HistText: An Application for leveraging large-scale historical textbases

Blouin Baptiste, Cécile Armand, Christian Henriot

HistText: An Application for leveraging large-scale historical textbases

Cécile Armand (Aix-Marseille University)
Baptiste Blouin (Aix-Marseille University)
Christian Henriot (Aix-Marseille University)

Abstract:

This paper introduces HistText, a pioneering tool devised to facilitate large-scale data mining in historical documents, specifically targeting Chinese sources. Developed in response to the challenges posed by the massive Modern China Textual Database, HistText emerges as a solution to efficiently extract and visualize valuable insights from billions of words spread across millions of documents. With a user-friendly interface, advanced text analysis techniques, and powerful data visualization capabilities, HistText offers a robust platform for digital humanities research. This paper explores the rationale behind HistText, underscores its key features, and provides a comprehensive guide for its effective utilization, thus highlighting its potential to substantially enhance the realm of computational humanities.

1. Introduction

Historians are currently grappling with the significant challenge posed by large-scale digital corpora. In the early 1990s, when historical documents were first converted into digital format as scanned images, the field of "digital history" emerged, allowing historians to still rely on established methods that had been developed over the past 150 years.[1] However, the landscape has changed considerably since then. The digital transformation of historical

[1] Michael J. Galgano, J. Chris Arndt, and Raymond M. Hyser, *Doing History: Research and Writing in the Digital Age*, 1st ed (Boston, MA: Thomson Wadsworth, 2008); Toni Weller, *History in the Digital Age* (London; New York: Routledge, 2013); Kristen Nawrotzki; Jack Dougherty, *Writing History in the Digital Age*, 2013; "The Promise of Digital History," *The Journal of American History* 95, no. 2 (September 2008), http://www.journalofamericanhistory.org/issues/952/interchange/.

sources has not only revolutionized access to and distribution of these sources, but has also transformed the way historical information is disseminated and historical narratives are constructed. In recent years, there has been an explosive proliferation of digital full-text resources, comparable to a "Big Bang" in the field of history. These resources have not only provided alternative versions of printed sources, but have also given rise to an infinite constellation of dematerialized and unconnected texts, encompassing billions of words.

The digital transformation affects all historical sources, especially the vast repository of textual documents (press, archives). In Western societies where alphabets are the norm and typewriters have been in use since the late nineteenth century, not only published materials but also archival documents, the very essence of historical research, are being digitized. For instance, in the United States, the National Archives and Records Administration (NARA) is making entire series available after digitization and OCR processing.[2] Furthermore, with the combination of OCR and AI, handwritten documents can now be transformed into full text as well. Even medieval manuscripts, although requiring meticulous curation, are undergoing their own digital transformation. New tools have been developed to tackle manuscript issues across different languages. Platforms like Transkribus offer various models and algorithms specifically designed for handwritten manuscripts.[3]

While Optical Character Recognition (OCR) technology has made remarkable advancements, resulting in the ability to convert scanned images into full text with increasingly fewer errors, the presence of errors persists to some extent, depending on the quality of the original document, OCR technology and algorithms.[4] Nevertheless, the vast corpora of digital texts, encompassing daily newspapers, periodicals, academic literature, books, dictionaries, encyclopedias, directories, and more, cannot be ignored or dismissed, even with some degree of error. These digital avatars raise new hermetic issues. In an ideal world, all digital versions of a source would be linked to its original image format to enable historians to go back to the original, albeit in digital image format. However, in the real world dominated by commercial providers, such links either do not exist or are tenuous at best. Nonetheless, with the proper metadata, it is still possible in most cases to trace back to the original document.

This is where academic research plays an invaluable role. The challenge in transforming scanned images into full text extends beyond OCR; it also involves text segmentation. While identifying chapters and paragraphs in a book is the primary concern and a fairly simple operation, in the case of journals or daily newspapers, segmenting scanned pages into individual articles, which serve as the fundamental units of research, becomes essential. Manual transformation of these massive quantities of newspapers found in libraries is simply impractical. Therefore, computational methods based on human annotations, metadata enrichment, and training models capable of automatically recognizing and labeling paragraphs are necessary. The Newseye project, conducted by a consortium of European computing labs, has successfully developed a workflow to address this issue.[5] The ENP-China project is currently conducting a campaign of annotations on a set of Chinese- and English-language newspapers and periodicals using the Coco annotation tool[6].

---

[2] Electronic documents at NARA can be accessed from here: https://www.archives.gov/research/electronic-records

[3] Transkribus: https://readcoop.eu/transkribus/

[4] Guillaume Chiron et al., "Impact of OCR Errors on the Use of Digital Libraries: Towards a Better Access to Information," in *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (IEEE, 2017), 1–4.

[5] Newseye project: https://www.newseye.eu/

[6] Brooks, Justin. 2019. "COCO Annotator." https://github.com/jsbroks/coco-annotator/.

The digitization of printed documents such as newspapers and books pales in comparison to the deluge of digital data that awaits historians. Today, students of contemporary history have access to an unimaginable amount of information on the internet. Ordinary individuals have been shaping their own history through their voices on digital platforms since the mid-1990s.[7] Mega-companies like Google, Facebook, Twitter, and Amazon in the West, as well as Alibaba, Tencent, ByteDance, and others in China, provide vast repositories of data for studying contemporary societies. Open-source platforms like Wikipedia, Internet Archive, Baidu (China), and numerous others fulfill a similar role in constructing boundless reservoirs of information. Additionally, public administrations themselves have embraced digitalization, posing the challenge of defining what constitutes a "document" when the billions of emails exchanged are stored for future reference. How will future historians cope with this overwhelming volume of historical data without proper preparation and adequate tools?

From its inception, the ENP-China project was conceived to confront the challenge of large-scale digital corpora and develop solutions to help historians navigate the digital deluge. It also placed the issue of digital source criticism and digital hermeneutics at the forefront of its work. The ENP-China team engaged in intense interdisciplinary work at the intersection of history, corpus linguistics, and computing, exploring a wide range of methods inspired by Natural Language Processing and AI. One of the project's major achievements is HistText, a user-friendly application that leverages AI to explore, mine, and interpret the infinite digital collections available to historians and other scholars in the humanities and beyond.

1.1 Background and Context

HistText emerged from the need to effectively utilize textual data within extensive collections of multilingual and heterogeneous texts, such as the corpora compiled for studying modern China in the Modern China Textual Database (MCTB). Although one of HistText's notable features is the exploration of Chinese historical texts, its capabilities extend to corpora in other languages as well. The development of this application is contextualized within the ERC-funded ENP-China project, which investigates the evolution of Chinese elites from the nineteenth century to 1949.[8]

From a methodological standpoint, the project aims to overcome the limitations of manual curation of historical documents by historians for collecting and processing historical information. The recent availability of large-scale full-text corpora, including newspapers, periodicals, archives, who's who directories, and Wikipedia, presents a unique opportunity to leverage methodologies derived from other disciplines, such as computational linguistics and natural language processing, for approaching historical corpora. The emergence and rapid evolution of large-scale language models have provided a crucial element for harnessing the computational methods' potential in textual analysis.

---

[7] Ian Milligan, *History in the Age of Abundance?: How the Web Is Transforming Historical Research* (Montreal: McGill-Queen's University Press, 2019).

[8] ENP-China project: https://www.enpchina.eu/. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 788476).

The ENP-China team — an interdisciplinary group of scholars in history, data analysis, and computing — faced the challenge of creating not just a digital work environment, but to provide all the members with the tools that met their respective objectives and expectations. One of the difficulties was to enable the non-programming historians with the capacity to harness the resources in the vast textual repository at their disposal. The choice was made to adopt the R programming language as the shared language between historians and computer scientists. Yet, we also realized that training in a programming language did not remove all the stumbling blocks on the road to designing queries that matched the needs of historical inquiry.

There was not just a learning curve, but also a cost in leaving the basic operations of data mining to individual processes. It could not lead to a process of accumulation to unlock the fruitful access to large-scale textual corpora. To overcome this limitation, the computer scientists in the ENP-China team designed an internal tool — originally named "enpchina R Library" — that provided ready-made functions to mine data in digital corpora. This library was made up of a series of search functions that covered the most repetitive tasks in querying a corpus. The initial set of search functions eventually led to further experiments through the sustained conversation between historians and computer scientists about fitting even more closely and more appropriately the potential of computational methods to the various steps of historical inquiry.

In this paper, we introduce the purpose and key features of HistText, in its R-based version for programming historians (and humanists) and the R-Shiny-based user-friendly interface. The R-based version for programming historians requires basic knowledge in R, but it offers a wide range of elaborate functions to explore and process data from historical texts. R-Shiny-based user-friendly interface includes the main functions from its parent version, with automatic statistical computing and various visualisations. Both can be used in parallel. In the second section of the paper, we review existing textual databases and the applications, incorporated therein or standalone, that present similar functionalities. The third section of the paper presents case studies to demonstrate how HistText operates and what results a scholar can obtain from its use. The last section discusses the broader contribution of HistText to computation methods in the humanities and further developments.

2. State of the Art

This section provides a brief overview of the existing textbases and text mining tools available to China scholars. It is not meant to be an exhaustive survey; rather, it aims to classify the existing resources in order to better situate MCTB-HistText in the field and to emphasize its specific, unique features. We provide some representative examples for each category, among the most popular and widely used by China scholars. Incidentally, whenever appropriate, we also refer to relevant cases outside the China field (2.3). We are aware that such a survey faces the risk of being out of date in a few months because tools are evolving quickly. Nevertheless, we believe it gives a fairly accurate state of the field as of July 2023.

2.1 Overview of Textbases

Building on Cooney et al. (2013), three main models of textbases can be distinguished (1) the representation model refers to textbases that emphasize the preservation and display of texts (2) mixed-mode textbases combine the display of texts with efforts to provide greater access to text contents, metadata, and annotation tools for semantic enrichment (3) data-driven textbases place a stronger emphasis on data mining and text analysis.[9]

Commercial providers and publishers, such as Airusheng[10] and Green Apple[11] for Chinese-language resources, ProQuest[12] and Brill[13] for English-language periodicals, fall under the first category. While these commercial platforms have significantly contributed to the accessibility of historical materials, they are primarily focused on document consultation and keyword search, offering limited possibilities to interact with the full text of documents, let alone manipulate text data. Moreover, while they provide access to a wide range of text collections, they usually require payment at costs that may be prohibitive to many institutions.

By contrast, university, research institutes, and public libraries such as Heidelberg University Early Chinese Periodicals Online (ECPO)[14], the Institute of Modern History (IMH) collections at Academia Sinica[15], the Shanghai Library[16], Chinese University of Hong Kong University Library[17] do not require payment for consultation. Most of these public textbases also fall under the representation model, although some are now shifting towards a mixed-mode approach, which appears to be better tailored to researchers' needs. The above-mentioned textbases have begun to incorporate automatic layout recognition, semantic enrichment and entity linking, as well as topic modeling and authorship attribution to enable more refined queries and to expand possibilities for exploration.[18] However, these functionalities are still in an experimental stage and the possibilities to interact with the texts remain limited. Notably, they do not allow researchers to select, structure, and build datasets in a personalized space, which is a key recommendation in recent surveys. Without the

[9] Charles Cooney, Glenn Roe, and Mark Olsen, "The Notion of the Textbase: Design and Use of Textbases in the Humanities," in *Literary Studies in the Digital Age. An Evolving Anthology* (New York, N.Y.: Modern Language Association, 2013).
[10] http://server.wenzibase.com/
[11] https://www.egreenapple.com/english/channels/136.html
[12] https://about.proquest.com/en/products-services/hnp_cnc/
[13] https://brill.com/browse?et=ponline&level=parent&pageSize=10&sort=datedescending&t=04-04
[14] https://kjc-sv034.kjc.uni-heidelberg.de/ecpo/
[15] https://mhdb.mh.sinica.edu.tw/mhpeople/index.php
[16] https://dhc.library.sh.cn/
[17] https://dsprojects.lib.cuhk.edu.hk/en/projects/
[18] Matthias Arnold and Henrike Rudolph, "Network Data in the Early Chinese Periodicals Online Database (ECPO)," *Journal of Historical Network Research* 5, no. 1 (September 8, 2021); Matthias Arnold, Duncan Paterson, and Jia Xie, "Procedural Challenges: The FAIR Principles and PRC Electronic Resources - a Case Study of Chinese Republican Newspapers," *International Journal of Digital Humanities* 4, no. 1 (February 1, 2023): 147–70.

possibility to create customized subcorpora, humanities researchers often cannot align their analyses with their research questions.[19]

2.3 Toolkits for Chinese

On the other hand, initiatives like C-Text,[20] Markus,[21] LoGaRT,[22] as well as toolkits developed by Chinese universities, such as gj.cool,[23] provide advanced tools for markup, semantic enrichment and text analysis, based on existing dictionaries and other lexical resources. However, they focus primarily on ancient Chinese texts and target specific genres of texts (literary, philosophical, gazetteers). They fail to address the specific set of challenges posed by modern Chinese texts spanning from the late Qing dynasty to the People's Republic of China (19th century-1949), namely (1) the unprecedented abundance and diversity of texts produced during the era of print capitalism and transnational exchanges under the "unequal treaties" regime; (2) the great instability of the Chinese language during this critical period of nation building and linguistic reform,[24] and (3) issues and biases that result from the transformation of texts into digital objects (noisy OCR, OLR, etc.).

To the best of our knowledge, MCTB-HistText is the only data-mining textbase focused on modern China that has seriously tackled these challenges to date.

2.3 Newspaper Interfaces

Because the periodical press initially constitutes the core of MCTB, HisText is more akin to newspapers interfaces like the British Newspaper Archive,[25] NewsEye,[26] and Impresso.[27] However, it presents three major differences, which reflect the distinct origins, ambitions, and team composition of these projects. Impresso and NewsEye are large-scale cultural heritage projects which involve computer scientists, historians, and librarians from major public libraries in Europe. They aim primarily at filling the increasing gap between users' expectations and current interfaces. The British Newspaper Archives is an ambitious project that digitized more than 68 million pages. Yet, their content is behind a paywall. On the other hand, MCTB is born from a specific research project (ENP-China) which focuses on the

---

[19] Maud Ehrmann, Estelle Bunout, and Marten Düring, "Historical Newspaper User Interfaces: A Review," in *IFLA WLIC 2019* (IFLA WLIC 2019, Athens, 2019); Eva Pfanzelter et al., "Digital Interfaces of Historical Newspapers: Opportunities, Restrictions and Recommendations," *Journal of Data Mining and Digital Humanities* HistoInformatics (January 2021).

