HistText: An Application forleveraging large-scale historical textbases

Baptiste Blouin[1], **Cécile Armand**[1], **Christian Henriot**[1]

[1]Aix-Marseille University, France

Corresponding author: Christian Henriot , Christian.R.Henriot@gmail.com

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This paper introduces HistText, a pioneering tool devised to facilitate large-scale data mining in historical documents, specifically targeting Chinese sources.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 large-scale Modern China Textual Database, HistText emerged 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 computational 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.

I 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 In the early 1990s, when historical documents were first converted into digital format as scanned images, the field of ”digital history” emerged, allowing historians to still rely on the established methods developed over the past 150 years.[1] However, the landscape has changed considerably since then. The digital transformation of historical 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.[2]

The digital transformation affects all historical sources, especially the vast repository of historical textual documents (periodicals, archives, diaries, etc.). 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. [3] Furthermore, with the combination of natural language processing (NLP) and computer vision (CV) techniques, handwritten documents can now be transformed into full text as well. Even medieval manuscripts, although requiring meticulous curation, are undergoing their own digital

[1] Cohen et al. [2008], Galgano et al. [2013], Dougherty and Nawrotzki [2016]

[2] Fickers and Tatarinov [2022], Rosenzweig [2011], Lansdall-Welfare et al. [2017], Cordell [2015]

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

transformation. New tools have been developed to tackle manuscript issues across different languages and scripts. Platforms like Transkribus offer various models and algorithms specifically designed for handwritten manuscripts.

While OCR technology has made remarkable advancements, resulting in the ability to convert scanned images into full text with increasingly fewer errors, the presence of misspelling 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, encyclopaedias, directories, and more, cannot be ignored or dismissed, even with some degree of error. These digital avatars raise new hermeneutic issues. In an ideal world, all digital versions of a source would be linked to its original image format, as national libraries usually do, 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 many 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 a crucial condition. Manual transformation of these massive quantities of newspapers found in libraries is simply impractical. Therefore, computational methods based on human annotations, metadata enrichment, and model training 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. The ENP-China project is currently conducting an optical layout recognition (OLR) annotation campaign on a set of Chinese- and English-language newspapers and periodicals using the Coco annotation tool [Brooks, 2019].

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.[5] 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 fulfil a similar role in constructing boundless reservoirs of information. Additionally, public administrations themselves have embraced digitization, 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 training and adequate tools?

From its inception, the ENP-China project was conceived to address 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

[4] Chiron et al. [2017]

[5] Milligan [2019]

Language Processing. One of the project's major achievements is HistText, a user-friendly application that leverages Artificial Intelligence to explore, mine, and interpret the infinite digital collections available to historians and other scholars in the humanities and beyond, with modern China as its starting point.

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), described in section 3.1. Although one of HistText's notable features is the exploration of Chinese historical texts, its capabilities extend to corpora in other languages such as English or French 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.[6]

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 knowledge bases like 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 transformer large-scale language models (LLM) have provided a crucial element for harnessing the potential of computational methods in textual analysis. [7]

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 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 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 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

[6] ENP-China project: https://www.enpchina.eu/.

[7] Radford et al. [2018]

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 computational methods in the humanities and further developments.

II RELATED WORK

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. [8] 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 China-related Textbases

Textbase, or textual database, is a term that denotes a coherent collection of semi- or unstructured digital documents. 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.

Commercial providers and publishers, such as Airusheng and Green Apple for Chinese-language resources, ProQuest and Brill 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, universities, research institutes, and public libraries such as Heidelberg University's Early Chinese Periodicals Online (ECPO), the Institute of Modern History (IMH) collections at Academia Sinica, the Shanghai Library, Chinese University of Hong Kong University Library 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 modelling and authorship attribution to enable more refined queries and to expand possibilities for exploration. [9] 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 possibility to create customized subcorpora, humanities researchers often cannot align their analyses with their research questions. [10]

[8] For a broader overview of text mining tools and digitized newspaper collections, please refer to: Ehrmann et al. [2019], Pfanzelter et al. [2021], Kumpulainen and Late [2022]

[9] Arnold and Rudolph [2021], Arnold et al. [2023]

[10] Ehrmann et al. [2019], Pfanzelter et al. [2021]

On the other hand, initiatives like C-Text, Markus, LoGaRT, as well as toolkits developed by Chinese universities, such as gj.cool, 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 [11], and (3) issues and biases that result from the transformation of texts into digital objects, such as noisy OCR and OLR.

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.

2.2 Newspaper Interfaces

Because the periodical press constitutes the core of MCTB, HistText is more akin to newspapers interfaces like the British Newspaper Archive, NewsEye [12], and Impresso [13]. However, it presents three major differences, which reflect the distinct origins, ambitions, and team composition of these projects. Impresso and NewsEye are large-scale 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 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 like NewsEye and Impresso. Nevertheless, our objectives and challenges remain distinct:

- MCTB comprises a broader range of textual sources (not just newspapers) and non-Western, low-resource languages, notably "transitional" Chinese) (see section 3.1). In contrast to projects based on cultural heritage online portals like Impresso and NewsEye, we did not start from the collections already available at certain libraries [14]. Instead, we built our own collections, following our specific research needs, discoveries, and how they evolved over time.
- 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 modelling, word embedding, text reuse).

[11] Kaske [2008], Magistry [2019], Tsu [2022], Blouin et al. [2023]

[12] NewsEye (2018-2022) was a research project aimed at advancing the state of the art and introducing new concepts, methods and tools for digital humanities by providing enhanced access to historical newspapers for a wide range of users.

[13] The first Impresso project (2017-2020) developed a scalable architecture for the processing of Swiss and Luxembourgish newspaper collections and created an interface with powerful search, filter and discovery functionalities based on semantic enrichments.

[14] Ehrmann et al. [2020b,a]

However, they do not address the downstream challenges of data transformation and analysis. By contrast, HistText aims to equip professional researchers 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 analysing 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.

- The MCTB-HistText articulation is more flexible than most 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 texts not connected with China). Furthermore, HistText is not just a visual interface, but an extensive open-source library. Consequently, the algorithmic models and training data produced during 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 (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 engineers, as it is often the case in many "digital humanities" projects.

III HISTTEXT AND THE MODERN CHINA TEXTUAL DATABASE

3.1 The Modern China Textual Database

The Modern China Textual Database, as one of the three integral pillars supporting research on the transformation of elites in modern China, complements the Modern China Biographical Database (MCBD) and the Modern China Geospatial Database (MCGD). While MCBD captures biographical data and MCGD handles geospatial information, MCTB specializes in textual corpora. This database forms the foundation for extracting historical insights and converting them into actionable data, positioning the ENP-China project at the forefront with its unique focus on Modern China's textual resources.

MCTB offers a diverse array of textual sources, from periodicals and directories to diaries, archives, and dictionaries, alongside digitized content from platforms like Wikipedia (Chinese, English, Japanese) and Baidu (Chinese). Noteworthy, Chinese periodicals within MCTB include the daily Shenbao (1872-1949), the monthly Eastern Miscellany (1903-1949), and the expansive ProQuest collection, which features English-language Chinese publications like 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, specialized journals like the Chinese Economic Bulletin/Monthly/Journal (1921-1936), 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 (IMH) at Academia Sinica (Taipei), 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 directories, all materials have been digitized and processed using OCR. Although some journals are still awaiting processing at the time of writing, they will be added to MCTB in due course. The second category within MCTB 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.

MCTB supports multiple languages, with a particular focus on Chinese, English, and French. MCTB facilitates data mining and analysis through various features. The texts within MCTB underwent pre-training to identify and index all named entities. Named entities refer to specific named objects or entities such as people, places, organizations, and dates. 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 conduct multifaceted queries to create their own corpora and further manipulate the textual data. This is the major difference with consulting platforms and even with mixed-mode interfaces like the IMH collection of Who's who directories.

With resources being open and freely accessible, excluding the commercially acquired Shenbao and ProQuest collections, MCTB aims to foster collaboration and knowledge sharing [15]. The text files within the database are indexed and stored on a SolR server, along with pre-computed Chinese words 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 historical textbases due to its incorporation of diverse genres, multilingual resources, and advanced data mining capabilities through HistText, enhancing its utility for researchers and scholars seeking comprehensive and versatile textual analysis tools.

3.2 HistText: A Meeting of Minds

3.2.1 From History to Computing

HistText represents the culmination of a longstanding and fruitful collaboration between historians and computer scientists that aimed at leveraging computational techniques to enhance 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:

- 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.
- Testing and Feedback Incorporation. To ensure the practicality and relevance of HistText, historians actively participated in testing the tooling and offering constructive feedback.

[15] As of 2023, MCTB includes more than 15 million documents and over 20 billion words. For detailed statistics on all corpora, please refer to: https://gitlab.com/enpchina/ENP-Datasets-Stats.

Workshops and hands-on sessions have been conducted 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 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), and the Association of Asian Studies (AAS) conference in Boston (March 2023), reaffirming the widespread impact and relevance of HistText.

