its entirety, it is difficult to narrow down to smaller excerpts of interest in its existing online forms. We create a pipeline that ingests the Hansard data and presents it in a new digital space, the HaToRI web app, that broadens the horizons of possibility for digital humanities research and collaboration. We present a case study in using the web app for a historical analysis of how ideas about land and property changed in the nineteenth-century British empire in the following sections. However, the tool is modular and can be applied with some modification to any number of analyses by selecting a different focal point other than the land topics or swapping out the Hansard corpus for another collection of text.

52 While historical scholarship on Hansard is rich and varied due to the completeness of the resource, there is a lack of research about the changing discourse of property in Hansard. In contrast, there is extensive research on four written reports specific to land and property [Bull 1996] [Black 1960] [Campbell 2005] [Connelly 2003] [Donnelly 1983] [Grigor 2000] [Steele 1974]. They were commissioned by the Queen and are commonly aliased by the name of the lords who wrote the report: Napier, Bessborough, Richmond, and Devon. A major shift occurred between the Devon and Bessborough Reports (1845-1881), from greater landlord protections to greater tenant protections, due in part to massive resistance in colonies of the Empire like Ireland, Scotland, and India [Sartori 2014]. The Encumbered Estates Act of 1849 disadvantaged tenants trying to improve their holdings by moving property with outstanding debts from Irish

to English owners; Gladstone’s Irish Land Act of 1881 advantaged tenants by introducing the first rent control law in history, in addition to redistributing landlord property to Irish tenants [Readman 2008] [Steele 1974]. We might attribute the large amount of scholarship on the reports to both the specificity of the subject matter of the reports and also, to a citation feedback loop, where historians produce much research about widely cited sources because of their ubiquitousness in the literature. While studying these documents so extensively is valuable, it creates bias in the historical analysis by leaving out more uninvestigated primary sources.

Contemporary ideas about property that favor tenant protections, a normalized cultural expectation that is often codified into law, have a long and bloody history that can in part be traced back to the revolts and reforms that occurred in the nineteenth-century British empire. A major shift from favoring landlords to tenants in areas like land distribution, eviction, fair rent, is theorized to have occurred from the first half to the second half of the century [Sartori 2014]. The evidence presented in the following section supports this theory. 53

Results and Visualizations

The historical analysis that follows is one of many possible threads of questioning into the Hansard body of text. We demonstrate how to explore the question of how the parliamentary discourse of property changed in the nineteenth-century, using the choices made in developing the code pipeline and the user- and visualization-driven search through the web app parameters. The threads of questioning, driven by scholars and research interests, can be approached in a new way that leans on digital tools for a deeper dive into large bodies of text. Historical claims are still made through close reading and analysis; but using topic-based search to guide close reading and analysis, we open a new and focused window into a previously under-studied primary source. In the next section, we both validate our hypothesis from the secondary literature and introduce new perspectives — of members of parliament — into the body of knowledge about British land reform. 54

Five Out of Five Hundred Topics are Relevant for Land and Property

An early choice made in the pipeline is what the ideal number of topics (k) is for the Hansard corpus. We pre-computed topic models at different increments of k (50, 100, 200, 500, 1000) for review. Humanists determined that the optimal number of topics is 500 because of the uniqueness and specificity of the topics. Some topics are difficult to label distinctly from other topics in the 1000 topic model, while most topics could be given unique labels in the 500 topic model. Five distinct land-related topics, their ten most heavily weighted words, and their designated labels are shown in Table 2. They are sorted by the year(s) in which they peaked, with early peaks on the left and late peaks on the right. The language used to describe land in Hansard mirrors the historical shifts of the period, from an agriculturally valued plot in the first half of the century to a rental relationship between landlords and tenants in the middle of the century and to legal protections that favored tenants in the latter half of the century — protections that included rent restrictions, freedom from arbitrary evictions, and land redistributions [Readman 2008]. 55

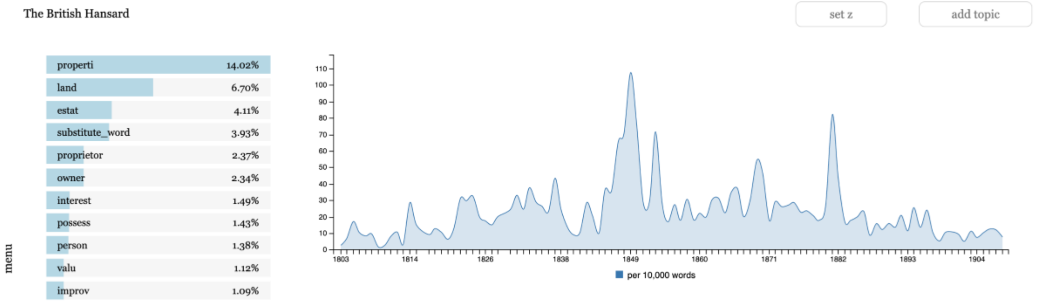
17. Land value measured in agricultural output	46. Land value measured as a tax for the church and as rental income for landlords	447. Property as a rental relationship between landlords and tenants	35. Negotiating the interests of landowners vs. tenants who seek to improve their rentals	180. Legal protections for tenants
1820, 1846	1827-1840	1838, 1892	1847-1850, 1882	1882-1904
farmer	tith	leas	properti	rent
agricultur	rent	year	land	land
land	clergi	rent	estat	case
price	charg	leasehold	proprietor	commiss
produc	land	lesse	owner	commission
farm	commut	land	interest	fix
cultiv	owner	term	possess	fair
year	collect	tenant	person	court
interest	landlord	properti	valu	chief
crop	payment	case	improv	applic

Table 2. Five land-related topics from our 500 topic model of Hansard. They are sorted on when they peaked. The first half of the century does not talk about land as much as a rental agreement but in terms of its productivity. The second half of the century considers the question of land as rental agreement and the legal protections and rights for both sides of the agreement.