[20] https://ctext.org/

[21] https://dh.chinese-empires.eu/markus/beta/

[22] https://www.mpiwg-berlin.mpg.de/research/projects/logart-local-gazetteers-research-tools

[23] https://gj.cool/

[24] Elisabeth Kaske, *The Politics of Language in Chinese Education, 1895-1919*, Sinica Leidensia (Leiden: Brill, 2008); Pierre Magistry, "Languages(s) of the Shun-Pao, a Computational Linguistics Account," in *10th International Conference of Digital Archives and Digital Humanities* (Taipei, Taiwan, 2019); Jing Tsu, *Kingdom of Characters: The Language Revolution That Made China Modern* (New York: Riverhead Books, 2022), Blouin et al., "Unlocking Historical Chinese: A Study on Word Segmentation in Transitional Chinese Texts," (unpublished manuscript).

[25] https://www.britishnewspaperarchive.co.uk/

[26] https://www.newseye.eu/

[27] https://impresso-project.ch/

transformation of elites in modern China. From the onset, ENP-China has placed a strong emphasis on data extraction and linking, with the specific objective to build two major biographical and geospatial databases to facilitate this study (MCBD, MGBC). It is only at a later stage that the project has expanded to include, almost accidentally, the development of a dedicated interface – HistText – with advanced functionalities similar to NewsEye and Impresso. Nevertheless, our objectives and challenges remain distinct:

1. MCTB comprises a broader range of texts (not just newspapers) and languages (including non-Western, low-resource languages, notably "transitional" Chinese) (see section 3.1). In contrast to cultural heritage projects like Impresso and NewsEye, we did not start from the collections already available at certain libraries. Instead, we built our own collections, following our specific research needs, discoveries, and how they evolved over time.

2. NewsEye and Impresso target a broad user base including both academic and non-academic users, who mostly utilize the interface to find and select a reasonable number of documents for close reading. These projects aim at developing user-friendly interfaces to support corpus building through advanced search functionalities, filtering options, and semantic enrichment (e.g., named entities, topic modeling, word embedding, text reuse). However, they do not address the downstream challenges of data transformation and analysis. By contrast, HistText caters primarily to professional researchers. It aims to equip them with a complete toolkit to efficiently transform digitized documents into data and gain full control over the datafication process within a single, integrated environment. This framework includes not only searching and corpus building, but also data extraction, consolidation, visualization, analysis, and beyond, scholarly interpretation, writing, and publication (see section 3.3). More specifically, HistText serves the needs of both cultural historians interested in analyzing concepts, discourses, and representations in vast corpora of historical texts over long periods of time, as well as social historians who need to retrieve biographical and social data from various text corpora to conduct prosopographical work and network analysis.

3. HistText is more flexible than cultural heritage interfaces. While MCTB initially focused on modern China and the study of elites, it can be expanded to include other textual materials (even not connected with China), provided they are available in plain text or digitized in high-quality resolution. Furthermore, HistText is not just a visual interface, but an extensive open-source library. Consequently, the algorithmic models and training data produced in the course of the project can be reused and developed further by other researchers to serve different research projects (see section 3.3.2).

The creation and development of MCTB-HistText has followed a bottom-up approach, driven primarily by ENP-China researchers' expanding needs. As the project unfolded, we gradually recognized that existing software did not adequately meet our needs and that *ad hoc* tools were necessary. This development has been possible because ENP-China relies on an interdisciplinary team. HistText is the result of a genuine, long-standing collaboration between historians and computer scientists specialized in NLP (see section 3.2). Such a collaboration requires a deep engagement with the alter disciplines on the part of both historians and computer scientists, which goes far beyond the passive reliance on

“engineer-servants”, as it is too often the case among self-proclaimed experts in “digital humanities”.

3. HistText and the Modern China Textual Database

3.1 The Modern China Textual Database

MCTD serves as one of the three interconnected pillars supporting research on the transformation of elites in modern China. Alongside the Modern China Biographical Database (MCBD) and the Modern China Geospatial Database (MCGD), MCTD plays a crucial role in facilitating the collection and analysis of historical information. While MCBD focuses on biographical data and MCGD on geospatial information, MCTD acts as the primary source for textual corpora, which serve as the foundation for extracting valuable historical insights and transforming them into usable data. With MCTD, the ENP-China project has designed the first textual database on Modern China.[28]

The Modern China Textual Database (MCTD) encompasses a diverse range of textual sources, including periodicals, directories, diaries, dictionaries, as well as full-text versions of Wikipedia and Baidu pages. These sources provide a wealth of historical information, with the majority being digitized versions of historical texts, while a few are born-digital resources. Among the periodicals included in MCTD are the Chinese daily newspaper *Shenbao* (1872-1949), the monthly *Eastern Miscellany* (1903-1949), and the extensive Proquest collection of English-language Chinese periodicals. The Proquest collection comprises a variety of daily, weekly, and monthly publications, including notable titles such as the *North China Herald* (1850-1949) and the *South China Morning Post* (1903-1997).

Furthermore, the ENP-China project has acquired journals such as the journal of the North China Branch of the Royal Asiatic Society (1863-1949), Chinese student journals published in the United States, Chinese Economic Bulletin, and a lot more. The collection also includes non-digitally born materials, such as Who's Who directories received from the Institute of Modern History (Academia Sinica), Annual Reports by foreign municipalities in Shanghai (in English and French), a substantial repository of China-related archives from the United States, and a collection of yearbooks (in English). With the exception of the Who's Who and directories, all materials have been digitized and processed using OCR (Optical Character Recognition). Although some journals are still awaiting processing at the time of writing, they will be added to MCTD in due course.

The second category within MCTD consists of digital-born materials, primarily comprising Wikipedia pages in both Chinese and English that feature biographies of individuals active in China during the designated time period. Additionally, there are plans to incorporate Baidu pages containing biographical information into the database.

[28] Charles Cooney, Glenn Roe, and Mark Olsen, “The Notion of the Textbase: Design and Use of Textbases in the Humanities,” in *Literary Studies in the Digital Age. An Evolving Anthology* (New York , N.Y.: Modern Language Association, 2013).

MCTB supports multiple languages, with a particular focus on Chinese, English, and French. While these languages are predominantly available in separate collections, the sources within MCTB also include bilingual publications, such as Who's Who directories, or documents that frequently cite Chinese terms, for instance, the Journal of the North China Branch of the Royal Asiatic Society, China Journal, and Yearbooks. MCTB facilitates data mining and analysis through various features. The texts within MCTB undergo pre-training to identify and index all named entities. Additionally, Chinese texts also benefit from pre-tokenization on a SolR server based on a robust and advanced model for transitional Chinese. These features allow researchers to create their own corpora and to control the parameters of their queries and further manipulations of the text. This is the major difference with consulting platforms and even with mixed-mode interface like the IHM collection of Who's Whos.[29]

MCTD represents the first comprehensive textual database dedicated to modern China, providing a valuable resource for scholarly research. With resources being open and freely accessible, excluding the commercially acquired Shenbao and Proquest collections, the database aims to foster collaboration and knowledge sharing. The text files within the database are indexed and stored on a SolR server, along with pre-computed tokens and named entities extracted from these texts. As an ongoing initiative, the database can continuously expand and evolve, welcoming spontaneous contributions from researchers. Overall, MCTB stands out among existing textbases due to its incorporation of diverse genres, multilingual resources, and advanced data mining capabilities, enhancing its utility for researchers and scholars seeking comprehensive and versatile textual analysis tools.

3.2 HistText: A Meeting of Minds

From History to Computing

As highlighted earlier (section 2.3), HistText represents the culmination of a longstanding and fruitful collaboration between historians and computer scientists that aimed at exploring machine learning in historical research. This symbiotic partnership has been instrumental in achieving optimized implementations, enhanced performance, and improved usability of HistText. The development of HistText benefited from historians' contributions at three pivotal levels:

(1) Providing text corpora and shaping research questions into computational tasks. Historians have played a pivotal role in providing valuable text corpora and articulating research questions that underpin the foundation of the HistText project. Through intensive discussions with computer scientists, these research questions were translated into well-defined computational tasks. This collaborative effort has ensured the alignment of computational methodologies with historical research goals, seeking optimal solutions encompassing implementation, performance, and usability.

(2) Testing and Feedback Incorporation. To ensure the practicality and relevance of HistText, historians actively participated in testing the tooling and offering constructive feedback. Workshops and hands-on sessions have been conducted

[29] http://mhdb.mh.sinica.edu.tw/

with colleagues and graduate students from diverse institutions, fostering an inclusive environment that empowers users outside the core team to contribute to the refinement and expansion of the framework. Noteworthy workshops and sessions have been held at esteemed institutions such as Paris-EHESS (December 2019), German Historical Institute in Washington, D.C. (May 2022), Summer School in Chinese Digital Humanities in Aix-en-Provence (June 2022), Leipzig University (November 2022), Institute of Modern History, Academia Sinica in Taipei (January 2023) in January 2023, and the Association of Asian Studies (AAS) conference in Boston (March 2023), reaffirming the widespread impact and relevance of HistText.

(3) Expertise-driven Annotation Campaigns. Historians' expertise is a valuable asset in the creation of ground truth and training data for advancing the capabilities of HistText. Their active involvement in annotation campaigns for tokenization, event detection, named entity recognition, and linking has ensured the generation of high-quality, validated data essential for training novel models.

HistText: A Development in Three Steps

- The development of HistText can be traced back to its inception by Pierre Magistry, currently an associate professor at Inalco, in 2019. Magistry laid the foundation for HistText with the creation of the R 'enpchina' library. This library offered essential functionalities for querying documents, retrieving full-text content, and extracting named entities from diverse corpora. Its primary objective was to empower historians by providing them with the ability to perform routine operations without the need to write code for each research endeavor. As historians gradually recognized the intricacies of their sources and the potential of programming languages, they engaged in discussions with computer scientists to enhance the 'enpchina' library and tailor it to their specific requirements for exploring digital corpora.

- Jeremy Auguste played a pivotal role in further refining the functionalities of HistText. He focused particularly on improving the 'extended' search and concordance features, enabling the introduction of filters to facilitate more precise narrowing down of results based on time, publications, fields, and other metadata. Additionally, Auguste spearheaded the development of the user interface in R-Shiny, designed to cater to non-programming users. During this process, new models for tokenization were also introduced. Baptiste Blouin, then a Ph.D. candidate, contributed to enhancing the named entity recognition (NER) capabilities for Chinese sources. As a culmination of these efforts, in February 2022, the decision was made to rename the 'enpchina' library as 'HistText'.

- Building upon this foundation, Baptiste Blouin further advanced HistText into a comprehensive application. Blouin made significant contributions to improving the R-Shiny interface, incorporating a diverse array of data visualizations that enhance the user experience. Importantly, behind the scenes, Blouin organized the implementation of several annotation campaigns focused on tokenization, named entity recognition, and event extraction in Chinese historical sources. The primary objective of these campaigns was twofold: to train models that yield more accurate results and to generate ground-truth datasets that serve as valuable resources for research purposes.

The interdisciplinary collaboration in HistText showcases the potential of combining historical expertise with computational methods, advancing research in both fields. The framework's continued evolution holds promising prospects for historical scholarship and computational linguistics.

3.3 HistText: Architecture and Key Features

HistText is based on the R programming language and presents two distinct work environments:

- a set of ready-made functions in an integrated environment (R Studio)
- a user-friendly R-Shiny interface for non programmers

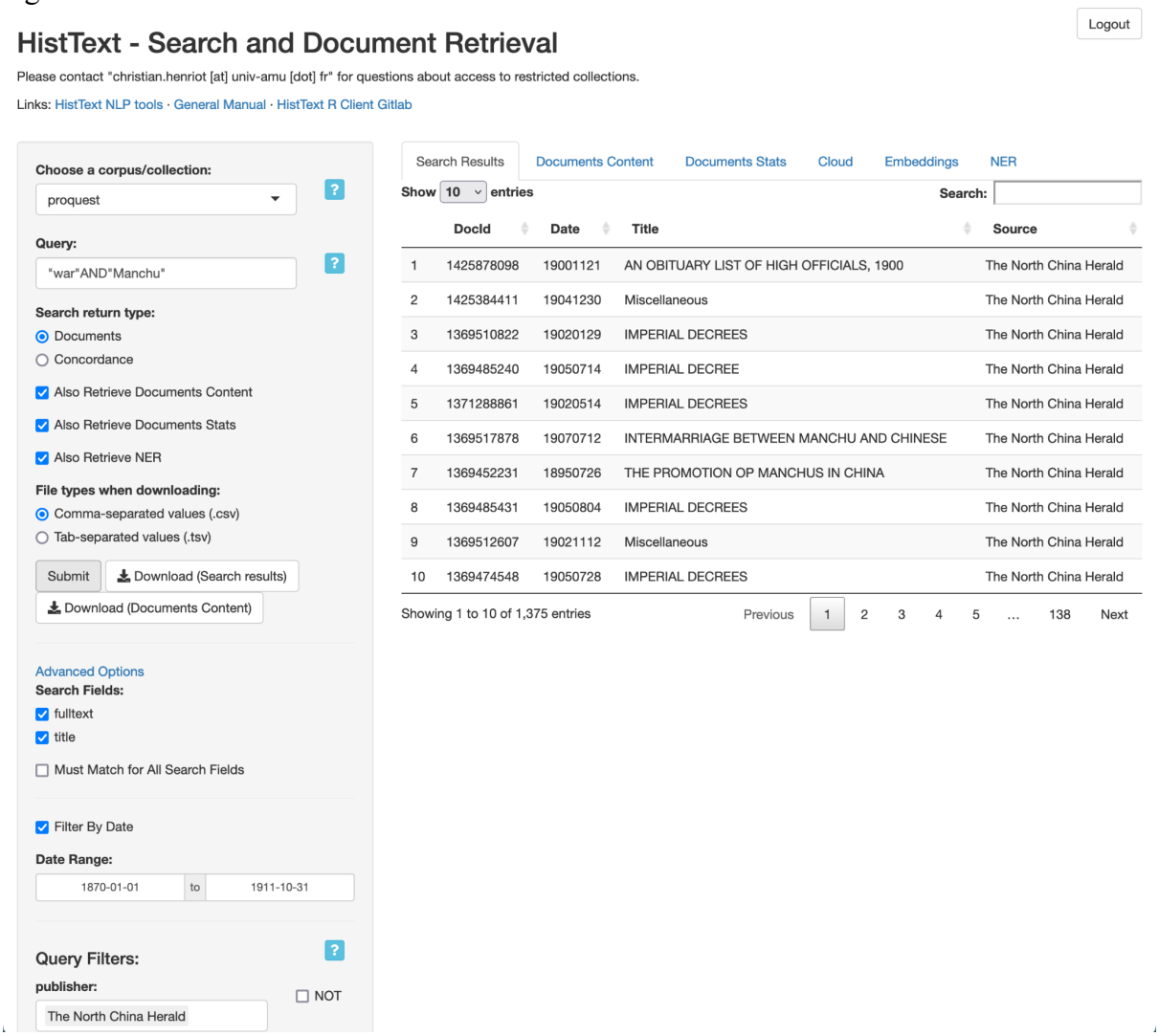
The combination of a HistText interface and a HistText R library is designed to smooth out the learning curve by researchers: the idea is to encourage them to gradually move from the user-friendly interface and its ready-made functions to the full appropriation of the extensive functions of HistText to gain full control over the process and unlock the full potential of their source corpora.