- 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 (NER) and linking has ensured the generation of high-quality, validated data essential for training novel models.

3.2.2 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 endeavour. 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 NER capabilities for Chinese sources. This phase led to rename in February 2022 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, based on LLM, 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 to advance 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 scholars 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 greater control over the process and unlock the vast potential of their source corpora.

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

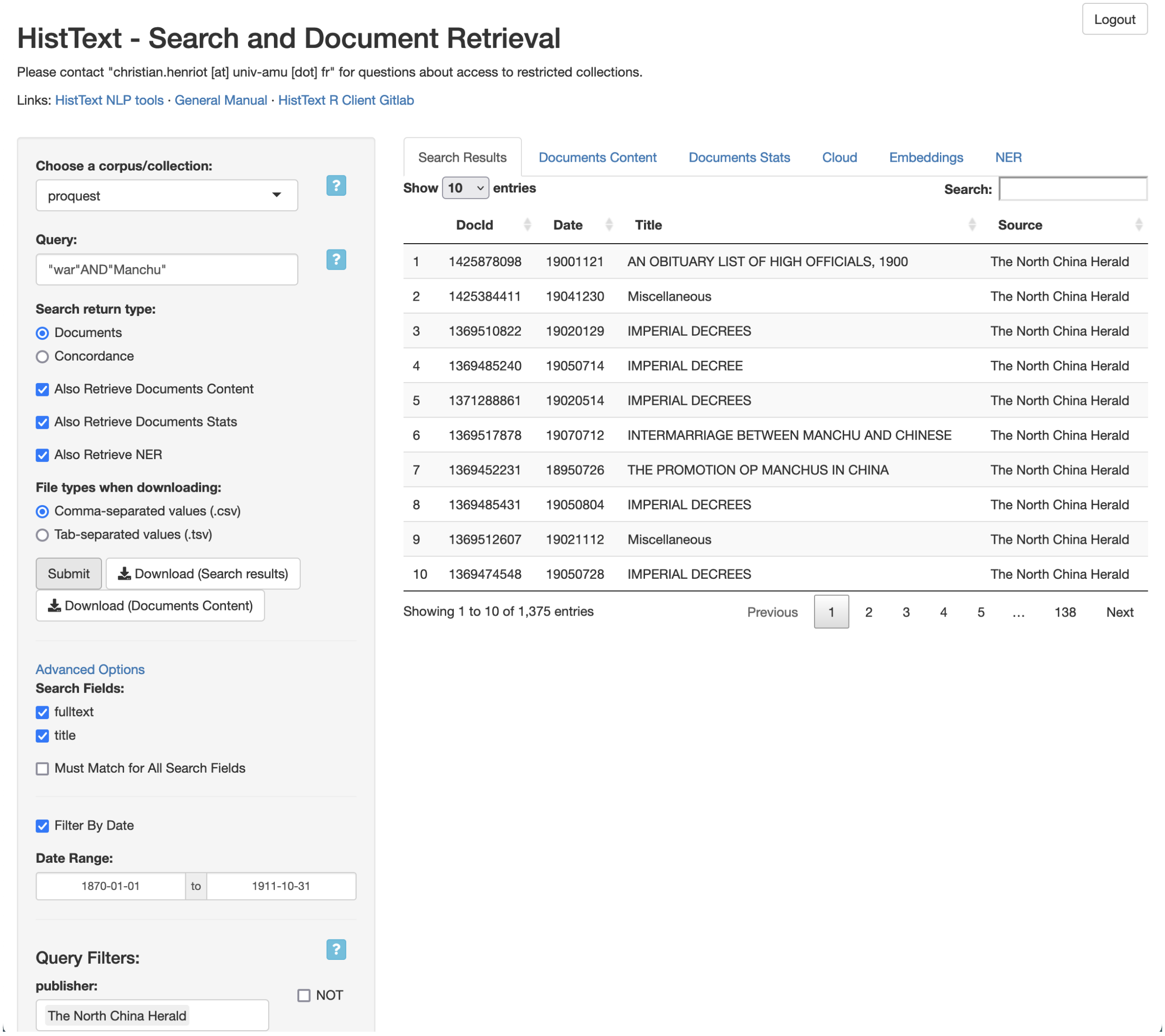
3.3.1 HistText as a user-friendly interface

The ENP-China project has developed a user-friendly interface on R-Shiny, allowing scholars without programming skills to utilize computational methods 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 8).

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 labelled as ’Search results’, presenting a list of identified documents (DocId, 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 comma-separated value (csv) or a tab-separated value (tsv) 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. If the tokens are not already pre-computed on a collection, the tokenization is done “on the fly” to allow researchers to use this version for further analysis.

The ’Documents Stats’ tab features a scrolling menu with six distinct statistical views of the dataset, each complemented by its specific visualization. The first one counts the number of occurrences of the queried term within each document, which can also be delineated by year to offer a temporal perspective. Additionally, it measures the frequency of query results relative to other tokens within the documents, providing insight into their significance over time. The tab also presents statistics related to document lengths. It enables users to determine the proportion



HistText - Search and Document Retrieval

Please contact "christian.henriot [at] univ-amu [dot] fr" for questions about access to restricted collections.

Links: HistText NLP tools · General Manual · HistText R Client Gitlab

Choose a corpus/collection: proquest

Query: "war"AND"Manchu"

Search return type:
● Documents
○ Concordance

☑ Also Retrieve Documents Content

☑ Also Retrieve Documents Stats

☑ Also Retrieve NER

File types when downloading:
● Comma-separated values (.csv)
○ Tab-separated values (.tsv)

Search Fields:
☑ fulltext
☑ title

☐ Must Match for All Search Fields

☑ Filter By Date

Date Range: 1870-01-01 to 1911-10-31

Query Filters:

publisher: The North China Herald ☐ NOT

Show 10 entries   Search:

	DocId	Date	Title	Source
1	1425878098	19001121	AN OBITUARY LIST OF HIGH OFFICIALS, 1900	The North China Herald
2	1425384411	19041230	Miscellaneous	The North China Herald
3	1369510822	19020129	IMPERIAL DECREES	The North China Herald
4	1369485240	19050714	IMPERIAL DECREE	The North China Herald
5	1371288861	19020514	IMPERIAL DECREES	The North China Herald
6	1369517878	19070712	INTERMARRIAGE BETWEEN MANCHU AND CHINESE	The North China Herald
7	1369452231	18950726	THE PROMOTION OP MANCHUS IN CHINA	The North China Herald
8	1369485431	19050804	IMPERIAL DECREES	The North China Herald
9	1369512607	19021112	Miscellaneous	The North China Herald
10	1369474548	19050728	IMPERIAL DECREES	The North China Herald

Showing 1 to 10 of 1,375 entries

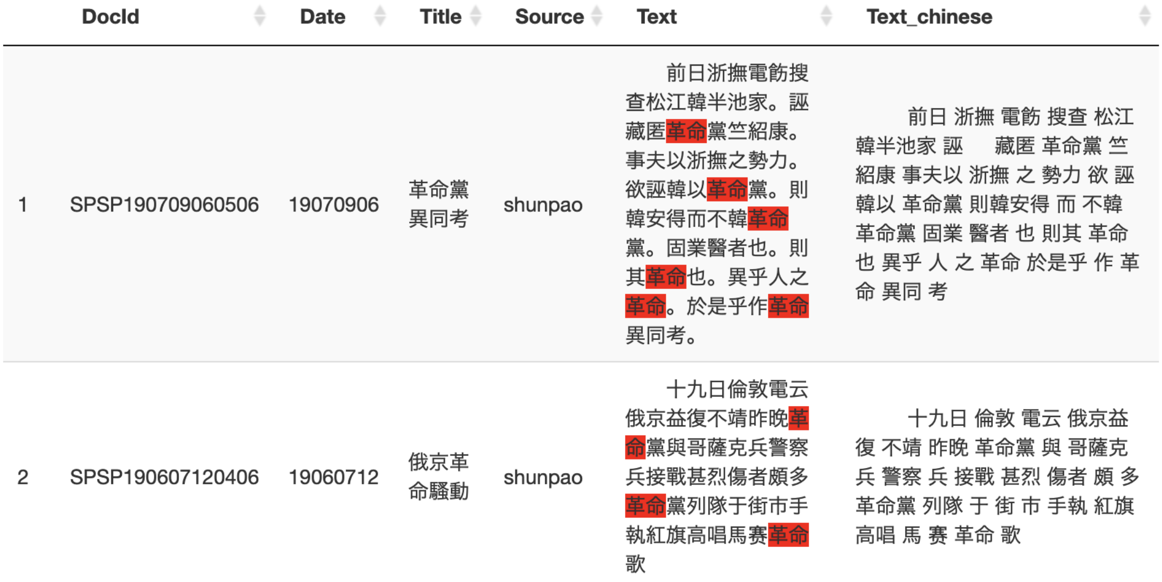
Figure 1: HistText Search and Document Retrieval Interface.

of results in the sub-collection in comparison to the entire collection, offering a perspective on the query's influence throughout the years. Finally, it provides insights into the annual number of results and the distribution of these results across various sources.