We build this process of preparing, reviewing, and labeling different k values into the code pipeline. While the computational cost of preparing different models is not reduced, the human effort of maintaining code and documentation on many models is greatly reduced. 56

Spikes in the discussion of land use and rent in 1850 and 1880

Figure 3 shows topic 35, which we named “Negotiating the interests of landowners vs. tenants who seek to improve their rentals”. This topic peaks twice in the late 1840s and early 1850s and then peaks again in 1870 and 1880, mirroring the historical arc of this topic. The Devon and Bessborough Commissions were ordered and written in 1845 and 1881 because of calls for land reform from peasant farmers and tenants [Sartori 2014]. This topic is one of several that point the scholar to the land question, and in combination they can paint a fuller, richer picture of the topic than any single excerpt of Hansard or written report. 57



properti	14.02%
land	6.70%
estat	4.11%
substitute_word	3.93%
proprietor	2.37%
owner	2.34%
interest	1.49%
possess	1.43%
person	1.38%
valu	1.12%
improv	1.09%

Figure 3. The most frequently occurring words in the topic are ranked and displayed on the left. The percentages are each word’s relative frequency in the topic and add up to 100%. The line plot shows the frequency of the topic (per 10,000 words) as a function of time. Common and custom stopwords (such as articles and conjunctions and proper nouns) are replaced with ‘substitute_word’ to drastically reduce our vocabulary without much information loss.

Topic modeling is better to extract a sub-corpus than word frequencies

58 We present a land and property example that illustrates how topic modeling goes beyond word frequencies for historical analysis. Figure 4 shows word frequencies for the most commonly occurring words in topic 35. The word frequencies are normalized by the total number of unigrams in each year and are shown per 10,000 words. Four peaks stand out: “person” in the early part of the period and “land” in the late part of the period. “Land” and “properti” have a few smaller peaks throughout the period. The pattern does have some historical significance, but it is certainly noisier than the peaks in topic frequency seen in Figure 3. There is no clear peak in the middle of the century where we would expect to see massive activity about this topic, which could potentially lead us to miss some interesting and important documents. Furthermore, if we were to search for documents of interest in Hansard using these frequently occurring keywords, we would likely incur many false negatives that are skewed towards the years in which the word counts peaked.

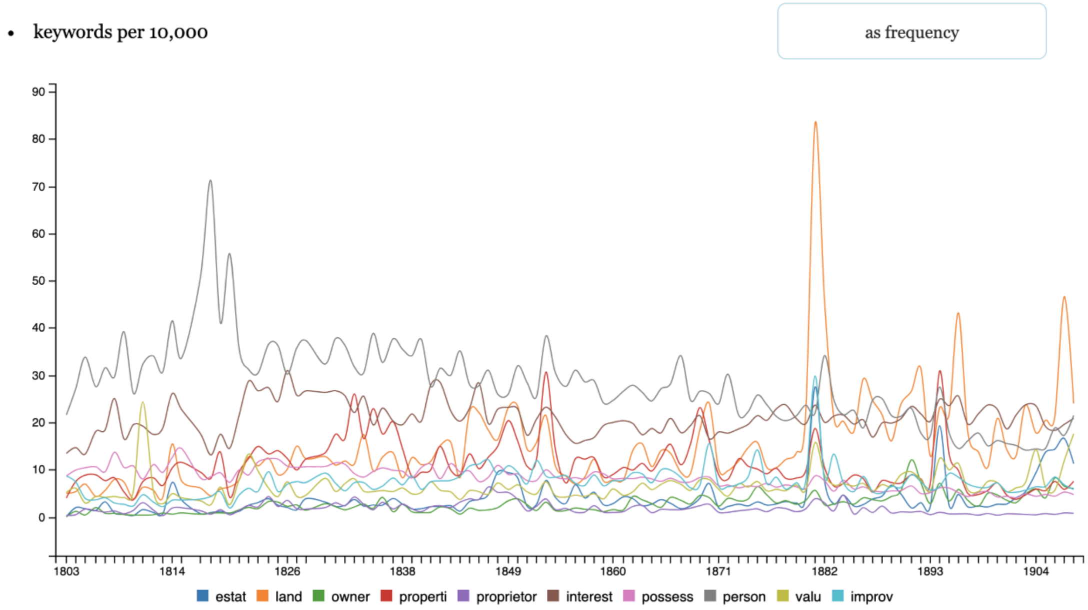


Figure 4. Unigram Frequencies. This figure shows the relative frequencies of the most commonly occurring unigrams in topic 35. The figure shows counts per 10,000 unigrams, normalized by the total number of unigrams in the year, on the y-axis and year on the x-axis. The diagram demonstrates a transition from a language of “person” and “interest”, visible in the dominance of yellow and pink lines around 1815, to a language where spikes in 1882, 1896, and 1907 were marked by the use of the terms “land”, “value”, and “improvement”.