In the following section, we shall first introduce the R-Shiny application and its main features. This will serve as a gateway to discuss the far more extensive capabilities of the HistText library.

3.3.1 HistText as a user-friendly interface

The ENP-China project has developed a user-friendly interface on R-Shiny, allowing scholars without programming skills to utilize language models and machine learning for exploring large-scale corpora and extracting data. The R-Shiny interface provides similar functionalities as the HistText R library, albeit with limited options for manipulating text data. It consists of two main pages: one for querying and retrieving documents (Figure 1), and another for implementing named entity recognition.

Figure 1. HistText Search and Document Retrieval Interface



On the Search page, users can perform simple keyword searches or more advanced queries using standard Boolean operators across all collections. The interface incorporates a set of filters that enable users to define specific time periods, fields (such as text, title, article type), and publications they wish to explore. In the example above, the search targeted “war” and “Manchu” in the *North China Herald* only, and for the period 1870-1911. The query yielded 1,375 records. The available fields and publications may vary depending on the selected collection. For example, the Shenbao collection includes only the title, text, and date fields, while the Proquest English-language periodicals collection additionally distinguishes publications (source) and article types (advertisement, feature article, obituary, etc.). It should be noted that HistText's functionality is contingent upon how the digital versions' providers structured the full-text documents.

Query results on the Search page are displayed in a tabular format labeled as 'Search results', presenting a list of identified documents (ID, Title). If selected, the full-text content of the articles can also be accessed in the 'Documents Content' section. The queried terms are highlighted in red (Figure 2). Both the search results and the results with the full-text documents can be downloaded as a csv or a tsb file. In the example below, we queried the term “革命” (revolution) in the Shenbao and extracted the full text of the articles. The first

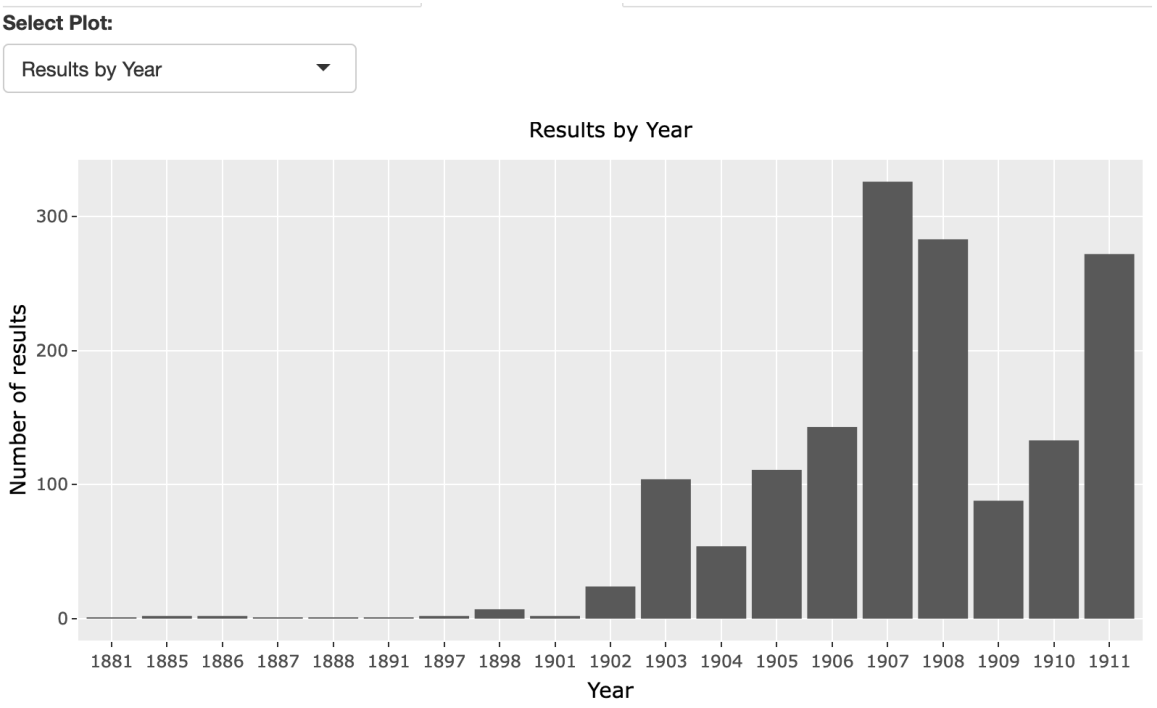
text column presents the original text of the article; the second text column (Text_chinese) presents the same text separated into tokens. The tokenization is done “on the fly” to allow researchers to use this version for further analysis.

Figure 2. Search Results with Documents ID, Date, Title, Full Text, and Tokenized Text

	DocId	Date	Title	Source	Text	Text_chinese
1	SPSP190709060506	19070906	革命黨異同考	shunpao	前日浙撫電飭搜查松江韓半池家。誣藏匿革命黨竺紹康。事夫以浙撫之勢力。欲誣韓以革命黨。則韓安得而不韓革命黨。固業醫者也。則其革命也。異乎人之革命。於是乎作革命異同考。	前日 浙撫 電飭 搜查 松江 韓半池家 誣 藏匿 革命黨 竺 紹康 事夫以 浙撫 之 勢力 欲 誣 韓以 革命黨 則韓安得 而 不韓 革命黨 固業 醫者 也 則其 革命 也 異乎 人 之 革命 於是乎 作 革命 異同 考
2	SPSP190607120406	19060712	俄京革命騷動	shunpao	十九日倫敦電云俄京益復不靖昨晚革命黨與哥薩克兵警察兵接戰甚烈傷者頗多革命黨列隊于街市手執紅旗高唱馬賽革命歌	十九日 倫敦 電云 俄京益 復 不靖 昨晚 革命黨 與 哥薩克 兵 警察 兵 接戰 甚烈 傷者 頗 多 革命黨 列隊 于 街 市 手執 紅旗 高唱 馬 賽 革命 歌

The 'Documents Stats' tab offers a scrolling menu with six distinct statistical perspectives on the dataset, each accompanied by different visualizations. In the example below, the graph shows the distribution of the number of mentions of “Wang Ching-wei” (Wang Jingwei) in the *China Weekly Review* between 1881 and 1911 (Figure 3).

Figure 3. Bart Chart of Results for Wang Jingwei in the *China Weekly Review (*1881-1911)



Furthermore, the 'Cloud' tab showcases a word cloud depicting the most frequently terms associated with the queried term. In the case of larger corpora, word embeddings have been calculated (for smaller corpora, a pre-defined set from Wikipedia is employed), which can be utilized to further refine queries by incorporating any terms linked to the word embeddings.[30] In the example below, we started the query from the term "革命" (Figure X.a), which came to be associated to several expressions in word embeddings. We selected the term "起義" (Uprising) to add to the query expression with OR (Figure 4).

[30] A *word embedding* is a learned representation for text where words that have the same *meaning* have a similar representation.

Figure 4.1 Word Embedding Enhanced Search Capability in HistText

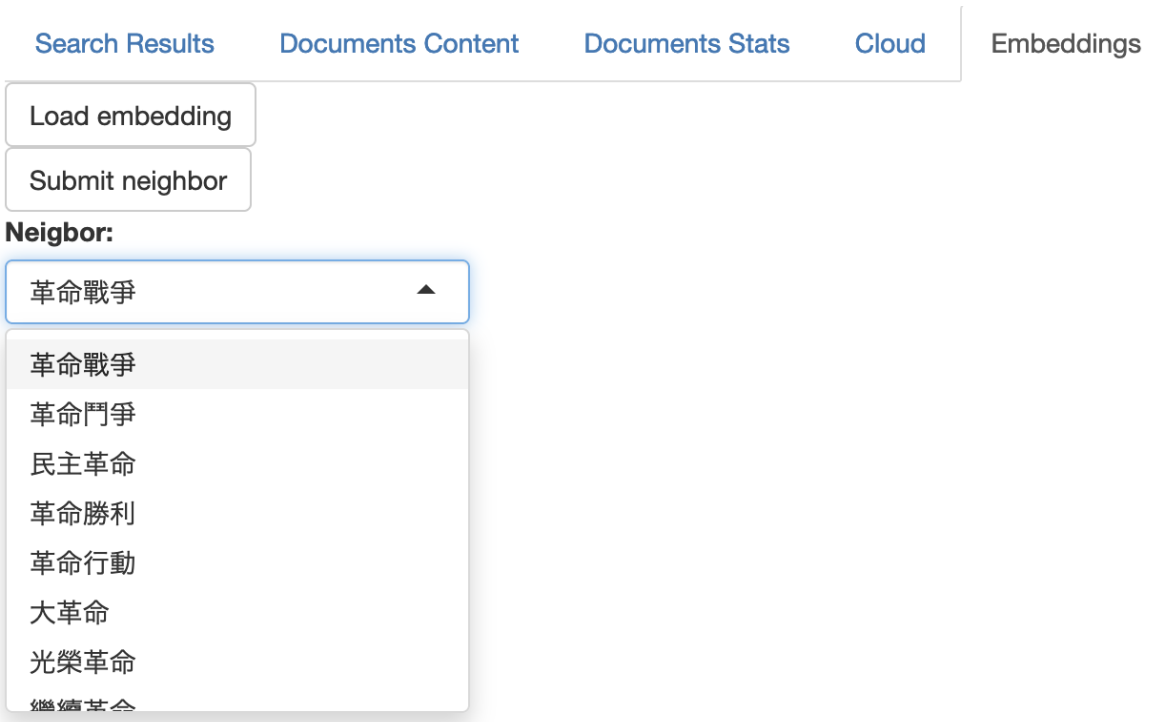
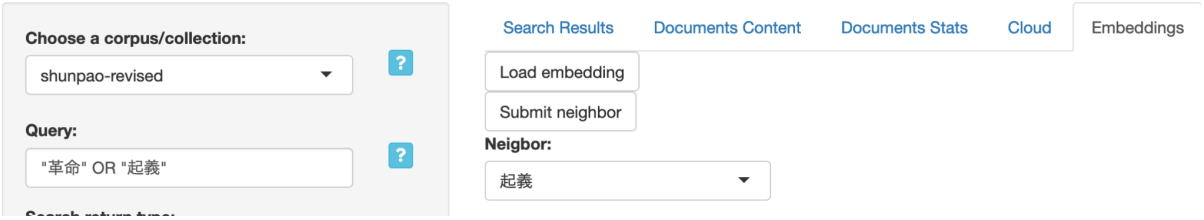
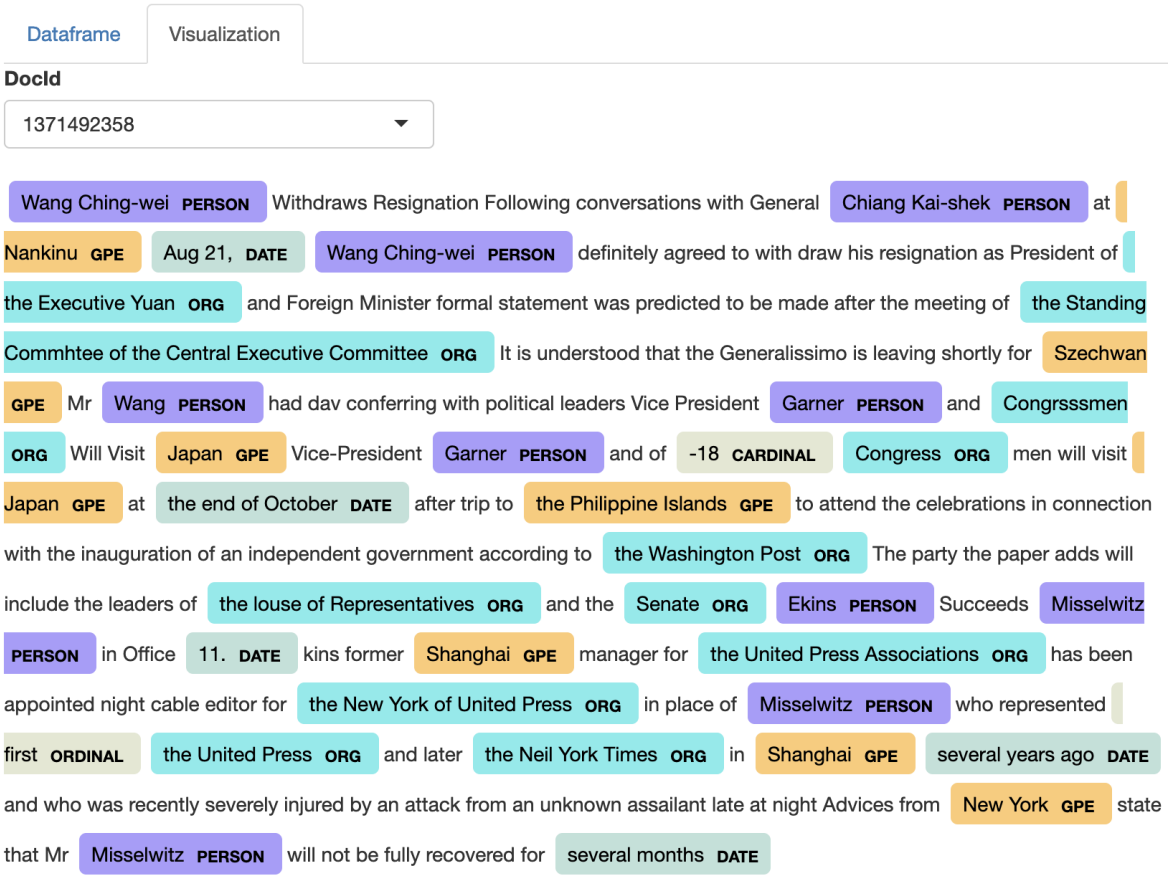


Figure 4.2 Word Embedding Enhanced Search Capability in HistText



Lastly, the Search page features a NER (named entity recognition) tab that allows users to gain insights into the named entities present in the first ten to one hundred documents of the dataset. These comprehensive visualizations assist scholars in contextualizing their produced dataset within the selected corpus, such as a specific periodical or a collection of books (Figure 5).