In the example below, the graph shows the distribution of the number of mentions of "Wang Ching-wei" (Wang Jingwei, 1883-1944) in the China Weekly Review between 1881 and 1911 (Figure 3).

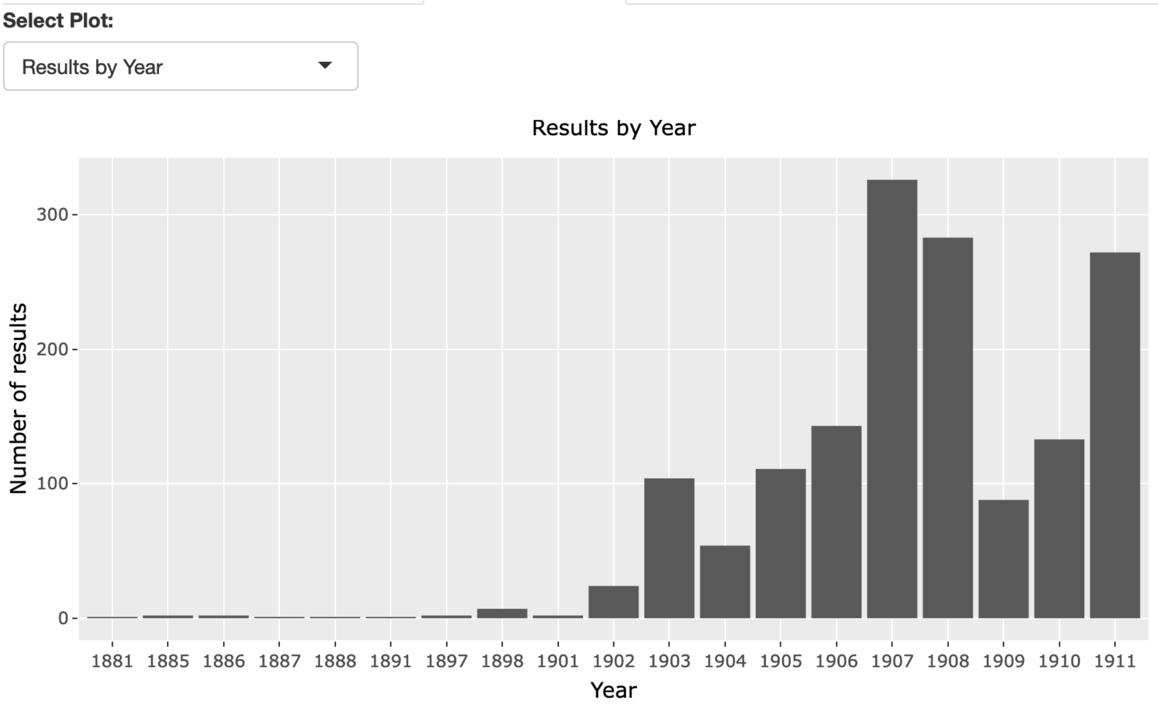
Furthermore, the 'Cloud' tab showcases a word cloud depicting the most frequent terms associated with the queried term. In the case of larger corpora, word embeddings have been calculated (for smaller corpora, a pre-defined set from Wikipedia is employed), which can be utilized to further refine queries by incorporating similar terms suggested by word embeddings [16]. In the example below, we started the query from the term "革命" (Figure 4), 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 5).

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



	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

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

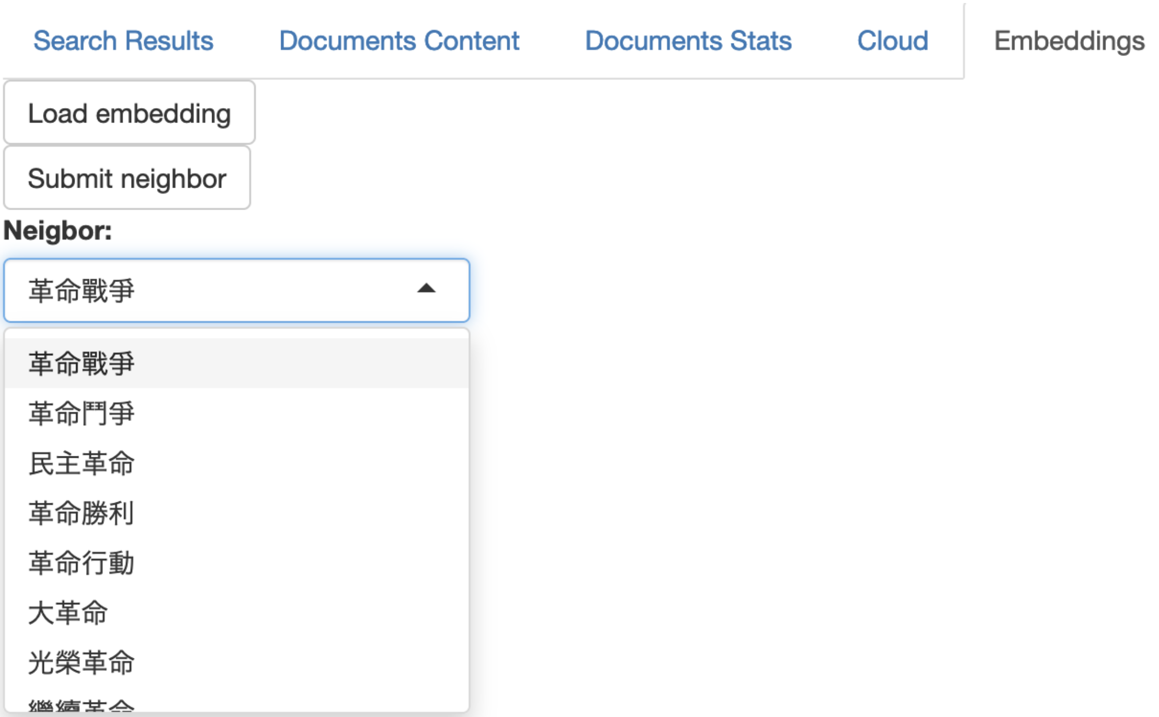


Select Plot: Results by Year

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

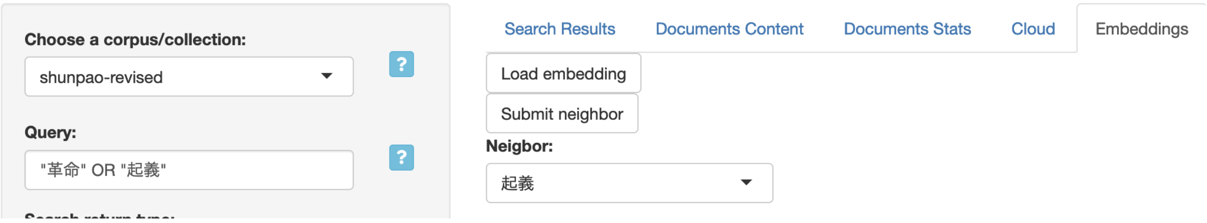
Lastly, the Search page features a NER tab that allows users to gain insights into the named entities present in the first ten documents of the dataset. These visualizations assist scholars in contextualizing their dataset within the selected corpus, such as a specific periodical or a collection of books (Figure 6).

Alternatively, HistText offers the possibility to query terms in their context with the "concordance" search function. One can use the same filters used for searching documents. The results are displayed with a snippet of the text appearing before and after the queried term or expression. Researchers can define the size of the context by specifying the number of characters,



Neigbor: 革命戰爭

Figure 4: Word Embedding Enhanced Search Capability in HistText.