D3 visualizations and web app

59 The screenshots shown in Figures 3 and 4 are taken from the HaToRI web app. The tool allows humanists to interact with their corpora in a user-friendly, visual, iterative, and exploratory process and avoids the time and resource penalties that traditional search and close reading of corpora incur. We will walk through the various features of the tool with accompanying screenshots in this next section with an emphasis on how the tool enhances digital humanities research by incorporating some of the values of code pipelines discussed in the introduction.

60 The tool approaches visual topic exploration at different levels or views: word, document, and topic. Taken together in a single interface, these levels give the user a three-dimensional view of their corpus that is less possible when searching for specific words, documents, or topics in silo. The n-gram plot in Figure 5 shows word frequencies per year for the top six words in a topic. Each line in the plot shows frequencies for a given word and each word in the plot is also shown on the right as a button; they can be removed from the plot by clicking on the words and the plot will dynamically re-scale based on the frequencies present in the current plot window. Words can be added by searching the vocabulary in the search bar. The search results filter down as the user types a query because corpora are often pre-processed differently, resulting in a vocabulary made up of stems or lemmas rather than actual words. The frequencies can be shown in absolute terms or relative, normalized to the total number of words in the year. The word view gives the user a

finer-grained view of the corpus, but leaves out the greater context and interaction of the words.

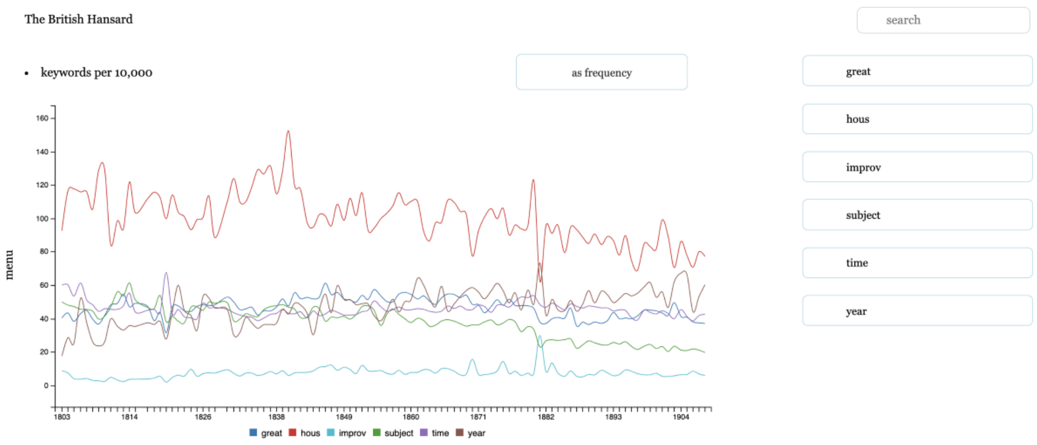


Figure 5. Word level exploration of the corpus. The default is to display the top six words in each topic. Users can add or remove words with the search bar and it auto-fills with unigrams in the corpus vocabulary. The user can switch between absolute and relative counts.

Figures 6 and 7 show topic views of the corpus. The home page shows the first view, where each topic is a word cloud showing the top words in the topic. The size of the word indicates its probability within the topic; bigger words are more likely to be sampled from the topic than smaller words. The page scrolls and loads one chunk of word clouds at a time. The user can filter to a smaller subset of topics by searching for unigrams in the search bar at the top right corner of the page. The alternate view shows a table view of the topics with an additional sort functionality that will sort the topics based on any of the table fields. They can be sorted by their ID, by mean peak year, alphabetically by words, or by their proportion in the corpus. The recommended navigational flow of the corpus exploration part of the tool is to start at one of these two topic overviews and drill down closer into a single topic. Figure 8 shows a detailed topic view. The view includes more top words than the overviews, a plot of the topic over time, a link to the n-gram plot, and a sorted table of documents that are most associated with the topic. The detail view allows the user to explore the topic from various different vantage points, but does not allow multi-topic exploration. All topics are shown in the plot on the left in Figure 9, and they are arranged based on their similarity to each other. Topics with very similar word distributions cluster together and topics with very dissimilar word distributions are very far apart.