Figure 5. An Annotated Article with Named Entities in HistText



Alternatively, HistText offers the possibility to query terms in their context with the "concordance" search function. One can use the same filters as for a document search. The results are displayed with a snippet of the text before and after the queried term or expression. Researchers can define the size of the context (number of characters) depending on what they want to examine (Figure 6).

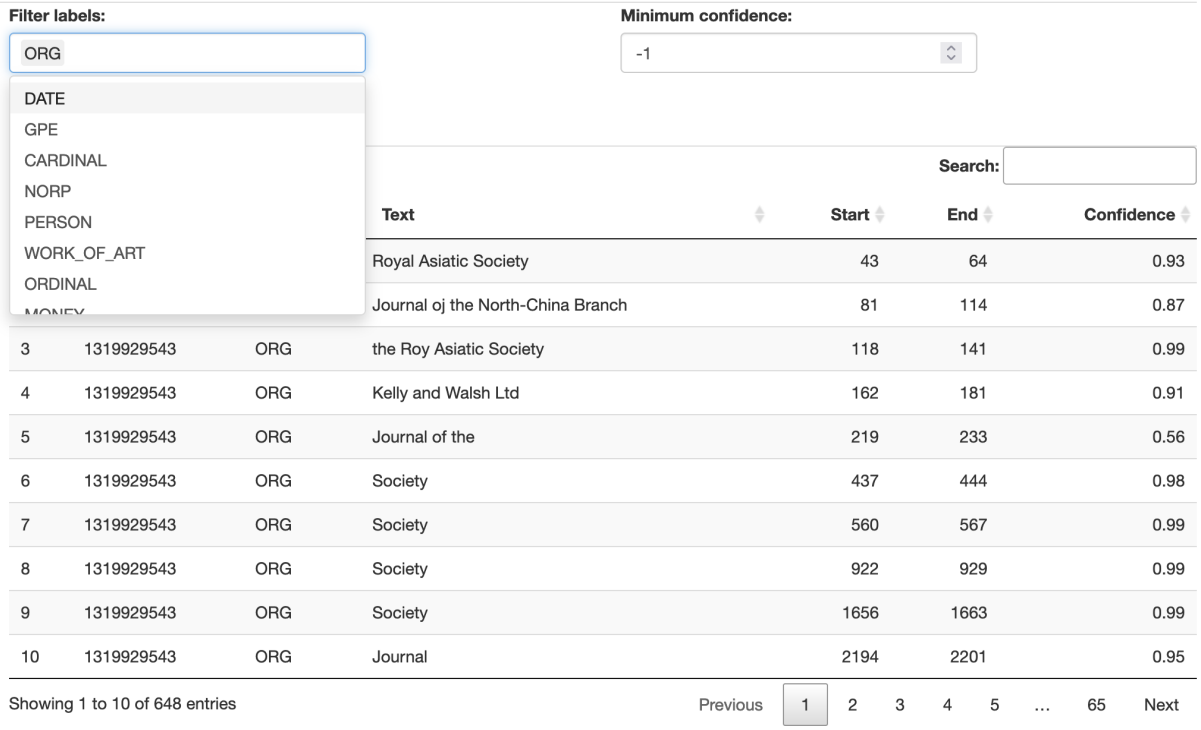
Figure 6. Concordance Search Results in HistText

	DocId	Date	Title	Source	Before	Matched	After
213	1369920598	18961016	THE DEATH OF M. GRIFFON	The North China Herald	nd acting French Vice-Consul Philippot merchant of Tientsin	Chollot	Engineer of Shanghai and Grevedon of the Customs these last
282	1369458889	18970528	Meetings	The North China Herald	circulation des bateaux indigenes et pour la sante publique	Chollot	Inglnieur de la Municipality Francaise presente sur cette q
283	1369458889	18970528	Meetings	The North China Herald	assages en Anglais Notre Conseil accepte les conclusions de	Chollot	comme il s d un travail qui interesse les deux Municipalit
284	1369458889	18970528	Meetings	The North China Herald	il est dispose contribuer au travail propose Le rapport de	Chollot	ete communi que au Capitaine du Port dont l opinion doit
285	1369458889	18970528	Meetings	The North China Herald	refore Mr Mayne will be instructed to consult with Monsieur	Chollot	and to ascertain from him and from personal inspection the
247	1369507733	18970709	THE SHANGHAI GENERAL CHAMBER OF COMMERCE	The North China Herald	from Bard was laid before the meeting enclosing letter from	Chollot	offering to submit to the Chamber plan for the improvement
248	1369507733	18970709	THE SHANGHAI GENERAL CHAMBER OF COMMERCE	The North China Herald	ld be very pleased to receive same on the terms proposed by	Chollot	After the transaction of other business Messrs Scott Gribbl
428	1369850272	18970723	READINGS FOR THE WEEK	The North China Herald	k at 1.30 on Wednesday -- It also learns with pleasure that	Chollot	Engineer of the French Municipality has been made Con- des
241	1369477525	18980114	LE BAL DES VOLONTAIRES ET DES POMPIERS	The North China Herald	of by theCommittee Messrs Tillot president Bottu secre tary	Chollot	Duval Gaillard Hdritte de Malherbe and Portier and carried
424	1369477354	18980321	READINGS FOR THE WEEK	The North China Herald	nd them copy of the engineer s specifications The report of	Chollot	on the extension to wards the river of the Place de l Est w

Showing 11 to 20 of 431 entries Previous 1 2 3 4 5 … 44 Next

The Natural Language Processing tools page permits users to upload their query results, specifically the documents with their full-text content, and apply named entity recognition to extract the named entities. The resulting data is presented in tabular format, displaying the Document ID, types of named entities, and their corresponding confidence indices. Users have the option to filter the results by selecting one or multiple types of named entities or by applying a confidence index filter (Figure 7). In the 'Visualization' tab, documents are displayed individually, with all annotated named entities presented by type using color codes as presented above. Users can select any document from the original dataset to examine the distribution of named entities in context, as opposed to relying solely on the tabular data in the 'Search results' tab. Lastly, HistText conducts statistical calculations on the results, which are visualized using various graphical representations.

Figure 7. HistText NER Results in Tabular Format with Named Entity Type Filt



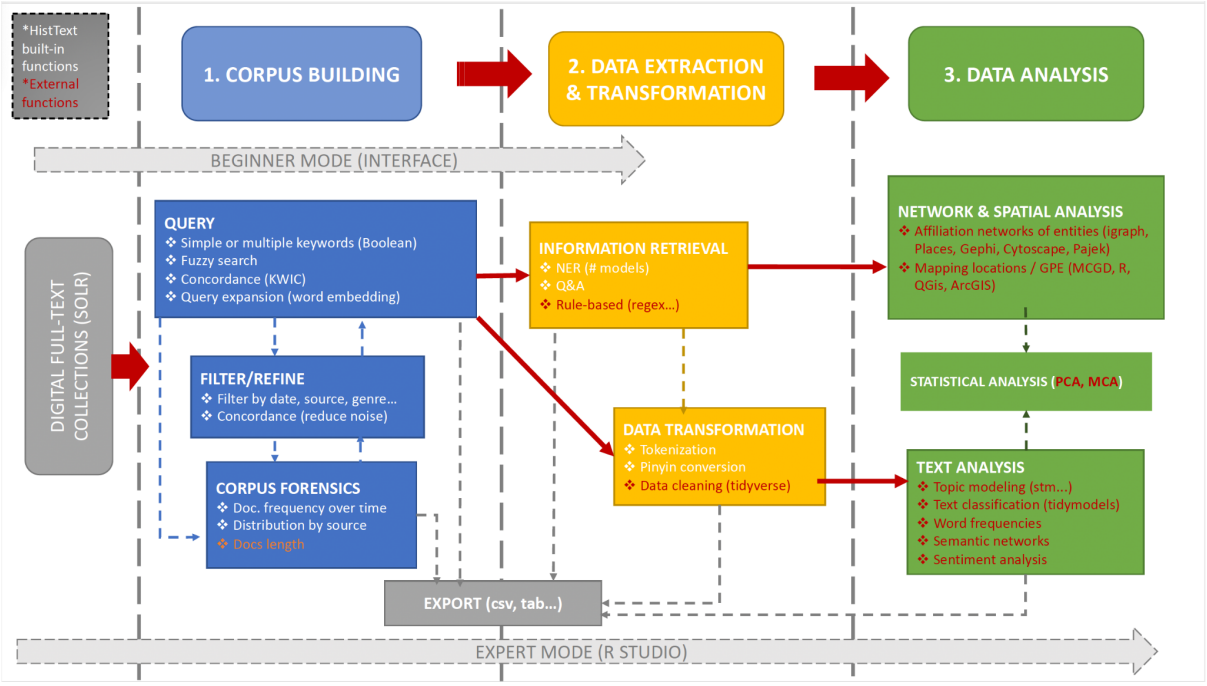
3.3.2 HistText as an R library

The HistText library offers a comprehensive set of functions — as of today 42 functions — within an R library, providing researchers with a range of advanced capabilities for efficient text analysis and extraction of information from historical documents. Compared to the public interface, the library enables greater control over the parameters, while it provides additional functions enabling more advanced operations of data processing.

The main functions of HistText serve to:

- build a customized corpus (query with advanced functionalities, including word embedding and concordance)
- explore metadata (plot distribution by year, etc) : this a key tool for refining the initial query and for digital source criticism/hermeneutics
  information extraction (NER)
- post-query analyses (topic model, text analysis)

Figure 8. The HistText Pipeline



This section highlights the key functionalities of the HistText library. The functions available in HistText can be grouped under six main sets of text analysis functions and one set of more technical functions. We shall not discuss the latter here that concerns server-related operations. The six main sets of functions are:

- Control functions
- Query functions
- Data extraction functions
- Advanced functions
- Chinese-specific functions
- Graph functions

Query functions

One of the primary functions of the HistText library is to facilitate the construction of a corpus based on specific queries. Researchers can utilize advanced search functions that go beyond simple keyword searches, enabling them to refine their queries and retrieve relevant textual data from the large-scale Modern China Textual Database. By leveraging these advanced search capabilities, researchers can build focused and targeted corpora tailored to their research questions.

The 'search_documents' function allows scholars to search for documents based on user-defined search terms, while the 'search_documents_ex' function provides additional options such as time ranges and specific fields for more refined searches. The search_concordance function and 'search_concordance_ex' perform Keyword in Context (KWIC) searches, extracting snippets of text that display the searched keyword or phrase within its contextual context. The 'get_documents' function retrieves full-text content of

selected documents based on document IDs, while the ‘count_documents’ function provides the number of documents that match a given query within a specified date range. Finally, the ‘view_document’ function enables researchers to directly view a single document within the RStudio environment by specifying the document ID. In addition to corpus construction, the HistText library offers functions for formatting textual data to make them compatible with computational algorithms. This process involves standardizing and preprocessing the text, ensuring that it can be effectively analyzed using various computational techniques. By preparing the textual data in a consistent and machine-readable format, researchers can apply computational algorithms and methods to extract insights and patterns from historical texts.

Data extraction functions

The HistText library includes functions for extracting named entities from the text. Named entities refer to specific named objects or entities such as people, places, organizations, and dates. By employing techniques such as named entity recognition, and using adapted language models for Chinese historical texts, the HistText library can automatically identify and extract these named entities from a wide range of historical sources. This feature significantly aids researchers in analyzing and understanding the roles, relationships, and occurrences of important entities within historical documents. The extracted data can be processed through other R libraries for data cleaning, building node and edge lists, topic modeling, etc. or used in third-party applications such as Cytoscape, Gephi, Orange, etc.

The data extraction functions in HistText offer researchers powerful capabilities for Named Entity Recognition (NER) analysis. The ‘ner_on_corpus’ function enables NER application on an entire corpus, automatically identifying and extracting named entities from the textual data, while the ‘ner_on_df’ function extends NER capabilities to specific columns within structured datasets. The run_ner function allows researchers to apply NER to individual text snippets, facilitating entity identification and analysis within specific contexts such as document titles, paragraphs, or sentences. These functions leverage various NER models and algorithms, including specific adaptations by the ENP-China team specifically tailored to historical texts (noisy OCR) and texts in Chinese (Chinese names). With these functions, researchers can effectively extract and analyze named entities, such as people, organizations, locations, and dates, from their textual data, supporting entity identification, entity linking, and entity-based analysis tasks.

Advanced functions

The implementation of question-and-answer (Q&A) queries is another valuable functionality provided by the HistText library. This feature enables researchers to target and extract specific content from natural-language texts based on user-defined queries. By formulating questions or prompts, researchers can use the Q&A feature to extract data from documents in natural language. The ‘qa_on_corpus’ function allows researchers to apply Question-Answering (QA) techniques to an entire corpus, automatically generating and retrieving answers to specific questions. This enables large-scale QA analysis and fine-tuning of the extraction process from historical documents. The ‘qa_on_df’ function extends QA capabilities to a specified column of a dataframe, allowing researchers to extract answers to predefined questions within their structured dataset. Additionally, the ‘extract_regexps_from_subcorpus’ function enables the application of regular expressions (regexps) to a collection of documents within a subcorpus, facilitating targeted data extraction, pattern matching, and efficient analysis.