Choose a corpus/collection: shunpao-revised

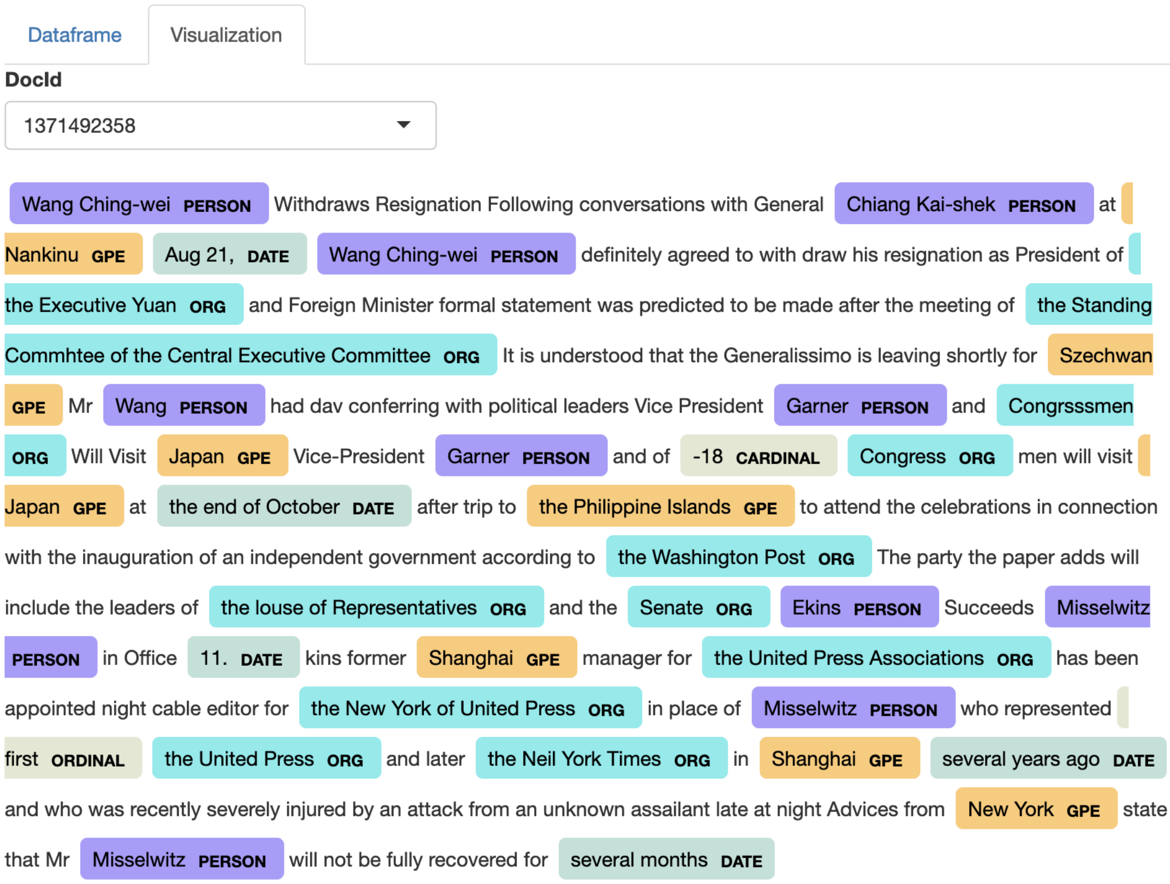
Query: "革命" OR "起義"

Neigbor: 起義

Figure 5: Word Embedding Enhanced Search Capability in HistText.

depending on what they want to examine (Figure 7).

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 a tabular format, displaying the Document identifier (DocId), types of named entities, and their corresponding confidence scores. 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 8). In the ’Visualization’ tab, documents are displayed individually, with all annotated named entities presented by type using colour codes, as presented above. Users can select any document from the original dataset to examine the distribution of named entities in their 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, including the distribution of entity types across all results, the year-wise distribution of these entity types, and their distribution across different documents. Additionally, the frequency of results is presented in relation to model confidence, facilitating the assessment of information reliability. Lastly, these visual representations offer detailed statistics about the positions of entity types within the texts, enhancing the depth of contextual analysis.

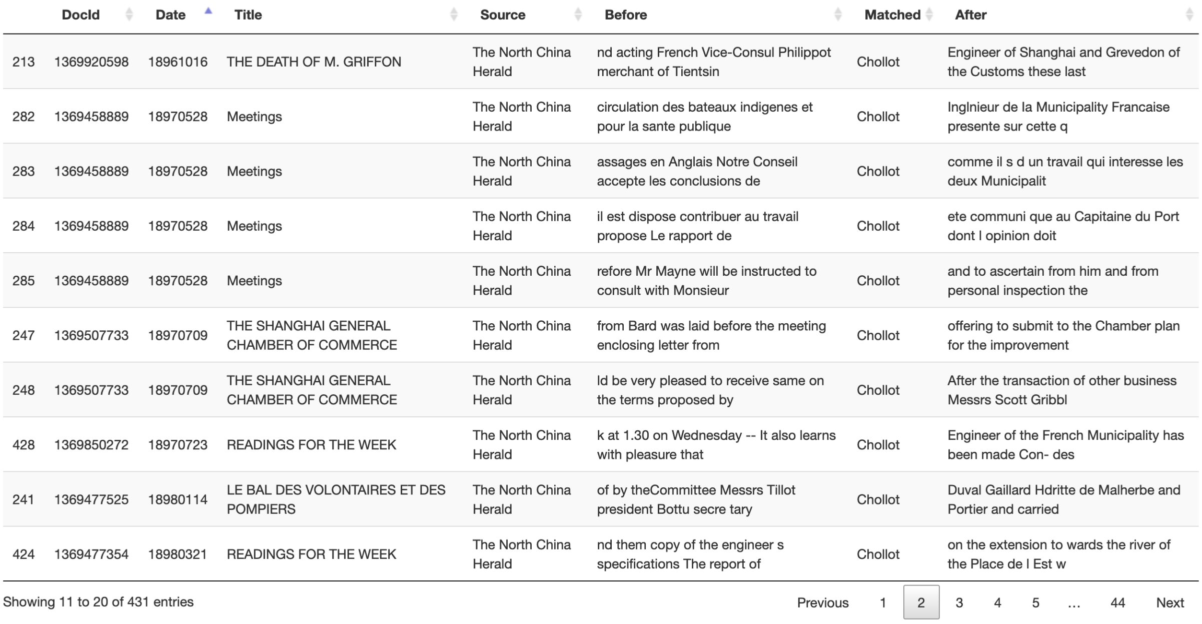


DocId

1371492358

Wang Ching-wei PERSON Withdraws Resignation Following conversations with General Chiang Kai-shek PERSON at Nankinu GPE Aug 21, DATE Wang Ching-wei PERSON definitely agreed to with draw his resignation as President of the Executive Yuan ORG and Foreign Minister formal statement was predicted to be made after the meeting of the Standing Commhtee of the Central Executive Committee ORG It is understood that the Generalissimo is leaving shortly for Szechwan GPE Mr Wang PERSON had dav conferring with political leaders Vice President Garner PERSON and Congrsssmen ORG Will Visit Japan GPE Vice-President Garner PERSON and of -18 CARDINAL Congress ORG men will visit Japan GPE at the end of October DATE after trip to the Philippine Islands GPE to attend the celebrations in connection with the inauguration of an independent government according to the Washington Post ORG The party the paper adds will include the leaders of the louse of Representatives ORG and the Senate ORG Ekins PERSON Succeeds Misselwitz PERSON in Office 11. DATE kins former Shanghai GPE manager for the United Press Associations ORG has been appointed night cable editor for the New York of United Press ORG in place of Misselwitz PERSON who represented first ORDINAL the United Press ORG and later the Neil York Times ORG in Shanghai GPE several years ago DATE and who was recently severely injured by an attack from an unknown assailant late at night Advices from New York GPE state that Mr Misselwitz PERSON will not be fully recovered for several months DATE

Figure 6: An Article Annotated with Named Entities in HistText.



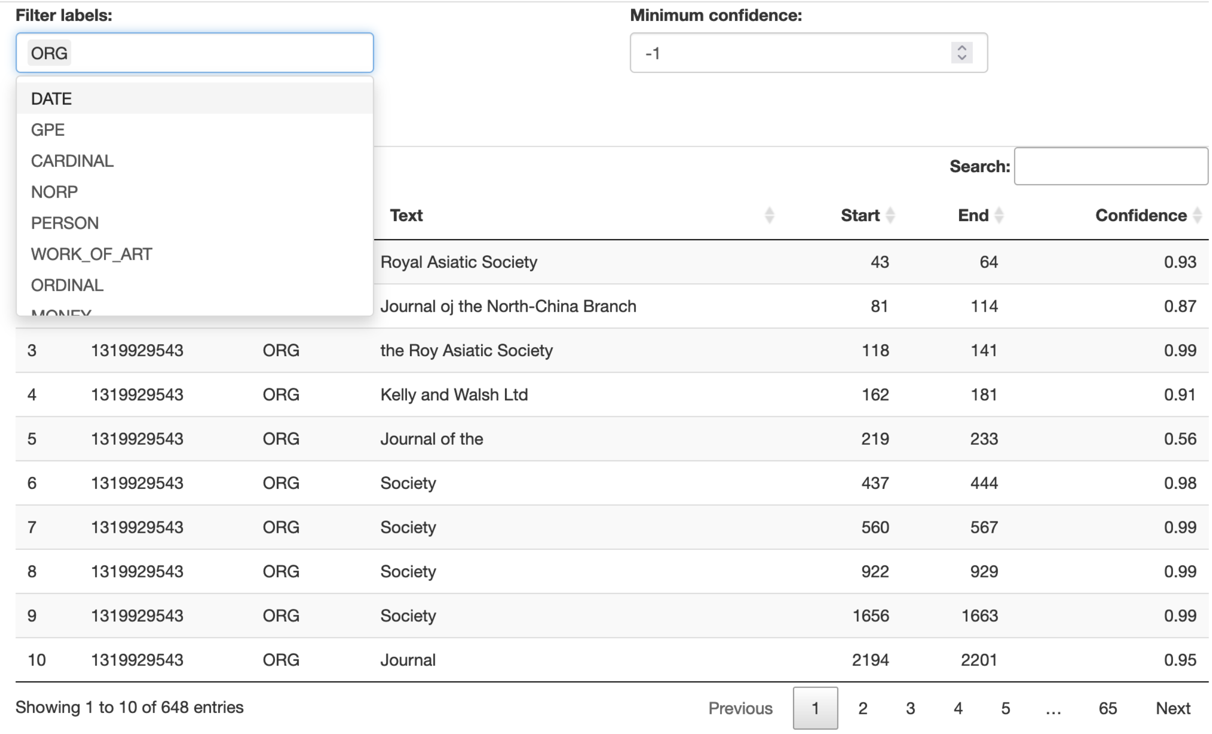
	DocId	Date	Title	Source	Before	Matched	After
213	1369920598	18961016	THE DEATH OF M. GRIFFON	The North China Herald	nd acting French Vice-Consul Philippot merchant of Tientsin	Chollot	Engineer of Shanghai and Grevedon of the Customs these last
282	1369458889	18970528	Meetings	The North China Herald	circulation 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

Figure 7: Concordance Search Results in HistText.