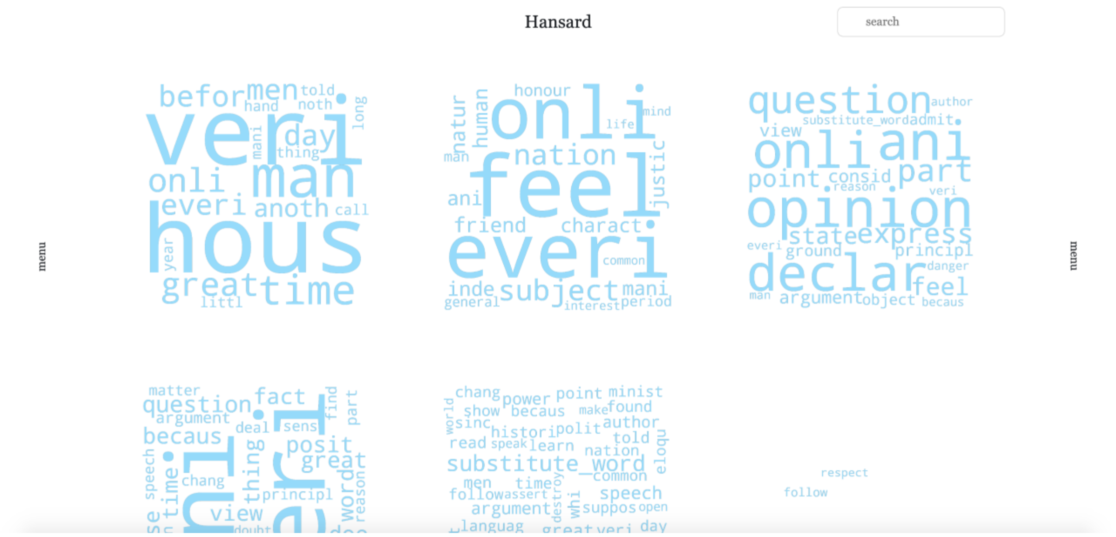
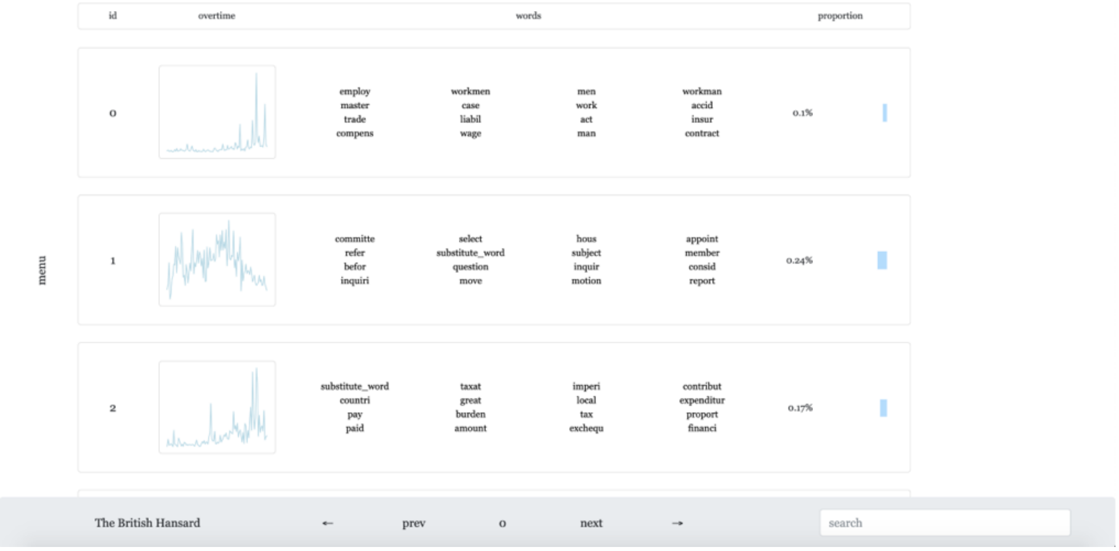
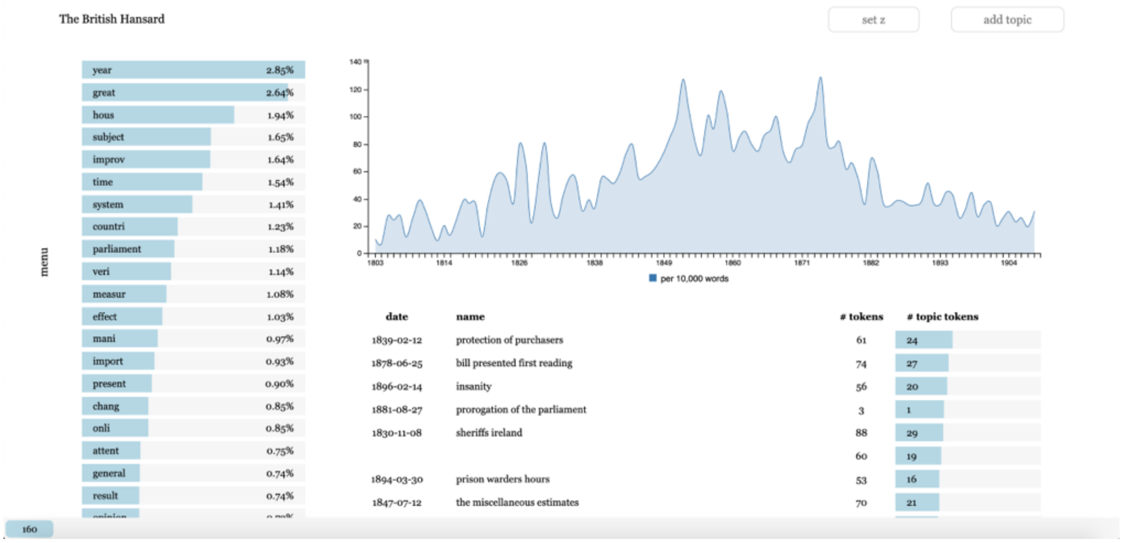


Figure 6. The front page of the web app for topic level corpus exploration. Each topic is a word cloud and starting point with which to move into a more detailed view of the topic. The word clouds show topic distribution over time on hover and they are searchable by word.