Chinese specific functions

HistText includes a range of specialized functions for handling Chinese texts. Researchers can utilize the 'list_cws_models' function to explore available Chinese Word Segmentation (CWS) models and choose the most suitable one for their analysis. The 'run_cws' function applies CWS to a given string, producing segmented words as output. Researchers can also apply CWS on entire corpora using the 'cws_on_corpus' function, enabling more detailed linguistic analysis. The 'cws_on_df' function allows segmentation of Chinese text within specific dataframe columns, incorporating contextual variables. Moreover, the 'sinograms_to_py' function converts Chinese characters into pinyin, facilitating language learning, linguistic analysis, and text normalization.

In summary, the HistText library in R offers a comprehensive suite of functions designed to support various stages of text analysis in historical documents. It allows researchers to build targeted corpora, format textual data for computational analysis, extract named entities, and implement question-and-answer queries. These functionalities enhance the researcher's ability to leverage computational techniques for insightful analysis of historical texts. The presentation above does not do justice to the amazing possibilities that the functions of HistText provide, but interested readers can refer to our online manual and to the descriptive fiche of the 42 functions. The design of HistText required strong interdisciplinarity. It was built on long-term, continuous interactions between historians and computer scientists rooted on humanists/historians' needs and informed with the latest computational possibilities (machine learning, large-scale language models).

3.4 HistText and Digital Hermeneutics

HistText demonstrates genuine efforts to address the challenges of what some scholars have termed "digital hermeneutics".[31] The notion broadly refers to the set of issues posed by the transformation of historical documents into digital artifacts and the need for greater transparency regarding the process of digitization and datafication. Digital hermeneutics covers two main aspects: source criticism and tool criticism.

Source criticism addresses issues related to the creation of digital sources and data, including source provenance (metadata), representativeness, and data quality (i.e., considering possible biases introduced during the digitization and datafication process, most commonly OCR noise). It is important to acknowledge that we can only provide limited information regarding the collections which we bought from external providers (ProQuest, *Shenbao*). We face many black boxes concerning the workflow they followed, the history of the collections, the process of selection (how the documents were selected, from which sources), curation (metadata, what constitutes a document unit) and digitization (OLR, OCR, manual transformation). However, we can offer greater transparency regarding the collections which we have created ourselves (journals, diaries) and the enrichment we have made on the inherited collections (e.g., repunctuation and re-OCRization of ProQuest collections,

[31] Fickers, Andreas, Tatarinov, Juliane, and van der Heijden, Tim, "Digital History and Hermeneutics - between Theory and Practice: An Introduction," in *Digital History and Hermeneutics. Between Theory and Practice.* (Berlin/Boston: De Gruyter, 2022), 1–19.

re-segmentation of articles in *Shenbao*). These post-processing operations are fully documented on the ENP-China GitLab.[32] For the newly created collections, users can access the OCR performance scores. For the inherited collections, we are able to assess the expected scores based on the re-OCRed versions of the corpora.

Furthermore, HistText provides several functions to critically assess potential biases in document distribution and content *within* the collections, based on the available metadata. A notable example is the possibility to display the distribution of documents by collection, title, and category over time, and to relate document frequencies to the overall number of documents in the collections (see section 3.3.1). The various functions aimed at source criticism are grouped under the "corpus forensics" step in the HistText pipeline (Figure 8). These functions are actionable through both the interface and the package, though the package generally offers more options and increased transparency compared to the interface.

Tool criticism calls for a greater awareness from humanities researchers regarding the tool they use for searching, extracting, and analyzing data, and a critical assessment of the potential biases these tools may introduce in the final results.[33] HistText offers several options to address these challenges. Firstly, all functions are fully documented and illustrated in the dedicated manuals,[34] while their creation and development are extensively documented on GitLab, including the training/testing data, annotation guidelines, and other resources.[35] Interested researchers can also refer to the specialized publications by Jeremy Auguste, Baptiste Blouin, and Pierre Magistry in relevant NLP journals and conference proceedings.

HistText itself includes a set of functions aimed at empowering researchers and giving them greater control over the tooling. For example, both the interface and the library provide advanced search and filtering options, the possibility to customize the window size for concordance, to display the confidence scores for NER, and to rely on word embeddings to overcome the limitations of simple keyword search. It is important to emphasize that the interface offers limited possibilities compared to the package. In the R Studio environment, researchers gain full command of the parameters applied for each function. Notably, the library offers the possibility to select and compare different models for NER and tokenization, to customize the pre-processing of text data, to select the desired number of topics, and many other advanced functionalities. Ultimately, we believe that digital tool criticism is hardly possible without a minimal degree of engagement with programming languages and computational sciences.

[32] https://gitlab.com/enpchina

[33] Marijn Koolen, Jasmijn van Gorp, and Jacco van Ossenbruggen, "Toward a Model for Digital Tool Criticism: Reflection as Integrative Practice," *Digital Scholarship in the Humanities* 34, no. 2 (June 1, 2019): 368–85.

[34] https://bookdown.enpchina.eu/rpackage/HistTextRManual.html

[35] https://gitlab.com/enpchina (all) , https://gitlab.com/enpchina/histtext-r-client and https://gitlab.com/enpchina/histtext-server-api for histtext.

4. Case Studies and Application Scenarios

4.1 Case Study 1: Mapping Language Evolution in the modern Chinese press

The Chinese written language experienced a tremendous transformation from the near-classical language of the administration and imperial publications to the near-contemporary Chinese of the late 1940s. This transformation happened almost seamlessly in the pages of the modern press. One can even argue that the press, especially the newspapers, actually created the modern Chinese language, which incrementally seeped into other print materials. Historians who use historical sources from this period face what can be labelled "transitional Chinese", a language that evolved continuously from the beginning of the first newspaper in 1872 to 1949.[36]

There is abundant literature describing this pivotal era from different perspectives and disciplines related to language, including the history of language policies,[37] the socio-linguistic aspects[38] or historical linguistics.[39] However, there has been no study that leveraged a complete corpus of almost 80 years of a daily newspaper, the Shenbao (申報), containing about 750 millions sinograms to account for the actual language practices and their evolution through time. The work presented here relied on advanced NLP tools for data extraction to provide an unprecedented data-driven account of language practices at that time.

In this section, we examine a single case that highlights the power of NLP tools for language analysis over time. The core issue was to determine the path of language transformation and the major shifts in language practices. The experiment consisted in processing the whole corpus of the newspaper on a yearly basis with language modeling methods to run hierarchical clustering on the extracted data. The hierarchical clustering succeeded in spotting different periods that were internally homogeneous but distinct from each other.

The main idea for this experiment was to use perplexity as the basis to define a metric to apply hierarchical clustering of the different parts of the corpus. To have enough data to estimate a language model on each subcorpus, but still have relatively fine grain clusters, we chose to split the corpus into one sub-corpus per year. With one sub-corpus and one language model per year, it was now possible to use the perplexity of the language models to define a distance measure between every two years of the *Shenbao*. Once the distance matrix was built, we could apply multidimensional scaling (MDS) to visualize the data on a two-dimensional plane and agglomerative clustering to define periods over the whole corpus.

---

[36] This section is based on Pierre Magistry, "Le(s)? chinois du Shun-pao 申報," 2021, https://hal.science/hal-04059911.

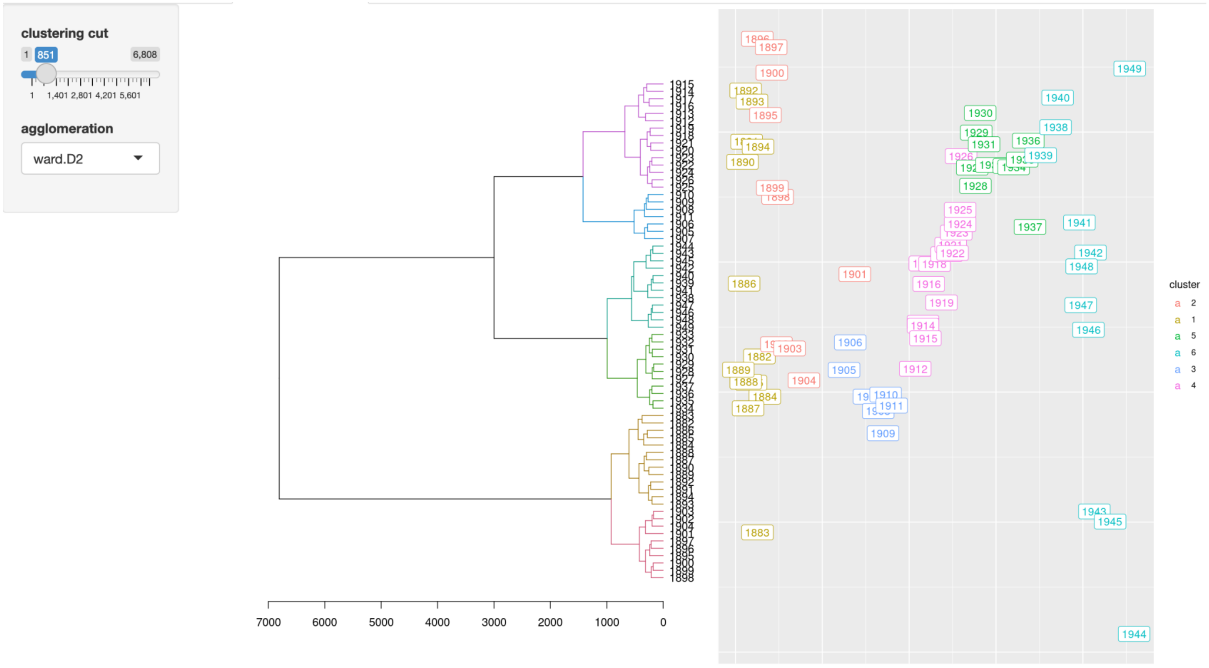
[37] Elisabeth Kaske, *The Politics of Language in Chinese Education, 1895-1919*, Sinica Leidensia (Leiden: Brill, 2008).

[38] Jeffrey Weng, "What Is Mandarin? The Social Project of Language Standardization in Early Republican China," *The Journal of Asian Studies* 77, no. 3 (2018): 611–33.

[39] W. South Coblin, "A Brief History of Mandarin," *Journal of the American Oriental Society* 120, no. 4 (2000): 537–52, https://doi.org/10.2307/606615; Richard Vanness Simmons, "Whence Came Mandarin? Qīng Guānhuà, the Běijīng Dialect, and the National Language Standard in Early Republican China," *Journal of the American Oriental Society* 137, no. 1 (2017): 63–88, https://doi.org/10.7817/jameroriesoci.137.1.0063.

The resulting plot of clustering dendrograms for the Ward is presented below. Yet the two clustering methods that we used tended to converge on relatively clear cuts after 1904, 1911 and 1937. The pre-1904 period is split either after 1894 or 1892 and another small disagreement occurs between 1921 and 1926 (Figure 9).

Figure 9. The Language Shifts of the *Shenbao*



4.2 Case Study 2: Building Social Networks from Historical Directories

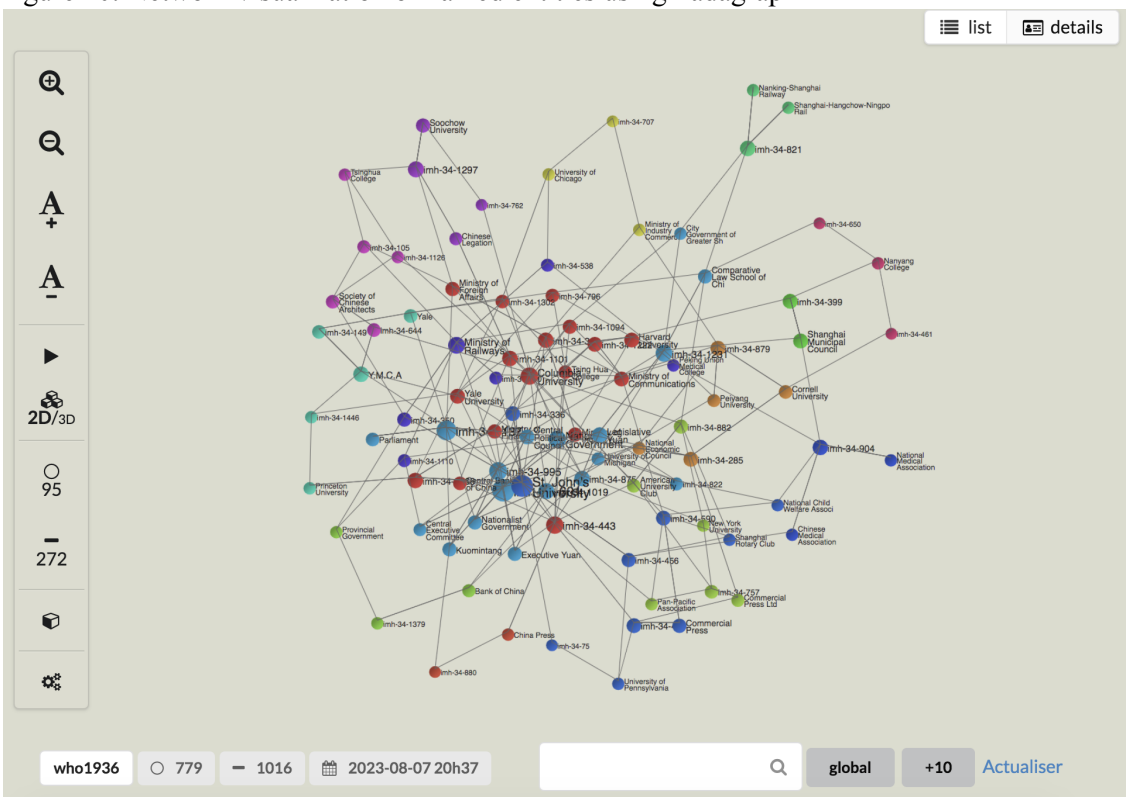
Context/motivation. This case study illustrates how researchers can use MCTB-HistText to gather biographical information on a group of people of varying size in order to reconstruct their career and analyze their networks of affiliations. It further illustrates two key powerful capabilities of HistText, namely, the capability to go beyond simple keywords and to search vectors of words and list of entities instead; and the possibility to combine HistText with existing R packages to effectively manipulate the data and conduct multidimensional analyses downstream.