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



Filter labels: ORG

Minimum confidence: -1

Search:

			Text	Start	End	Confidence
			Royal Asiatic Society	43	64	0.93
			Journal oj the North-China Branch	81	114	0.87
3	1319929543	ORG	the Roy Asiatic Society	118	141	0.99
4	1319929543	ORG	Kelly and Walsh Ltd	162	181	0.91
5	1319929543	ORG	Journal of the	219	233	0.56
6	1319929543	ORG	Society	437	444	0.98
7	1319929543	ORG	Society	560	567	0.99
8	1319929543	ORG	Society	922	929	0.99
9	1319929543	ORG	Society	1656	1663	0.99
10	1319929543	ORG	Journal	2194	2201	0.95

Showing 1 to 10 of 648 entries

Figure 8: HistText NER Results in Tabular Format with Named Entity Type Filter.

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 (Figure 9).

The main functions of HistText serve to:

- Build a customized corpus, allowing refined queries with advanced functionalities, including word embedding and concordance.
- Explore metadata through various visualizations and statistics: this is a key tool not only for refining the initial query, but also for enabling digital hermeneutics and source criticism.
- Information extraction (e.g., NER, events)
- Post-query analyses (such as topic modelling and network analysis)

This section highlights the key functionalities of the HistText library. The functions available in HistText can be grouped under six main sets of analytical functions and one set of more technical functions (Control). The six main sets of functions are:

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

The following sections focus on the functions that are more immediately relevant for humanists (query, data extraction, advanced, and Chinese-specific functions). Other, more specific functions (Control, Graph) are described in detail in the Appendix B.

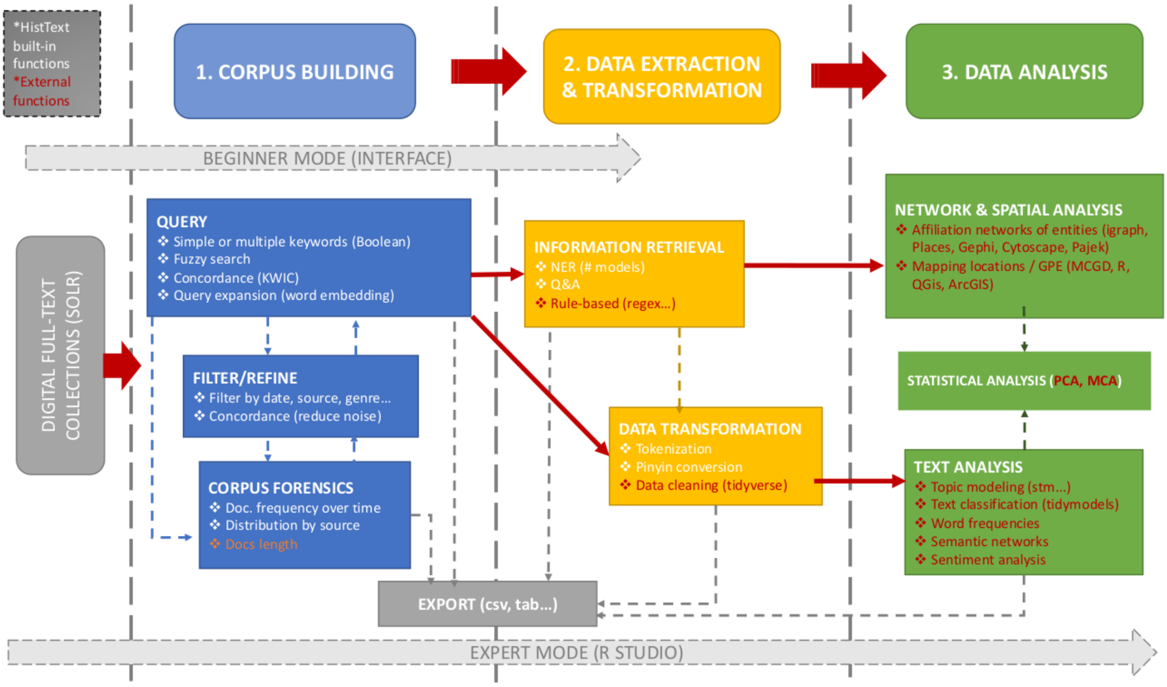

*HistText built-in functions
*External functions
1. CORPUS BUILDING
2. DATA EXTRACTION & TRANSFORMATION
3. DATA ANALYSIS
BEGINNER MODE (INTERFACE)
DIGITAL FULL-TEXT COLLECTIONS (SOLR)
QUERY
Simple or multiple keywords (Boolean)
Fuzzy search
Concordance (KWIC)
Query expansion (word embedding)
FILTER/REFINE
Filter by date, source, genre...
Concordance (reduce noise)
CORPUS FORENSICS
Doc. frequency over time
Distribution by source
Docs length
INFORMATION RETRIEVAL
NER (# models)
Q&A
Rule-based (regex...)
DATA TRANSFORMATION
Tokenization
Pinyin conversion
Data cleaning (tidyverse)
NETWORK & SPATIAL ANALYSIS
Affiliation networks of entities (igraph, Places, Gephi, Cytoscape, Pajek)
Mapping locations / GPE (MCGD, R, QGis, ArcGIS)
STATISTICAL ANALYSIS (PCA, MCA)
TEXT ANALYSIS
Topic modeling (stm...)
Text classification (tidymodels)
Word frequencies
Semantic networks
Sentiment analysis
EXPORT (csv, tab...)
EXPERT MODE (R STUDIO)

Figure 9: The HistText Pipeline.

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 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 the text of 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 pre-processing the text, ensuring that it can be effectively analysed using various computational techniques. By preparing the textual data in a consistent and machine-readable format, researchers can further apply computational algorithms and methods to extract insights and patterns from historical texts.

Data extraction functions

By employing techniques such as named entity recognition (NER), and using adapted models for Chinese historical texts, the HistText library can automatically identify and extract named entities (e.g., persons, organizations, events, locations) from a wide range of historical sources.

More specifically, 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). At present, we have two default models for English and Chinese (pre-trained spaCy [Honnibal et al., 2020] models), in addition to six models specially trained to handle historical source materials. [17] Of these six, two are dedicated to English, three are tailored for Chinese, and one is designed specifically for French.

NER and other data extraction functions in HistText can significantly aid researchers in analysing and understanding the roles, relationships, and occurrences of important actors and entities within historical documents. The extracted data can be processed through other R libraries for data cleaning and network analysis, or used in third-party applications such as Cytoscape, Gephi, Orange, etc.

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. Specifically, the ‘qa_on_corpus’ function allows researchers to apply Question-Answering techniques to an entire corpus, automatically generating and retrieving answers to specific questions. For example, employing questions such as ’Where did name was position?’ or ’Where did name study at?’ allows for the extraction of specific biographical details. The ‘qa_on_df’ function extends QA capabilities to a specified column of a data frame, 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 documents, facilitating targeted data extraction, pattern matching, and more efficient analysis.

Chinese specific functions

Finally, 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 data frame 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 complete 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

[17] Blouin [2022]

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 range of possibilities that HistText functions provide, but interested readers can refer to our online manual [18] and to the descriptive table of the 42 functions (see Appendix B).

3.4 HistText and Digital Hermeneuticss

HistText demonstrates genuine efforts to address the challenges of what Fickers and Tatarinov have termed "digital hermeneutics". 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, re-segmentation of articles in Shenbao). These post-processing operations are extensively documented on the ENP-China GitLab. 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 9). These functions are actionable through both the interface and the library, though the library 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 analysing data, and a critical assessment of the potential biases these tools may introduce in the results [Koolen et al., 2019]. HistText offers several options to address these needs. Firstly, all functions are fully documented and illustrated in the dedicated manuals, while their creation and development are extensively documented on GitLab, including the training and testing data, annotation guidelines, and other resources. Interested researchers can also refer to the specialized publications [Blouin and Magistry, 2020, Blouin et al., 2021, 2022] by Jeremy Auguste, Baptiste Blouin, and Pierre Magistry in relevant conference proceedings.