id	overtime	words				proportion
0		employ master trade compens	workmen case liabil wage	men work act man	workman accid insur contract	0.1%
1		committe refer befor inquiri	select substitute_word question move	hous subject inquir motion	appoint member consid report	0.24%
2		substitute_word countri pay paid	taxat great burden amount	imperi local tax exchequ	contribut expenditur proport financi	0.17%

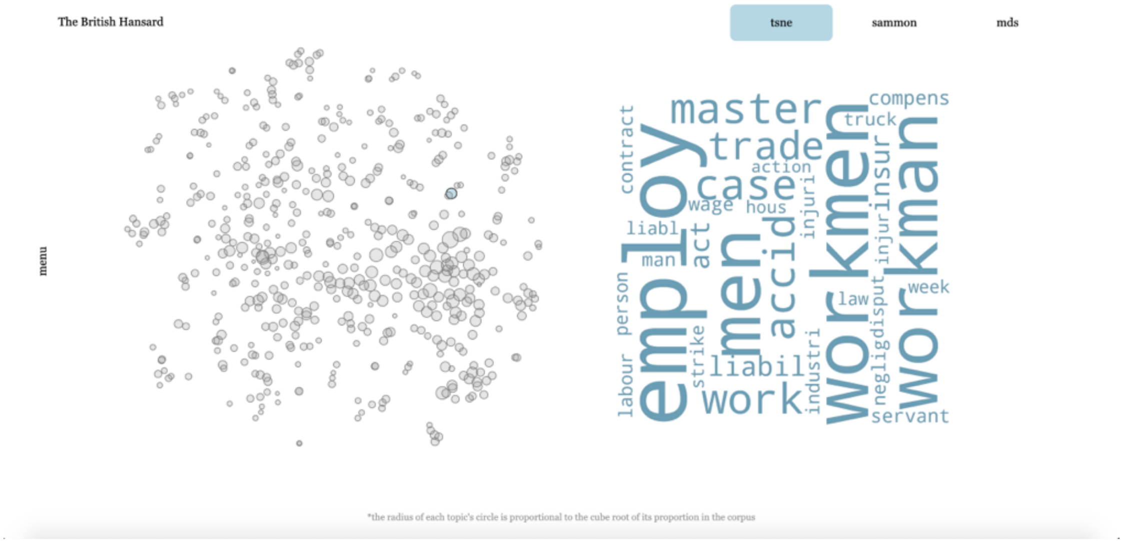
Figure 7. Alternate table view of home page. Sortable by topic ID, mean peak time, alphabetically by words, and by proportion in corpus.



year	2.85%
great	2.64%
hous	1.94%
subject	1.65%
improv	1.64%
time	1.54%
system	1.41%
countri	1.23%
parliament	1.18%
veri	1.14%
measur	1.08%
effect	1.03%
mani	0.97%
import	0.93%
present	0.90%
chang	0.85%
onli	0.85%
attent	0.75%
general	0.74%
result	0.74%

date	name	# tokens	# topic tokens
1839-02-12	protection of purchasers	61	24
1878-06-25	bill presented first reading	74	27
1896-02-14	insanity	56	20
1881-08-27	prorogation of the parliament	3	1
1830-11-08	sheriffs ireland	88	29
		60	19
1894-03-30	prison warders hours	53	16
1847-07-12	the miscellaneous estimates	70	21

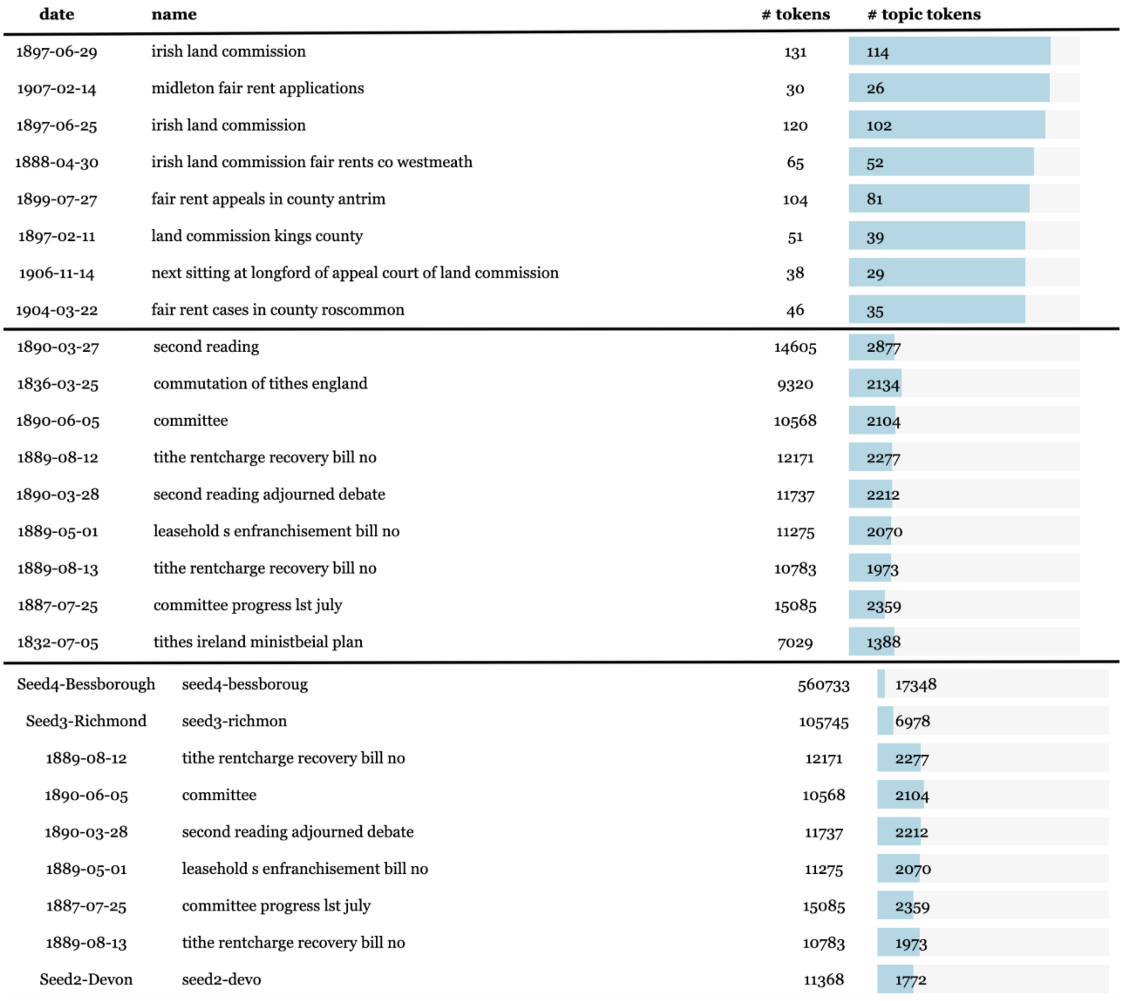
Figure 8. Detail view of topic. Ranks words and ranks documents, and for both shows relevance to the topic. Topic relevance over time plot.