Workflow. This research proceeds in four steps: (1) We used the HistText *search_documents* function to search a list of 418 individuals, specifically the 418 members of the American University Club (AUC) of China, in the IMH (Institute of Modern History) collection of *who's who* directories. Based on the results, we applied the *get_documents* function to retrieve the full text of the biographies. (2) We applied the function *ner_on_corpus* to extract the named entities from the biographies and we filtered the results to retain only the organizations. Optionally, researchers can utilize *Padagraph* to explore the relations between entities and documents in a graph form (Figure 10). (3) We relied on the *tidyverse* suite to clean and standardize the data. We occasionally referred to the HistText interface to visualize

the entities in their original context (see Figure 5 in section 3.3.1), which proved particularly helpful to manually complete missing data, correct noisy output and verify ambivalent entities. (4) Once we had a consolidated dataset of individuals and their affiliations, we called specific R libraries to conduct formal networks analysis and visualization. More specifically, we used *igraph, Places,* and *networkD3* to identify structural equivalences, detect communities (Figure 11), compare networks metrics, and to visualize strong ties and frequent flows (Figure 12).[40]

Results. This research has provided valuable insights on the formation of alumni networks in modern China, a topic which has been largely overlooked in previous studies of elites. A data-driven study reveals the crucial role these networks played in establishing the influence of American returned students in Chinese society during the Republican era, through recommendation practices among peers, and deliberate recruiting strategies in government institutions, particularly foreign affairs and the economic bureaucracy. Our findings further complicate the narrative of imperialism, highlighting the significance of interactions between Chinese and Americans through co-membership in elite clubs (Rotary), secret societies (Freemasonry), and service in mission-affiliated institutions (YMCA, St. John's University, St. Luke's Hospital). Overall, this research contributes to a more nuanced understanding of the complex relationships between China and the United States, through an exploration of their educational foundations and career extensions across national boundaries.The complete results be published in Cécile Armand, "Bonding minds, bridging nations: Sino-American alumni networks in the Era of Exclusion (1882-1936)", in Henriot (Ed.), *Modern China in Flux: Networks, Mobility, and Transformation*, Berlin, De Gruyter (scheduled for publication in 2024).

Figure 10. Network visualization of named entities using Padagraph



[40] The data and full code are available on GitHub: https://github.com/carmand03/american-university-men-china

Figure 11. Communities of American-educated Chinese from *Who's Who in China*, 1936.

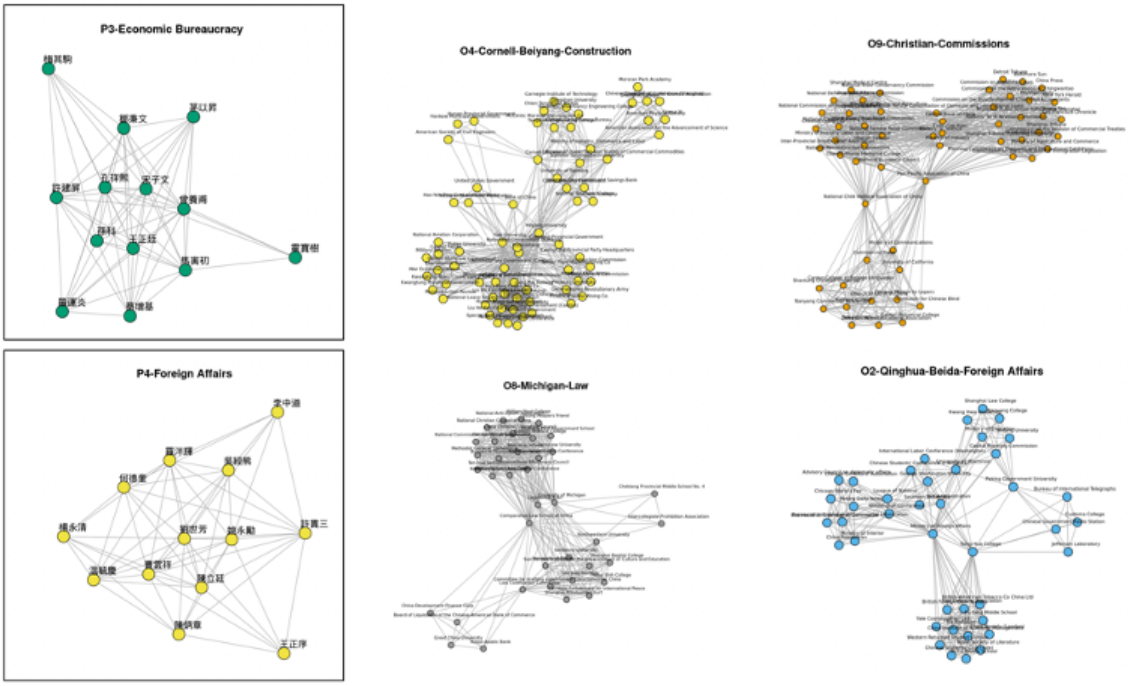
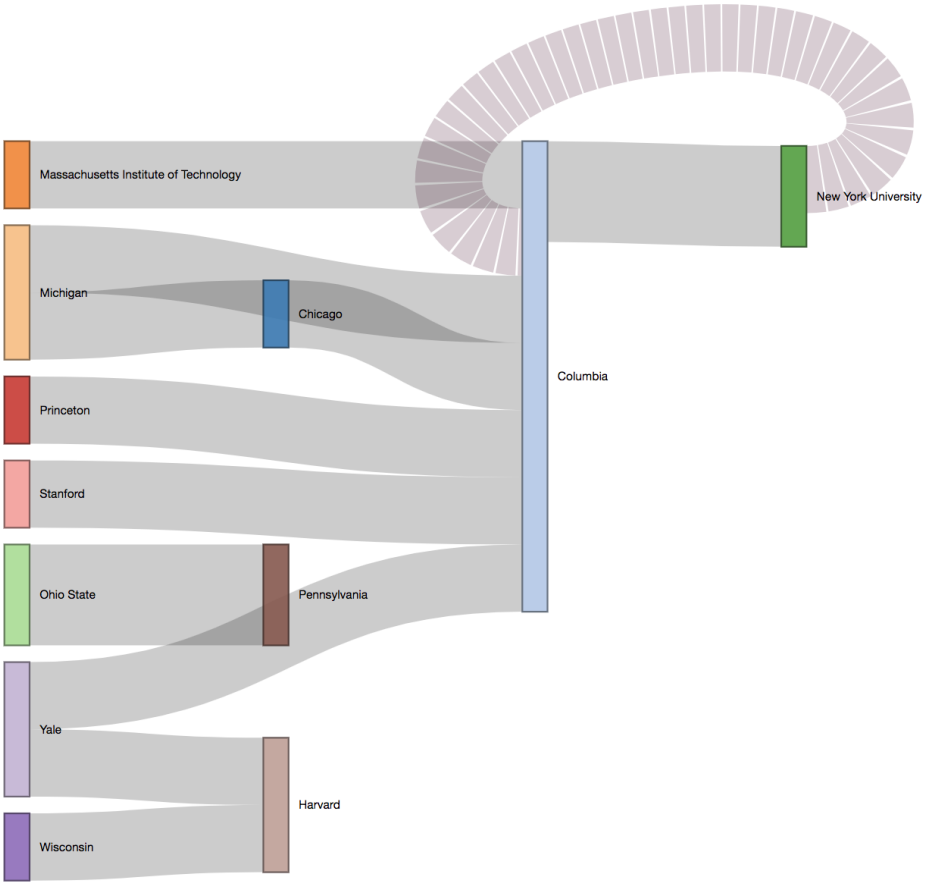


Figure 12. Most important flows of Chinese students in the United States (1883-1935)



4.3 Case Study 3: Topic modeling in Historical Newspapers

Context/motivation. This case study is an experiment in using different computational methods to explore a simple question: who were the individuals that appeared in the *Shenbao* in the first twenty years of its existence, especially who were the individuals mentioned repeatedly, whether and how they related to each other, to what institutions they were connected, and what they were involved in? I combined HistText with an array of R libraries to build a dataset, extract the data from the *Shenbao* daily and to conduct the analysis in three steps: statistical analysis, network analysis, and topic modeling. As an illustration, we shall only present the last step.

Workflow. My quest started with a search (*search_documents_ex*) based on two very common terms in any text in Classical Chinese: 之 and 也. This produced respectively 123,274 and 71,651 results (194,925). When boiled down to unique documents (emin_search1u), there remained 130,733 documents. We retrieved the full text for all the unique documents (*get_documents_ex*) and applied NER to extract named entities (*ner_on_corpus*). The results required further filtering: first, to filter out articles that were extracts from the Peking Gazette as well as advertisements by publishers; second, to filter strictly based on the length of the articles (articles with less than 500 characters) that seem in majority to contain individual articles. The final sample contained 87,997 articles. To establish the file for topic modeling, we built a new sample with articles that contained validated names to which we added the full text of related articles. All the operations were done with HistText, except data cleaning and file joins (Tidyverse). After removing the duplicates we obtained a sample with 50,565 documents. We processed all the articles with the tokenizer for transitional Chinese that we have been developing. The resulting file had 47,780 rows that contained the tokens for all the documents that served for topic modeling.

Results. This exploration eventually highlighted the major themes and topics that emerged from the analysis of the *Shenbao* in its first twenty years of existence. We implemented three different models (15-, 20-, and 30-topic models) to examine the impact of the models on the nature of the topics. The distribution of topics in three correlation models showed a consistent pattern of semantic proximity between certain topics, depending on the degree of granularity (Figure 13). The main themes were the issues of social order and justice, with several sub-themes or topics relating to this. The Mixed Courts, involving the local judicial system, were the most prevalent topic, covering issues like delinquency, petty crimes, and commercial disputes (Figure 14). Topic modelling suggests that the prevalence of social disorder topics could reflect the instability of local society after Shanghai opened to foreign trade. However, it might more likely reflect how the newspaper gathered newsworthy information. Besides, several other loosely connected topics were covered, including issues involving local Chinese officials, the Chinese Army, shipping and consular matters, international affairs, and literary texts. There were a few changes over time in the relative importance of topics (see Topic 10 [Xian officials] and 12 [Chinese Army] in Figure 15) but overall, the nature of news the *Shenbao* chose to publish was relatively stable (Figure 15). The topics that emerged provide insight into how the Shenbao operated and the process of news-making during its first twenty years. The *Shenbao* had to rely on existing cultural and social forces, including the highly educated literati who had not made it into the imperial bureaucracy. This exploration led us to conclude that the *Shenbao*, though a major source for exploring Shanghai's social history, had its biases. It was crafted by and for a particular

segment of the elites, with the newspaper serving as a conduit for the literati to narrate the novelty and uncertainties of urban life in Shanghai and secondarily in the surrounding towns and cities. The full study is to be published in Christian Henriot, "Eminent Chinese of the Shenbao (1872-1891). A digital investigation of news reporting and newspaper-making in late imperial China", *Journal of Digital History* (scheduled for publication in 2024).

Figure 13. Inter Topic Map and Most Prevalent Terms (topic 7)