[18] HistText Interface: https://bookdown.enpchina.eu/Histtext/HistText_interface.html; HistText Manual: https://bookdown.enpchina.eu/rpackage/HistTextRManual.html

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 the possibility to display the confidence scores for NER. It is important to emphasize that the interface offers limited possibilities compared to the library. In the R Studio environment, researchers gain greater control of the arguments 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 and many other specific functionalities. Ultimately, we believe that digital tool criticism is hardly possible without a minimal degree of engagement with programming languages and computational sciences.

IV CASE STUDIES AND APPLICATION SCENARIOS

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

Context & motivation. The Chinese written language experienced a tremendous transformation from the near-classical language of the administration and imperial publications in the 1850s 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. [19].

There is abundant literature describing this pivotal era from different perspectives and disciplines related to language, including the history of language policies [Kaske, 2008], the sociolinguistic aspects [Weng, 2018] or historical linguistics [Coblin, 2000, Simmons, 2017]. 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 million sinograms, to account for the actual language practices and their evolution through time. The work presented here relied on NLP tools for data extraction to provide an unprecedented data-driven account of language practices at that time.

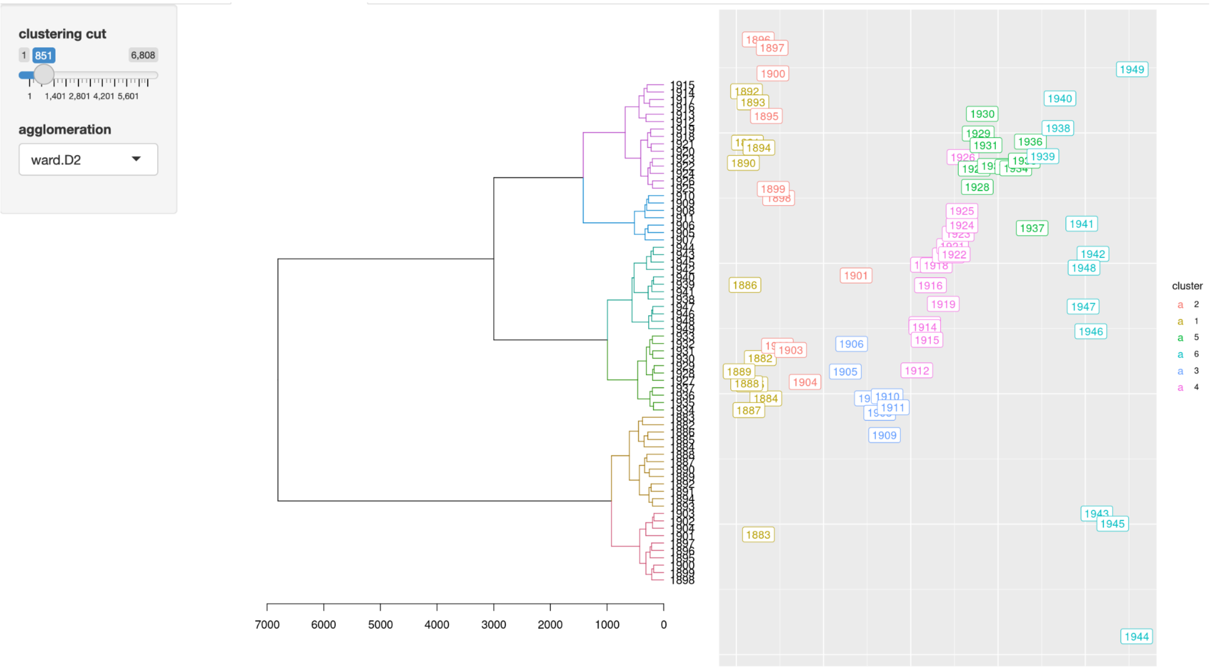
In this section, we examine a single case that highlights the power of NLP tools for language analysis over time. The core issue was to determine the path of language transformation and the major shifts in language practices. The experiment consisted in processing the whole corpus of the newspaper on a yearly basis with language modelling 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.

Workflow. The main idea for this experiment was to use perplexity [20] as the basis to define a metric to apply hierarchical clustering on the different parts of the corpus. To have enough data to estimate a language model on each subcorpus, but still have relatively fine-grained 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 possible to use the perplexity of the language models to define a distance metric between every two years of the Shenbao. Once the distance matrix was built, we applied multidimensional scaling (MDS) to visualize the data on a two-dimensional plane and agglomerative clustering to define periods over the whole corpus.

[19] This section is based on Magistry [2021], Blouin et al. [2023]

[20] A measure used to assess how well a language model predicts the likelihood of a sequence of words, with lower perplexity indicating better prediction performance.

Results. The resulting plot of clustering dendrograms using the Ward method is presented below[21]. It shows 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 10). With minor overlaps, this approach provides for the first time a temporal mapping that delineates very clearly the successive linguistic shifts of the modern Chinese language as it developed sui generis in the press.



clustering cut: 851

agglomeration: ward.D2

Figure 10: The Language Shifts of the Chinese newspaper Shenbao (1872-1949).

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 to reconstruct their career and analyse 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 libraries to effectively manipulate the data and conduct multidimensional analyses downstream.

Workflow. This research proceeded 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 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 11). (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 6 in section 3.3.1), which proved particularly helpful to manually complete missing data, correct noisy output and verify ambiguous entities. (4) Once we had a consolidated dataset of individuals and

[21] Ward's method is a popular method applied in hierarchical cluster analysis.

their affiliated organizations, 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 12), compare networks metrics, and to visualize strong ties and frequent flows (Figure 13). [22]

Results. This research has provided valuable insights on the formation of transnational alumni networks in modern China, a topic which has been largely overlooked in previous studies of elites. Our 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 in 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 (Young Men's Christian Association, 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 [23].

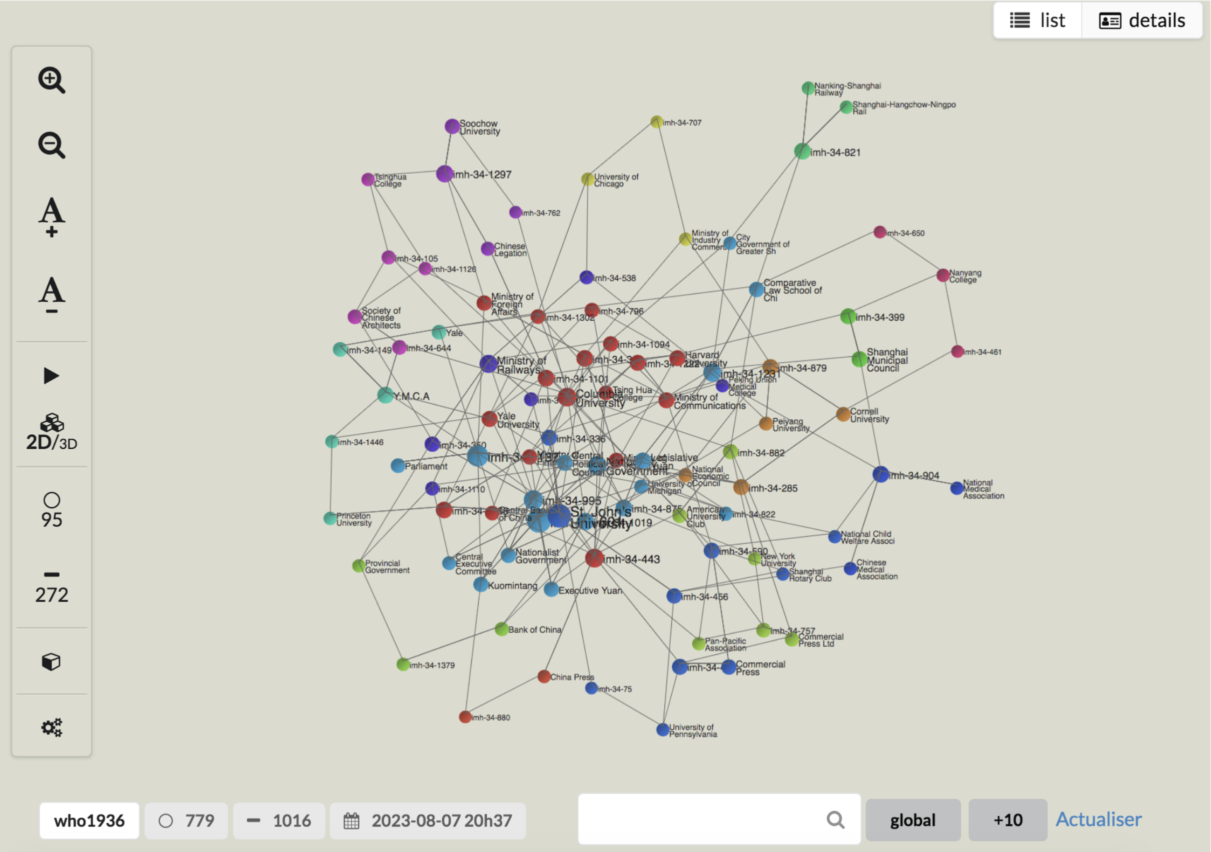


Figure 11: Network visualization of named entities using Padagraph.