*the radius of each topic's circle is proportional to the cube root of its proportion in the corpus

Figure 9. Topic Clustering view shows how similar topics are to each other.

Figure 10 shows ranked documents relevant to the five land-related topics shown in Table 2 for z = [0, 100, 500]. For land topics, larger z values produce a better ranking than smaller z values. There are documents from the early and late periods of the century, near the peaks of land-related activity that we expect. Also, three out of four written reports, denoted by the prefix “seed-”, are in the top-ranked documents list when z = 500; we expect the written reports to be ranked highly because the historical literature confirms that they are key documents summarizing land use in the nineteenth-century. The low proportion of topic tokens out of total tokens in the written reports — 0.031, 0.066, and 0.156 for the Bessborough, Richmond, and Devon reports, respectively — indicates that the seed reports contain a lot of noise and shows the weakness of an earlier document-based approach compared to our topic-based approach [Lee et al. 2018] [Tangherlini and Leonard 2013].



date	name	# tokens	# topic tokens
1897-06-29	irish land commission	131	114
1907-02-14	midleton fair rent applications	30	26
1897-06-25	irish land commission	120	102
1888-04-30	irish land commission fair rents co westmeath	65	52
1899-07-27	fair rent appeals in county antrim	104	81
1897-02-11	land commission kings county	51	39
1906-11-14	next sitting at longford of appeal court of land commission	38	29
1904-03-22	fair rent cases in county roscommon	46	35

1890-03-27	second reading	14605	2877
1836-03-25	commutation of tithes england	9320	2134
1890-06-05	committee	10568	2104
1889-08-12	tithe rentcharge recovery bill no	12171	2277
1890-03-28	second reading adjourned debate	11737	2212
1889-05-01	leasehold s enfranchisement bill no	11275	2070
1889-08-13	tithe rentcharge recovery bill no	10783	1973
1887-07-25	committee progress lst july	15085	2359
1832-07-05	tithes ireland ministbeial plan	7029	1388

Seed4-Bessborough	seed4-bessboroug	560733	17348
Seed3-Richmond	seed3-richmon	105745	6978
1889-08-12	tithe rentcharge recovery bill no	12171	2277
1890-06-05	committee	10568	2104
1890-03-28	second reading adjourned debate	11737	2212
1889-05-01	leasehold s enfranchisement bill no	11275	2070
1887-07-25	committee progress lst july	15085	2359
1889-08-13	tithe rentcharge recovery bill no	10783	1973
Seed2-Devon	seed2-devo	11368	1772

Figure 10. Top nine most relevant documents ranked using a Wilson-score based sorting method with a single parameter, z. From top to bottom, the tables show results for z = [0, 100, 500].

Summary

In this paper, we introduce the pipeline concept as a means for scholars to conduct transparent, modular, and interoperable research and we create a web app that exemplifies these values. We make the HaToRI instance of the web app digitally available, and present a use case for doing enhanced historical analysis on the Hansard text corpus using the app. We explored how topic modeling can be used for sub-corpus extraction when other methods likely fail due to the sheer size of the corpus or nuances in the topics of interest that cannot be accurately captured by keywords or terms. To illustrate the advantages of topic modeling in sub-corpus extraction, we extracted a sub-corpus specific to land use, rent, and property from the Hansard British Parliamentary debates of the 19th century. There are two historically relevant periods when land use and property was exhaustively discussed in the Parliament: the 1850's and 1880's. We illustrate that topic-based sub-corpus extraction identifies these two eras based on the frequencies of the relevant topics, while keywords tend to wash out the signal and identify false peaks in the frequency plot. We described our pre-processing and post-processing steps in detail and provide a web app with a set of visualizations and functionalities to aid humanist research on the Hansard corpus. These functionalities include topic clustering based on similarity, document ranking based on topic prevalence, and frequency plots. The source codes for the pipeline and the web app are publicly available and can be modified to work with other corpora. We believe that topic modeling, in particular, as well as our novel document ranking algorithm, are well suited to extract a relevant sub-corpus for in-depth studies, especially if the corpus is prohibitively long to read by humans.