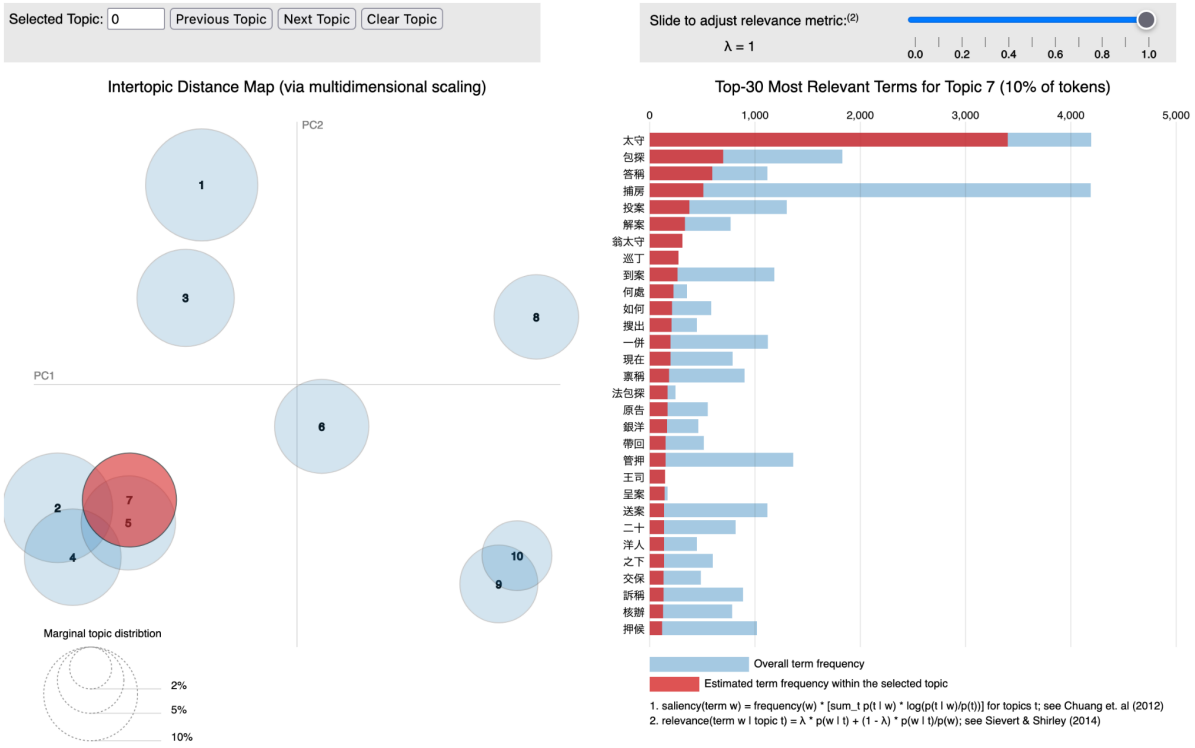


Figure 14. Correlation graph of the 15 topics in the 15-topic model

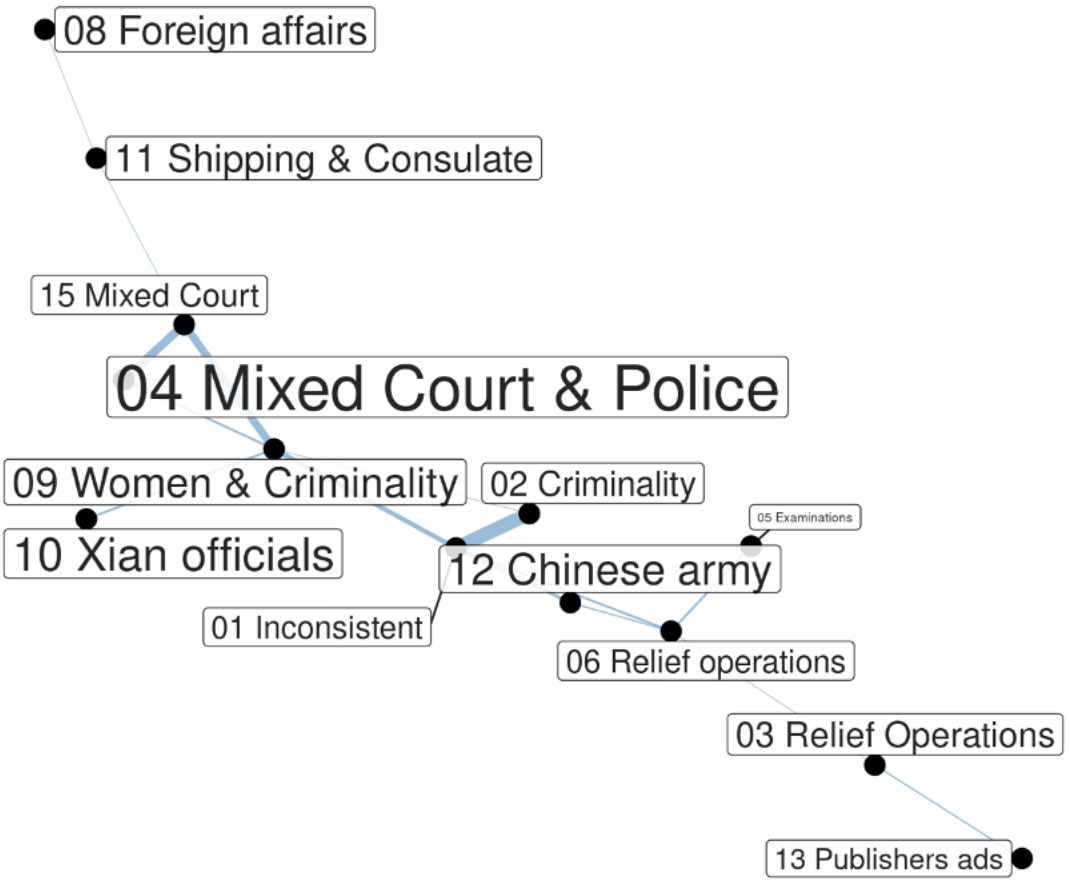
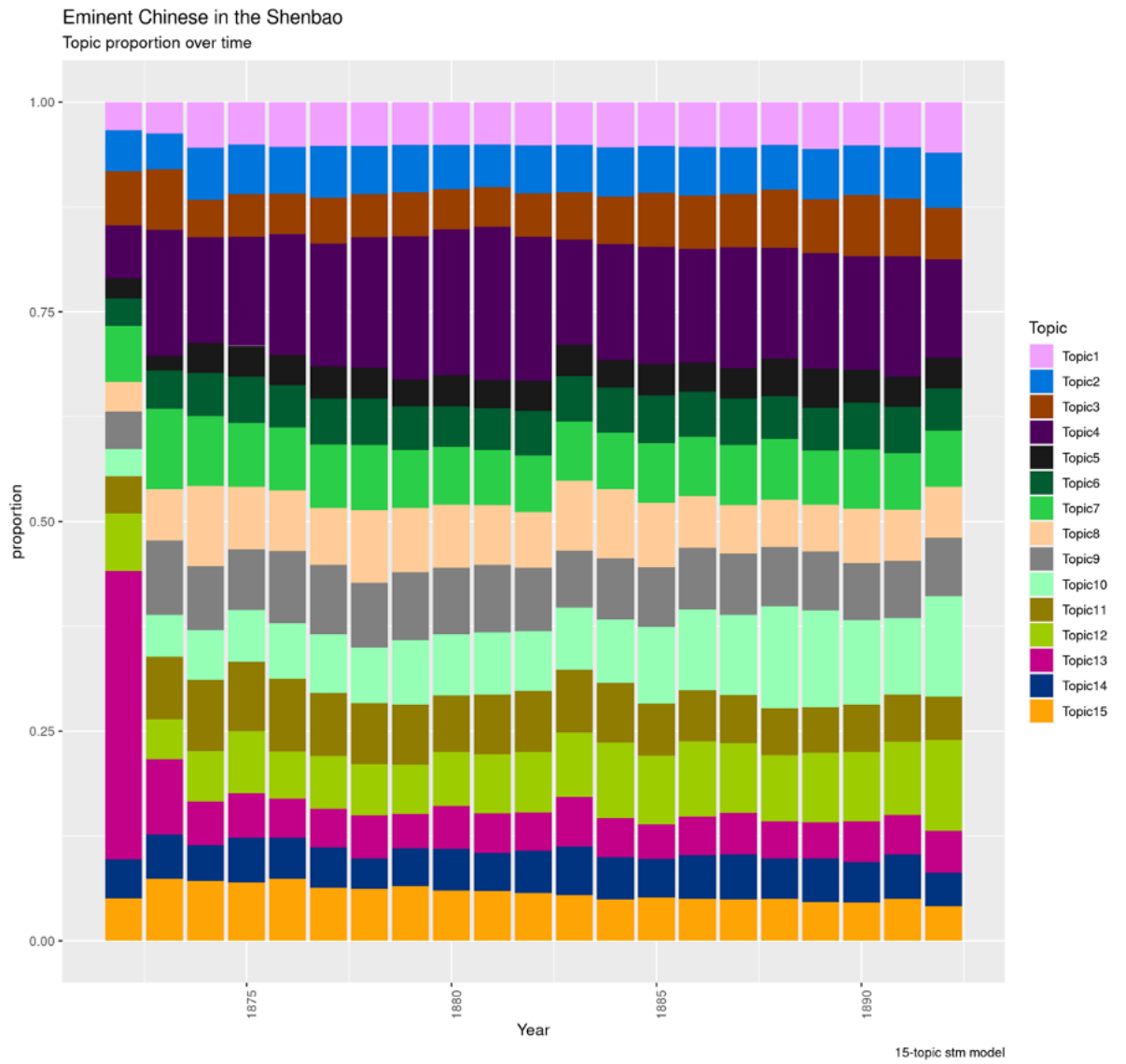


Figure 15. Topic proportion over time (1872-1892)



5. Discussion

HistText makes a significant contribution to the field of computational humanities by addressing the challenges associated with data mining in large-scale historical digital corpora, particularly focusing on Chinese sources. With the Modern China Textual Database containing an enormous volume of documents and words, the need for efficient extraction of insights from such vast repositories had become paramount. HistText offers a unique and innovative solution through its user-friendly interface, advanced text analysis techniques, and powerful data visualization capabilities.

By streamlining the process of data mining, HistText enables digital humanists to delve into historical texts more effectively and uncover valuable information buried in the vast corpus. The application's ability to handle the complexities of historical language, including archaic

terms, variant spellings, graphic variants, and diverse writing styles, greatly enhances researchers' capacity to explore and analyze historical documents in a comprehensive and systematic manner. The capabilities offered by HistText unlock the path to implement sophisticated methods such as network analysis, topic modeling or sentiment analysis. The application's data visualization features facilitate the identification of patterns, trends, and connections within the corpus, thereby enabling researchers to formulate new research questions, generate hypotheses, and gain a deeper understanding of historical contexts.

By providing a user-friendly interface, HistText lowers the entry barrier for researchers, allowing even those without extensive computational skills to utilize advanced text analysis techniques effectively. It is our contention that such accessibility will fosters interdisciplinary collaborations and encourage scholars from various fields to engage with historical texts using computational methods.

HistText, despite its significant contributions and potential, has certain limitations and expectations that should be taken into consideration. One of the limitations of HistText is its focus on full-text analysis, excluding the incorporation of images. While textual analysis provides valuable insights, historical documents often contain visual elements that can enhance the understanding of the context and meaning. Incorporating image analysis capabilities would broaden the application's scope and enable researchers to extract insights from both textual and visual components of historical documents.

Another challenge for HistText lies in the issues of access and copyright. Many historical documents are subject to copyright restrictions, making them inaccessible for analysis or limiting the availability of certain text collections. Overcoming these barriers requires collaborations with libraries, archives, and other institutions to ensure legal access to the texts and compliance with copyright regulations. Moreover, efforts should be made to establish partnerships with organizations that specialize in digitizing and preserving historical documents, allowing HistText to expand its collection and improve access for researchers.

A central concern and expectation by the ENP-China project is the transferability of HistText to other texts and communities of researchers beyond the study of modern China. While HistText's primary focus is on Chinese historical texts or China-related English-languages sources, its underlying methodologies can be adapted to analyze texts from different regions and time periods. By expanding its language support and incorporating diverse corpora, HistText can serve as a versatile tool for researchers working with historical documents from various cultural and linguistic backgrounds. HistText is fully open source and available on GitLab.

Furthermore, building a multidisciplinary community of users is crucial for the growth and development of HistText. Historians, computational linguists, literary scholars, and researchers from other disciplines can all benefit from the application's capabilities. We believe that creating forums, workshops, and collaborative platforms that encourage knowledge sharing and interdisciplinary exchanges will foster a vibrant community of users who can contribute to the further refinement and enhancement of HistText.

Looking ahead, there are several promising avenues for the future development and enhancement of HistText. One potential direction is the integration of more advanced machine learning algorithms and natural language processing techniques to further improve the accuracy and efficiency of text analysis. By harnessing the power of the more recent

large-scale models, HistText can expand its capabilities to handle complex linguistic patterns, automate the extraction of information, and facilitate the identification of meaningful patterns within historical texts. Furthermore, incorporating additional languages and expanding the scope beyond Chinese sources would broaden the applicability and reach of HistText, enabling comparative studies and cross-cultural analysis.

In conclusion, HistText represents a significant advancement in the field of computational humanities, particularly in the domain of data mining in historical documents, with a focus on Chinese sources. By streamlining the process of extracting insights from vast repositories, HistText offers a user-friendly interface, advanced text analysis techniques, and powerful data visualization capabilities. Moving forward, further development and enhancements, such as integrating advanced machine learning algorithms, expanding language support, and refining analytical techniques, hold promise for the continued growth and impact of HistText. The application's potential for facilitating interdisciplinary research and fostering collaborations across different domains underscores its importance as a valuable tool for digital humanists and researchers working with historical texts.

References

Chiron, Guillaume, Antoine Doucet, Mickaël Coustaty, Muriel Visani, and Jean-Philippe Moreux. "Impact of OCR Errors on the Use of Digital Libraries: Towards a Better Access to Information." In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 1–4. IEEE, 2017.

Coblin, W. South. "A Brief History of Mandarin." *Journal of the American Oriental Society* 120, no. 4 (2000): 537–52. https://doi.org/10.2307/606615.

Cooney, Charles, Glenn Roe, and Mark Olsen. "The Notion of the Textbase: Design and Use of Textbases in the Humanities." In *Literary Studies in the Digital Age. An Evolving Anthology*. New York , N.Y.: Modern Language Association, 2013.

Dougherty, Kristen Nawrotzki; Jack. *Writing History in the Digital Age*, 2013. http://hdl.handle.net/2027/spo.12230987.0001.001.

Galgano, Michael J., J. Chris Arndt, and Raymond M. Hyser. *Doing History: Research and Writing in the Digital Age*. 1st ed. Boston, MA: Thomson Wadsworth, 2008.

Kaske, Elisabeth. *The Politics of Language in Chinese Education, 1895-1919*. Sinica Leidensia. Leiden: Brill, 2008.

Magistry, Pierre. "Le(s)? chinois du Shun-pao 申報," 2021. https://hal.science/hal-04059911.

Milligan, Ian. *History in the Age of Abundance?: How the Web Is Transforming Historical Research*. Montreal: McGill-Queen's University Press, 2019. https://0-ebookcentral-proquest-com.cataleg.uoc.edu/lib/bibliouocsp-ebooks/detail.action?docID=5732673.

Simmons, Richard Vanness. "Whence Came Mandarin? Qīng Guānhuà, the Běijīng Dialect, and the National Language Standard in Early Republican China." *Journal of the American Oriental Society* 137, no. 1 (2017): 63–88. https://doi.org/10.7817/jameroriesoci.137.1.0063.

"The Promise of Digital History." *The Journal of American History* 95, no. 2 (September 2008). http://www.journalofamericanhistory.org/issues/952/interchange/.

Tsu, Jing. *Kingdom of Characters: The Language Revolution That Made China Modern*. New York: Riverhead Books, 2022.

Weller, Toni. *History in the Digital Age*. London; New York: Routledge, 2013.

Weng, Jeffrey. "What Is Mandarin? The Social Project of Language Standardization in Early

Republican China." *The Journal of Asian Studies* 77, no. 3 (2018): 611–33.

Resources

We have released the public version of the code on Gitlab. The current version of HistText 1.XX comes with a complete HistText Manual that we have prepared to describe all the functions and to provide ready-made examples of scripts. The HistText online interface has two access pages: one for the Search and Query functions, one for Named Entity Extraction.The functions of the interface are fully described in the HistText User Guide.