Figure 11 shows the links between the biographies of 48 American-educated Chinese sourced from the Who's Who in China (1936) and the organizations extracted from their biographies

[22] The data and full code are available on GitHub

[23] This study will appear in Armand [2024].

using named entity recognition, more specifically, the spaCy model available in HistText. The network was then projected as an interactive graph using the "Padagraph" tool developed by Pierre Magistry. This figure represents only part of the network, centered on Saint John's University, a major missionary institution based in Shanghai, which played a pivotal role in the training and hiring of American-educated Chinese. The different colours represent clusters of more densely connected nodes. The size of the nodes is proportionate to the number of ties they have.

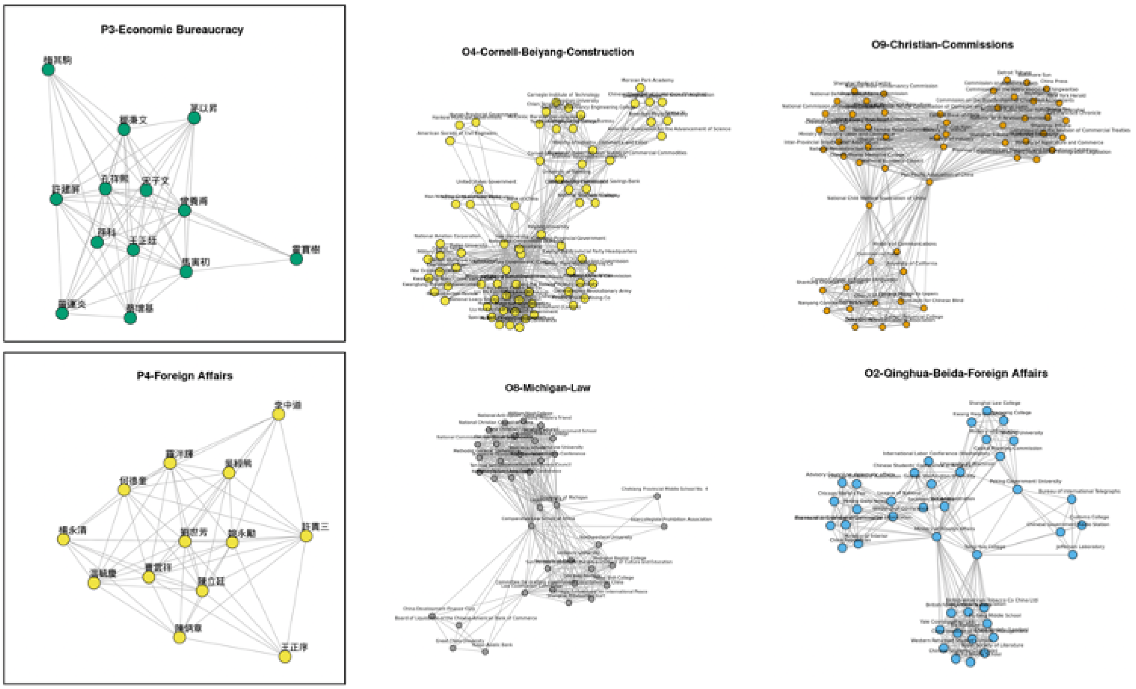


Figure 12: Communities of American-educated Chinese from Who's Who in China, 1936.

Figure 12 represents the largest clusters in the affiliation network of 48 Chinese graduates of American universities. In this context, individual affiliations encompassed educational institutions, professional positions, and membership in clubs and associations. The data was extracted from the Who's Who in China (1936) using the NER function available in HistText. The "igraph" and "Places" R libraries were used to build the networks and detect communities. The initial network of affiliations was projected into two networks: one network of individuals linked by their shared affiliations (left) and one network of institutions linked by the individuals who were affiliated to them (right). Each network was then clustered into communities of more densely connected individuals and institutions using the Louvain algorithm [Blondel et al., 2008] – a popular algorithm for detecting communities in large networks. The figure shows two significant communities of individuals (1) economic bureaucrats (P3) and (2) officials in foreign affairs (P4), each corresponding to two main communities of institutions (1) economic bureaucrats either studied at Cornell and Beiyang universities and worked in China's reconstruction projects after their return from the United States (O4) or who were affiliated to Christian institutions and were employed in various commission as technical experts (O9); (2) officials in foreign affairs who were trained in law at Michigan University (O6) or who were affiliated with Qinghua and Beida Universities and were employed in the Ministry of Foreign Affairs (O2).

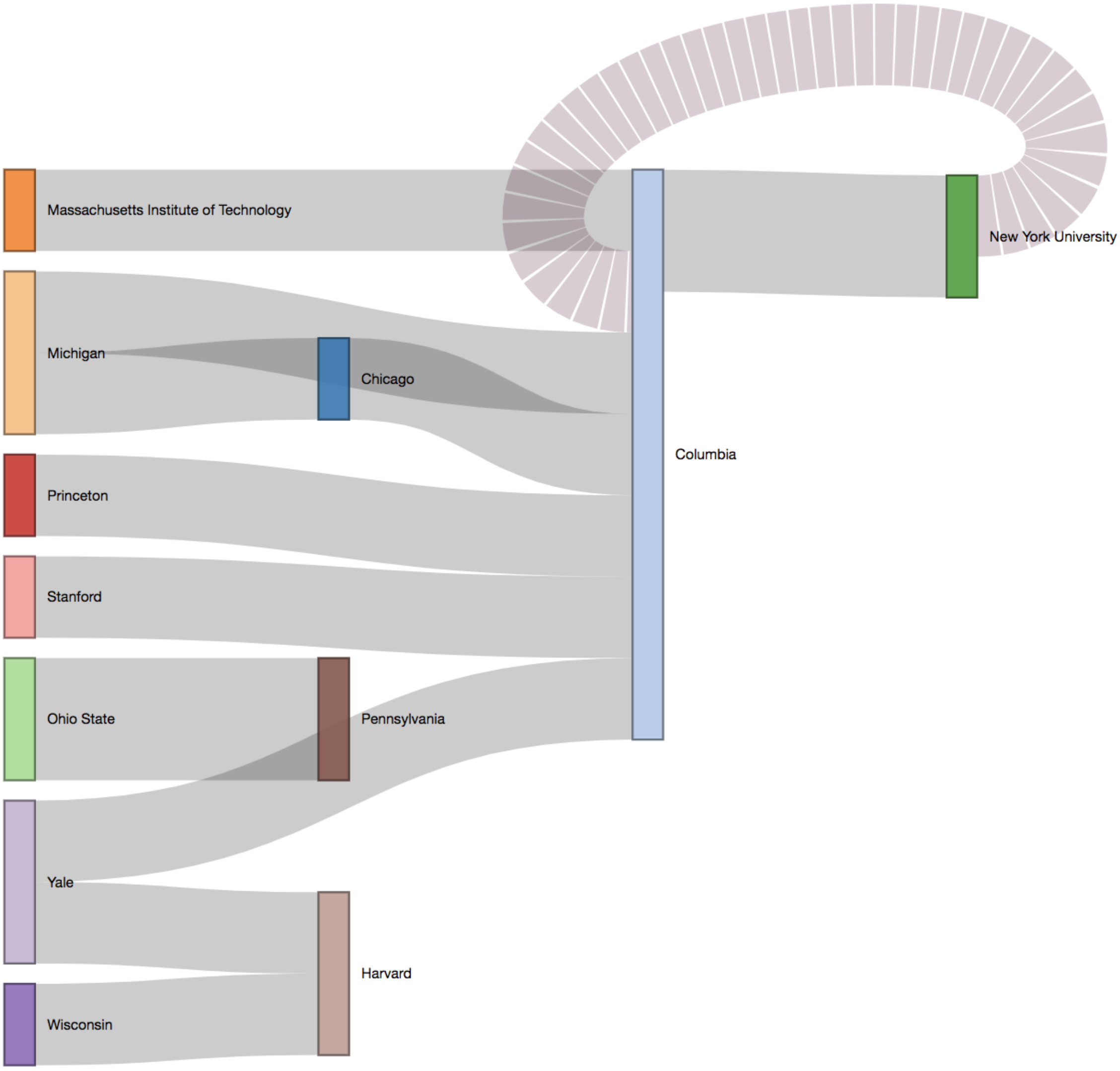


Figure 13: Most important flows of Chinese students in the United States (1883-1935).

Figure 13 was created from the list of students and the universities they attended using the "NetworkD3" R package. Notably, it shows that Columbia University drew many Chinese students from a wide range of other American universities (including Chicago, Massachusetts Institute of Technology, Michigan, Princeton, Stanford, and Yale). This highlights the fact that after graduation, many Chinese students went to Columbia to pursue postgraduate training.