Acknowledgments

Thanks to Mark Howison at Brown University for facilitating the collaboration with Prof. Jo Guldi and contributing early work to this project. This research was conducted using computational resources and services at the Center for Computation and Visualization, Brown University.

Notes

[1] https://docs.cortext.net/

[2] The code is open-sourced (MIT license), so it is free to use, modify, and share. The pipeline code can be found at https://github.com/brown-data-science/inquiry-for-philologic-analysis, and the web-app code, both front end and back end, at https://github.com/brown-data-science/hansard_api. The visualizations and front end are built using Javascript frameworks (`d3.js` and `vue.js`) and the back end is deployed using Docker containerization. The web app is hosted on github.io at https://eight1911.github.io/hansard.

[3] https://pages.github.com/

[4] http://www.hansard-archive.parliament.uk/

[5] http://www.hansard-archive.parliament.uk/

Works Cited

Agresti and Coull 1998 Agresti, A. and Coull, B. A. *Approximate is Better than “Exact” for Interval Estimation of Binomial Proportions*, *The American Statistician*, 52 (1998): 119-126.

Alexander and Struan 2017 Alexander, M. and Struan, A. “Digital Hansard: Politics and the Uncivil”, *Digital Humanities*, (2017): 378-380

Berry and Browne 2005 Berry, M. W. and Browne, M. “Email Surveillance Using Non-negative Matrix Factorization”, *Computational and Mathematical Organization Theory*, 11.3 (2005): 249-264.

Black 1960 Black, R. D. C. *Economic Thought and the Irish Question, 1817-1870*. University Press, Cambridge, MA (1960).

Blaxill 2013 Blaxill, L. “Quantifying the Language of British Politics, 1880–1910”, *Historical Research*, 86 (2013): 313-41.

Blei and Lafferty 2006 Blei, D. M. and Lafferty, J. D. “Dynamic Topic Models”, In *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, June 2006.

Blei and Lafferty 2009 Blei, D. M. and Lafferty, J. D. “Topic Models”, In A. Srivastava and M. Sahami (eds), *Text Mining: Theory and Applications*, Taylor and Francis, London (2009).

Blei et al. 2003 Blei, D. M., Ng, A. Y. and Jordan, M. I. “Latent Dirichlet Allocation”, *Journal of Machine Learning Research*, 3 (2003): 933-1022.

Brooke et al. 2015 Brooke, J., Hammond, A. and Hirst, G. “GutenTag: An NLP-driven Tool for Digital Humanities Research in the Project Gutenberg Corpus”, In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, Denver, CO, June 2015.

Bull 1996 Bull, P. Land, *Politics and Nationalism: A Study of the Irish Land Question*. Gill and Macmillan, Dublin (1996).

Butterfield et al. 2016 Butterfield, A., Ngondi, G.E. and Kerr, A. A *Dictionary of Computer Science*, Seventh Edition. Oxford University Press, Oxford (2016).

Campbell 2005 Campbell, F. J. M. *Land and Revolution Nationalist Politics in the West of Ireland, 1891-1921*. Oxford University Press, Oxford, New York (2005).

Connelly 2003 Connelly, S. J. “Jacobites, Whiteboys, and Republicans: Varieties of Disaffection in Eighteenth-Century Ireland”, *Eighteenth-Century Ireland/Iris an Da Chultur*, 18 (2003): 63-79.

Donnelly 1983 Donnelly, J. S. 1983. “Irish Agrarian Rebellion: The Whiteboys of 1769-76”, *Proceedings of the Royal Irish Academy, Section C: Archaeology, Celtic Studies, History, Linguistics, Literature* (1983): 293-331.

Edmond 2013 Edmond, J. “CENDARI’s Grand Challenges: Building, Contextualising and Sustaining a New Knowledge Infrastructure”, *International Journal of Humanities and Arts Computing*, 7.1-2 (2013): 58–69. doi:10.3366/ijhac.2013.0081

Edmond et al. 2015 Edmond, J., Bulatovic, N. and O'Connor, A. "The Taste of 'Data Soup' and the Creation of a Pipeline for Transnational Historical Research", *Journal of the Japanese Association for Digital Humanities* 1.1 (2015): 107-122. doi:10.17928/jjadh.1.1_107

Edwards et al. 2007 Edwards, P.N., Jackson, S.J., Bowker, G.C., Knobel, C.P. "Understanding Infrastructure: Dynamics, Tensions, and Design", *Human and Social Dynamics* (2007). NSF Grant 0630263.

Eubanks 2015 Eubanks, V. *Automating Inequality: How High-tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, New York, NY (2015).

Goldstone and Underwood 2014 Goldstone, A. and Underwood, T. "The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us", *New Literary History*, 45.3 (2014): 359-384. doi: 10.1353/nlh.2014.0025

Grigor 2000 Grigor, I. F. *Highland Resistance*. Mainstream Publishing, Edinburgh (2000).