Histext Functions

Function names	Extended description
Control Functions	
accepts_date_queries	This function allows researchers to check if a particular corpus accepts date queries. Date queries are essential for conducting temporal analyses and investigating historical events within specific timeframes. By utilizing the accepts_date_queries function, researchers can determine whether a corpus supports the querying of historical documents based on specific dates or date ranges. This information enables researchers to tailor their analyses and retrieve relevant textual data corresponding to their desired temporal context.
get_default_ner_model	The get_default_ner_model function is an essential control function provided by the HistText application. Named Entity Recognition (NER) plays a significant role in extracting and identifying named entities, such as people, organizations, locations, and dates, from textual data. This control function allows researchers to retrieve the name of the default NER model associated with a given corpus. By accessing this information, researchers can ensure that they are using the appropriate NER model for their specific corpus, enhancing the accuracy and effectiveness of their named entity extraction tasks.

list_corpora	The list_corpora control function provides researchers with an overview of the available collections in SolR, the search platform used by HistText. This function enables researchers to retrieve a comprehensive list of the corpora or text collections accessible within the application's database. By utilizing the list_corpora function, researchers can quickly identify and explore the diverse range of text collections available for analysis.
Query Functions	
search_documents	The search_documents function is a fundamental query function that serves to search for specific documents within a corpus based on user-defined search terms. This is the most basic, yet essential tool to build a dataset for analysis. By inputting relevant keywords or phrases, researchers can retrieve documents that contain the specified search terms. Researchers to narrow down their focus and extract relevant textual data by using the common Boolean operators.
search_documents_ex	The search_documents_ex function extends the capabilities of the basic search_documents function within the HistText application. It provides additional options such as defining a time range, targeting specific fields, etc. to conduct more refined document searches. By leveraging the search_documents_ex function, researchers can fine-tune their search queries and retrieve more precise and targeted results.
search_concordance	The search_concordance function in the HistText application enables researchers to perform Keyword in Context (KWIC) searches within textual corpora. This function extracts concordances or snippets of text that display the searched keyword or phrase within its contextual context. Researchers can adjust the size of the surrounding text to gain a better understanding of the context in which it appears.

search_concordance_ex	Similar to the search_documents_ex function, the search_concordance_ex function expands upon the basic KWIC search functionality provided by the search_concordance function. It offers extended capabilities for researchers to refine their KWIC searches. This extended KWIC search function allows researchers to incorporate advanced search parameters, such as defining a time range, targeting specific fields, etc.
search_concordance_on_df	The search_concordance_on_df function in the HistText application enables researchers to perform KWIC searches on a custom dataframe of their choice. This function allows researchers to apply the KWIC search technique to a dataset that they have already prepared and imported into R. By specifying the target dataframe and the desired search keyword or phrase, researchers can extract relevant concordances from their custom dataset. This functionality is particularly useful when researchers have curated a specific corpus or dataset outside of the standard HistText corpora.
get_documents	The get_documents function allows researchers to retrieve the documents selected through the search_documents function with their full-text content based on the document IDs. Researchers can simply input the list of document IDs of interest, and the function will fetch the corresponding documents
count_documents	The count_documents function serves as a helpful query function within the HistText application, providing researchers with the ability to determine the number of articles or documents that match a given query, based on date criteria. By specifying the query parameters and a specific date range, researchers can obtain an accurate count of the documents that fall within the defined query criteria.
count_search_documents	The count_search_documents function allows researchers to determine the number of documents that can be returned by a particular query without retrieving the actual documents. This function aids researchers in understanding the potential size and scale of their query results, enabling them to gauge the feasibility and magnitude of their research endeavors before executing resource-intensive queries.

view_document	The view_document function offers researchers the ability to view a single document directly within the RStudio environment. By specifying the document ID, researchers can display the targeted document in the R-Studio environment. This function will work only if there is a direct ID correspondence between the document ID on the SolR server and the document ID on the online collection. We have tested this function successfully with the Proquest collection.
Data extraction functions	
ner_on_corpus	The ner_on_corpus function enables researchers to apply Named Entity Recognition (NER) on an entire corpus. By utilizing this function, researchers can automatically identify and extract named entities, such as people, organizations, locations, and dates, from the textual data within a corpus. The ner_on_corpus function utilizes various NER models and algorithms, especially for Chinese, to analyze the corpus. The ENP-China team carried out several annotation campaigns on Chinese press and biliangual directory corpora to adapt and improve existing models (especially Ontonote).
ner_on_df	The ner_on_df function extends the NER capabilities to a specified column of a dataframe. Researchers can apply NER to a specific column within their structured dataset, such as a column containing textual data. This function allows for targeted NER analysis within a specific context or domain of interest.
run_ner	The run_ner function provides researchers with the ability to apply Named Entity Recognition (NER) to a single string or text snippet. This function is particularly useful when researchers want to analyze individual pieces of text, such as a document title, a paragraph, or even a sentence. By applying the run_ner function to a specific string, researchers can extract named entities within that particular context, aiding in tasks such as entity identification, entity linking, or entity-based analysis.

run_qa	The run_qa function allows researchers to apply Question-Answering (QA) techniques to a given string or text snippet. This function utilizes advanced natural language processing algorithms and models to identify and extract relevant answers to user-defined questions from the provided text. By leveraging the run_qa function, researchers can obtain specific information or insights from textual data, enabling them to gain a deeper understanding of the content and extract valuable knowledge from historical documents.
qa_on_corpus	The qa_on_corpus function empowers researchers to apply Question-Answering (QA) techniques on an entire corpus. By utilizing this function, researchers can automatically generate and retrieve answers to specific questions from the textual data within a corpus. This functionality enables researchers to perform large-scale QA analysis, allowing them to fine-tune the extraction process from historical documents in natural language.
qa_on_df	The qa_on_df function extends the Question-Answering (QA) capabilities to a specified column of a dataframe. Researchers can apply QA techniques to a specific column within their structured dataset, extracting answers to predefined questions.
extract_regexps_from_subcorpus	The extract_regexps_from_subcorpus function allows researchers to apply a collection of regular expressions (regexps) to a collection of documents within a subcorpus. Regular expressions provide a powerful tool for pattern matching and extraction in textual data. This process enables researchers to extract specific information, identify patterns, or perform targeted data extraction based on predefined regexps, facilitating efficient data extraction and analysis tasks.
Advanced functions	

list_search_fields	The list_search_fields function within the HistText application allows researchers to retrieve a comprehensive list of possible search fields for a given corpus. By running this function, researchers can access the metadata associated with the corpus and identify the available fields that can be used for conducting searches. This information assists researchers in understanding the structure and organization of the corpus, enabling them to formulate more targeted and specific search queries.
get_search_fields_content	The get_search_fields_content function provides researchers with the ability to retrieve the content associated with each search field within a corpus. By using this function, researchers can access the actual data or information contained within each search field.
list_filter_fields	The list_filter_fields function serves as a valuable tool for researchers to obtain a list of possible filter fields specific to a given corpus. Filters are used to refine and narrow down search results based on specific criteria or attributes. Researchers can identify the available filter options associated with a corpus, enabling them to effectively apply filters and retrieve more precise and relevant search results.
list_ner_models	The list_ner_models function provides researchers with a comprehensive list of available Named Entity Recognition (NER) models hosted on the server. NER models are crucial for identifying and extracting named entities from textual data. By running the list_ner_models function, researchers can access information about the various NER models supported by the HistText application. This functionality allows researchers to choose the most appropriate NER model for their specific corpus or analysis needs.
list_possible_filters	The list_possible_filters function allows researchers to retrieve a list of possible filter values for a given filter field within a corpus. Filters provide researchers with a means to narrow down search results based on specific criteria or attributes. By utilizing the list_possible_filters function, researchers can explore the available options and values associated with a particular filter field.

list_precomputed_corpora	The list_precomputed_corpora function provides researchers with a list of corpora that have precomputed annotations available. Precomputed annotations refer to the process of performing certain analyses or annotations on a corpus in advance, making them readily available for researchers to utilize. By running the list_precomputed_corpora function, researchers can identify the corpora that already have precomputed annotations, saving them time and computational resources in performing certain analyses or annotations.
list_precomputed_fields	The list_precomputed_fields function allows researchers to obtain a list of fields within a given corpus that have precomputed annotations. Precomputed annotations are valuable resources that provide researchers with enriched information and insights about the corpus. By utilizing the list_precomputed_fields function, researchers can identify the specific fields within a corpus that have precomputed annotations, enabling them to leverage these precomputed data for their analyses or research tasks.
list_qa_models	The list_qa_models function provides researchers with a comprehensive list of available Question-Answering (QA) models hosted on the server. QA models are designed to generate answers to specific questions based on textual data. By running the list_qa_models function, researchers can access information about the different QA models supported by the HistText application.
load_pdf_as_df	The load_pdf_as_df function is a powerful tool that allows researchers to extract and load the text from a PDF document into a structured data frame. PDF documents often contain valuable textual information relevant to research and analysis. By using the load_pdf_as_df function, researchers can convert the content of PDFs into a structured format that is more amenable to computational analysis. This function facilitates the seamless integration of PDF data into the research workflow, enabling researchers to leverage the rich textual content of PDF documents for various text mining and data extraction tasks.

proquest_view	The proquest_view function is a useful feature in the HistText application that enables researchers to view and explore specific entries from the ProQuest Corpus. By utilizing the proquest_view function, researchers can access and examine individual entries from the ProQuest Corpus directly within the HistText interface. This functionality allows researchers to conveniently navigate from data and data frames to the original source document.
Chinese-specific functions	
list_cws_models	The list_cws_models function provides researchers with a comprehensive list of available Chinese Word Segmentation (CWS) models hosted on the server. CWS is a fundamental task in Chinese text processing that involves dividing a sequence of Chinese characters into individual words or tokens. By running the list_cws_models function, researchers can access information about the different CWS models supported by HistText. This functionality allows researchers to select the most suitable CWS model for their specific corpus or analysis requirements. The ENP-China project carried out an annotation campaign on a large corpus of text from various periods to create a robust model for all periods.
run_cws	The run_cws function enables researchers to apply Chinese Word Segmentation on a given string. By utilizing the run_cws function, researchers can process Chinese text and obtain segmented words or tokens as output.
get_default_cws_model	The get_default_cws_model function provides researchers with the name of the default Chinese Word Segmentation (CWS) model associated with a given corpus. Different corpora may have specific CWS models tailored to their characteristics or requirements. Researchers can retrieve the default CWS model associated with a particular corpus, ensuring consistent and reliable word segmentation results within their analysis.

cws_on_corpus	The cws_on_corpus function allows researchers to apply Chinese Word Segmentation on an entire corpus. By running the cws_on_corpus function, researchers can segment Chinese text within the corpus into individual words or tokens. This enables more granular analysis and exploration of Chinese textual data at a linguistic level, facilitating further computational processing and analysis tasks such as text analysis, topic modeling, etc.
cws_on_df	The cws_on_df function enables researchers to apply Chinese Word Segmentation on a specified column of a dataframe containing Chinese text. By utilizing the cws_on_df function, researchers can segment Chinese text within a specific column into individual words or tokens. This function is particularly useful when working with structured data containing Chinese text, allowing researchers to segment and analyze the text within the context of the dataframe's other attributes and variables.
sinograms_to_py	The sinograms_to_py function provides researchers with the capability to convert sinograms (Chinese characters) into pinyin, which is the romanized representation of Chinese pronunciation. By running the sinograms_to_py function, researchers can obtain the corresponding pinyin representation for a given set of Chinese characters. This conversion is valuable for various applications, such as language learning, linguistic analysis, text normalization, or to transform Chinese names to alphabetical names to be included in a database.
wade_to_py	The wade_to_py function allows researchers to convert Wade-Giles romanization of Chinese characters into pinyin, which is the modern standard for romanized Chinese pronunciation. By utilizing the wade_to_py function, researchers can transform Chinese text in Wade-Giles romanization into its corresponding pinyin representation.
Graph functions	

get_padagraph_url	The get_padagraph_url function is a convenient feature that allows researchers to send a tidygraph object to Padagraph, a graph visualization platform, and retrieve the corresponding URL. Tidygraph is a popular R package for working with graph data structures and performing graph analysis. By using the get_padagraph_url function, researchers can seamlessly transfer a tidygraph to Padagraph for interactive visualization and exploration. The function returns the URL that can be shared or accessed to view the graph visualization in the Padagraph interface.
in_padagraph	The in_padagraph function enables researchers to send a tidygraph object directly to Padagraph and display it within the R-Studio environment. By utilizing the in_padagraph function, researchers can visualize the graph structure and properties of the tidygraph in an interactive and visually appealing manner. This feature facilitates the exploration and analysis of complex networks or relationships encoded in the tidygraph, allowing researchers to gain insights and identify patterns more effectively.
load_in_padagraph	The load_in_padagraph function provides researchers with the capability to load a previously saved graph object and send it to Padagraph for visualization and analysis. Researchers can save a tidygraph object to a file using the save_graph function, and then use the load_in_padagraph function to load the saved graph into Padagraph. This functionality allows researchers to resume their graph analysis and visualization in Padagraph without the need to recreate the graph from scratch, ensuring continuity and efficiency in the research workflow.
save_graph	The save_graph function enables researchers to save a tidygraph object into a file, preserving its structure, properties, and associated data. By using the save_graph function, researchers can store the tidygraph object for future reference or sharing with collaborators. This feature ensures the reproducibility of graph analysis and allows researchers to revisit and reuse the graph data without the need to recreate it from the original data source. Saved graph files can be loaded back into HistText or other environments using the

	load_in_padagraph function for further analysis or visualization.
Server functions	
query_server_get	The query_server_get function is a versatile tool that enables researchers to retrieve a resource from the server using the HTTP GET method. Researchers can use this function to request and obtain specific data or information from the server. The query_server_get function facilitates communication between the application and the server.
query_server_post	The query_server_post function serves to send a file to the server using the HTTP POST method. Researchers can use this function to upload files or data to the server, which can be utilized for various purposes such as data storage, processing, or further analysis. The query_server_post function enhances the flexibility and functionality of HistText by enabling seamless file transfer and interaction with the server.
set_config_file	The set_config_file function is a configuration tool that allows researchers to specify the server URL and other necessary information. By using the set_config_file function, researchers can set up the application to connect with a specific server, ensuring proper communication and data exchange. This function ensures that HistText is configured to access the desired server.
get_error_status	The get_error_status function is a helpful utility that enables researchers to retrieve the error status associated with a server response. When interacting with the server, responses may contain error codes or messages indicating issues or problems encountered during the request. By using the get_error_status function, researchers can access and analyze the error status of a response, facilitating effective troubleshooting and debugging when necessary.
get_server_status	The get_server_status function provides researchers with the ability to check and retrieve the status of the server. By utilizing this function, researchers can obtain information about the server's operational state, including its availability and responsiveness.