Guldi 2018 Guldi, J. "Critical Search: A Procedure for Guided Reading in Large-Scale Textual Corpora", *Journal of Cultural Analytics* (2018). doi: 10.31235/osf.io/g286e

Harris 1954 Harris, Z. "Distributional Structure", *Word*, 10.2-3 (1954): 146-62.

Introna and Nissenbaum 2006 Introna, L.D. and Nissenbaum, H. "Shaping the Web: Why the Politics of Search Engines Matters", *The Information Society: An International Journal* 16.3: 169-185 (2006).

Janicke et al. 2015 Janicke, S., Franzini, G., Cheema, M. F. and Scheuermann, G. "On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges", *Eurographics Conference on Visualization State of the Art Report* (2015): 21.

Kaplan 2015 Kaplan, F. "A Map for Big Data Research in Digital Humanities", *Frontiers in Digital Humanities*, 2 (2015). doi:10.3389/fdigh.2015.00001

Le and Mikolov 2014 Le, Q. V. and Mikolov, T. "Distributed Representation of Sentences and Documents", In *Proceedings of the Thirty-first International Conference on Machine Learning Beijing*, CN, June 2014.

Lee et al. 2018 Lee, A. S., Guldi, J. and Zsom, A. "Measuring Similarity: Computationally Reproducing the Scholar's Interests", arXiv:1812.05984 [cs.CL] (2018).

Llora et al. 2008 Llora, X., Acs, B., Auvil, L. S., Capitanu, B., Welge, M. E. and Goldberg, D. E. "Meandre: Semantic-Driven Data-Intensive Flows in the Clouds", In *IEEE Fourth International Conference on eScience*, Indianapolis, IN, December 2008. doi: 10.1109/eScience.2008.172

Manning and Schutze 1999 Manning, C. D. and Schutze, H. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA (1999).

Mattern 2013 Mattern, S. "Infrastructural Tourism", *Places Journal* (2013). doi: https://doi.org/10.22269/130701

Mattern 2015 Mattern, S. "Deep Time of Media Infrastructure", In Lisa Parks and Nicole Staroeislski (eds.), *Signal Traffic: Critical Studies of Media Infrastructures*. University of Illinois Press, Champaign, IL (2015).

Mattern 2016 Mattern, S. "Scaffolding, Hard and Soft - Infrastructures as Critical and Generative Structures", *Spheres Journal for Digital Cultures*, 3 (2016).

McCallum 2002 McCallum, A. K. "MALLET: A Machine Learning for Language Toolkit", *Scientific Research*, (2002). http://mallet.cs.umass.edu.

Mikolov et al. 2013 Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J. "Distributed Representations of Words and Phrases and Their Compositionality", *Advances in Neural Information Processing Systems*, (2013).

Murdock et al. 2017 Murdock, J., Allen, C., Borner, K., Light, R., McAlister, S., Ravenscroft, A., Rose, R., Rose, D., Otsuka, J., Bourget, D., Lawrence, J., and Reed, C. "Multi-level Computation Methods for Interdisciplinary Research in the HathiTrust Digital Library", *PLoS One*, 12.0 (2017).

Nissenbaum 2010 Nissenbaum, H. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press, Stanford, CA (2010).

Noble 2018 Noble, S.U. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, New York, NY (2018).

O'Neill 2016 O'Neil, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*.

Broadway Books, New York, NY 2016.

Readman 2008 Readman, P. *Land and Nation in England: Patriotism, National Identity, and the Politics of Land, 1880-1914*. Boydell Press, Woodbridge, UK (2008).

Rule et al. 2015 Rule, A., Cointet, J. and Bearman, P. S. “Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014”, In *Proceedings of the National Academy of Sciences*, 112.35 (2015): 10837-44.

Sartori 2014 Sartori, A. *Liberalism in Empire: An Alternative History*. University of California Press, Berkeley (2014).

Sinclair and Rockwell 2016 Sinclair, S. and Rockwell, G. “Voyant Facts”, Hermeneuti.ca: Computer-Assisted Interpretation in the Humanities (2016). http://hermeneuti.ca/VoyantFacts.

Steele 1974 Steele, D. *Irish Land and British Politics: Tenant-Right and Nationality, 1865-1870*. Cambridge University Press, London (1974).

Svensson 2011 Svensson, P. “From Optical Fiber to Conceptual Cyberinfrastructure”, *Digital Humanities Quarterly*, 5.1 (2011).

Tangherlini and Leonard 2013 Tangherlini, T. and Leonard, P. “Trawling in the Sea of the Great Unread: Sub-corpus topic modeling and Humanities research”, *Poetics*, 41.6 (2013): 725-749.

Tucker 2004 Tucker, A. B. *Handbook of Computer Science*. CRC Press, Boca Raton (2004).

Vaidhyanathan 2012 Vaidhyanathan, S. *The Googlization of Everything: (and why we should worry)*. University of California Press, Los Angeles, CA (2012).

Vaidhyanathan 2018 Vaidhyanathan, S. *Anti-Social Media: How Facebook Disconnects Us and Undermines Democracy*. Oxford University Press, New York, NY (2018).

d’Alessandro et al. 2016 d’Alessandro, B., O’Neil, C. and LaGatta, T. “Conscientious Classification: A Data Scientist’s Guide to Discrimination-Aware Classification”, *Big Data*, 5.2 (2017). doi: http://doi.org/10.1089/big.2016.0048