4.3 Topic modelling in Historical Newspapers

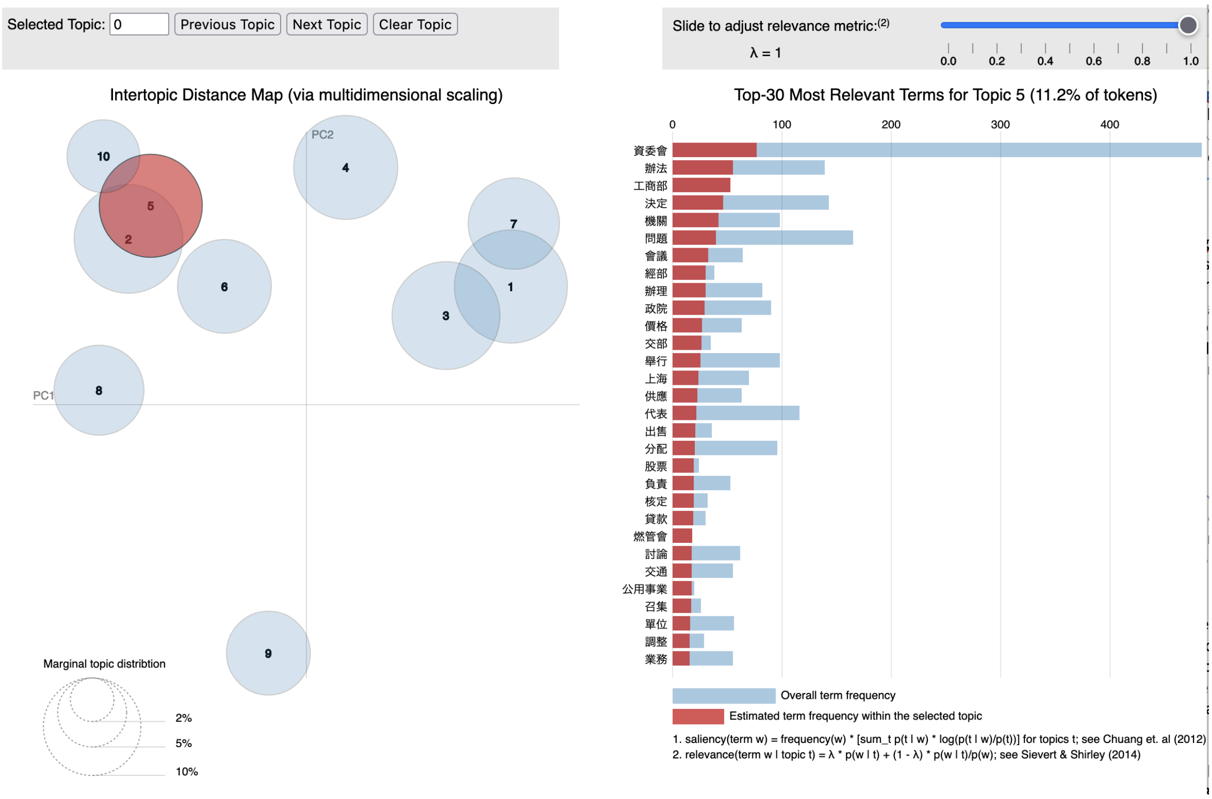
Context & motivation. This case study is an experiment in using different computational methods to explore a simple question: who were the individuals that appeared in the Shenbao in the first twenty years of its existence, especially who were the individuals mentioned repeatedly, whether and how they related to each other, to what institutions they were connected, and what they were involved in? We 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 modelling. As an illustration, we shall only present the last step.

Workflow. Our quest started with a search (search_documents_ex) based on two very common terms in any text in Classical Chinese: 之 (zhi) and 也 (ye). This produced 123,274 and 71,651 results (194,925) respectively. When filtered on 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 in 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 dataset for topic modelling, we built a new sample with articles that contained the validated names of individuals to which we added the full text of related articles. All the operations were done with HistText, except data cleaning and file joins that were done with the Tidyverse library. After removing the duplicates that resulted from joining the validated names and the articles, 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 modelling based on Latent Dirichlet Allocation LDA) using R stm and stminsights libraries [24] 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 14). [25]

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: mostly 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 15). 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 reflected more likely how the newspaper gathered newsworthy information. Besides, several other loosely connected topics were covered, including affairs 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 16) but overall, the nature of news the Shenbao chose to publish was relatively stable (Figure 16). The topics that emerged provide insight into how the Shenbao operated and the process of news-making during its first twenty years. The newspaper relied on existing cultural and social forces, including the highly educated literati who had not made it into the imperial bureaucracy. This exploration through topic modeling allowed us to identify hidden biases in this major source for Shanghai's social history, that close reading in previous studies had not perceived. 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.

[24] LDA is a probabilistic method for classifying documents based on the most frequent word co-occurrences. Structural Topic Modeling (STM) is a general framework for topic modeling with document-level covariate information. The covariates can improve inference and qualitative interpretability and are allowed to affect topical prevalence, topical content or both. https://www.structuraltopicmodel.com

[25] The full study is to be published in Henriot [2024].



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 14: Inter Topic Map and Most Prevalent Terms (topic 7).

Figure 14 represents the relationships between various topics. Central to the map is ”topic 7” (“Xian officials” also present in Figure 15 below as topic 10)), and accompanying it are the most prevalent terms associated with this topic, highlighting its key themes or concepts (produced with R stm LDAvis function).

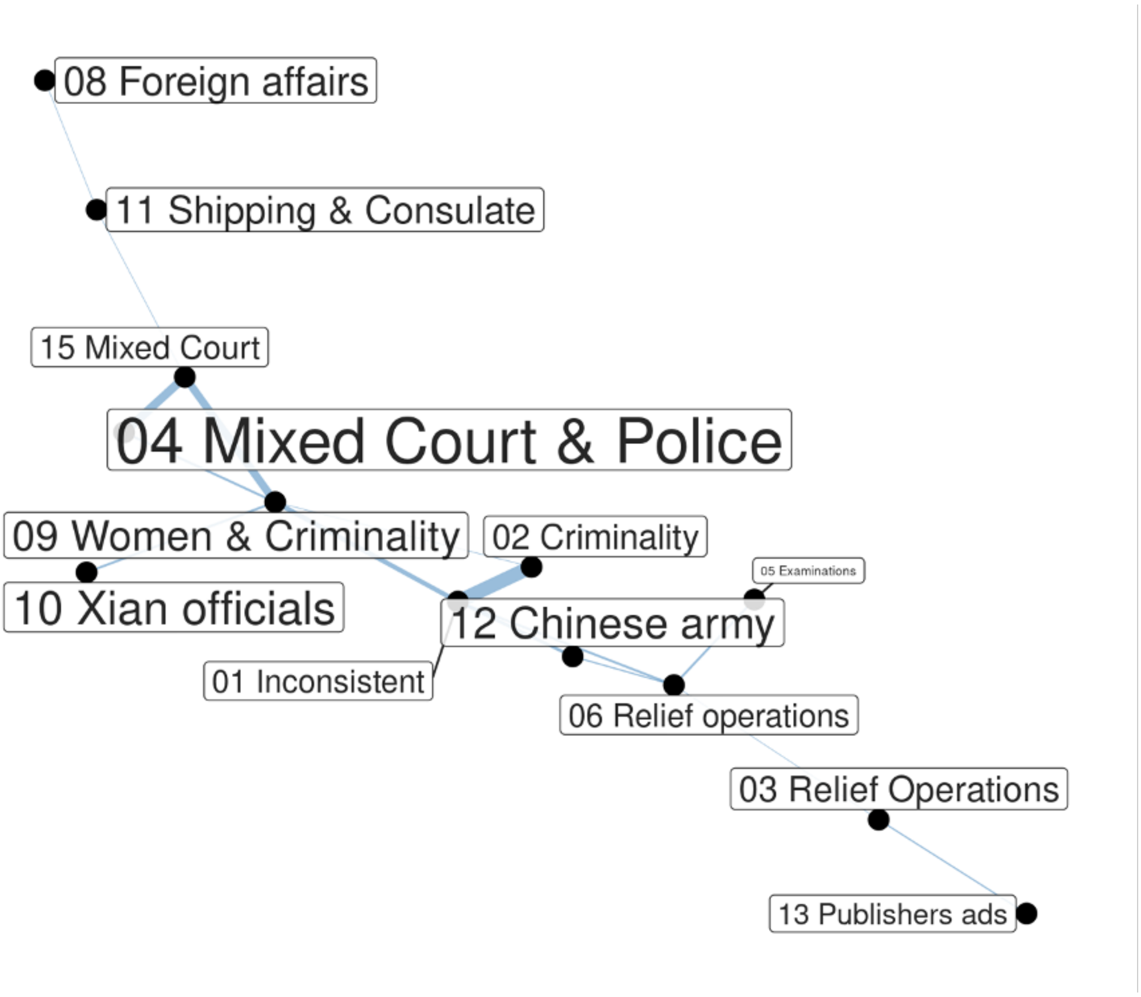


Figure 15: Correlation graph of the 15 topics in the 15-topic model.

Figure 15 presents the relationships between the 15 distinct topics derived from the 15-topic model. It visually represents how closely related each topic is to the others, with connections or lines indicating correlations between specific topics based on the words they share. The strength or weight of each correlation is depicted through varying line thicknesses as can be seen for the topics related to "crime" (produced with R stminsights).

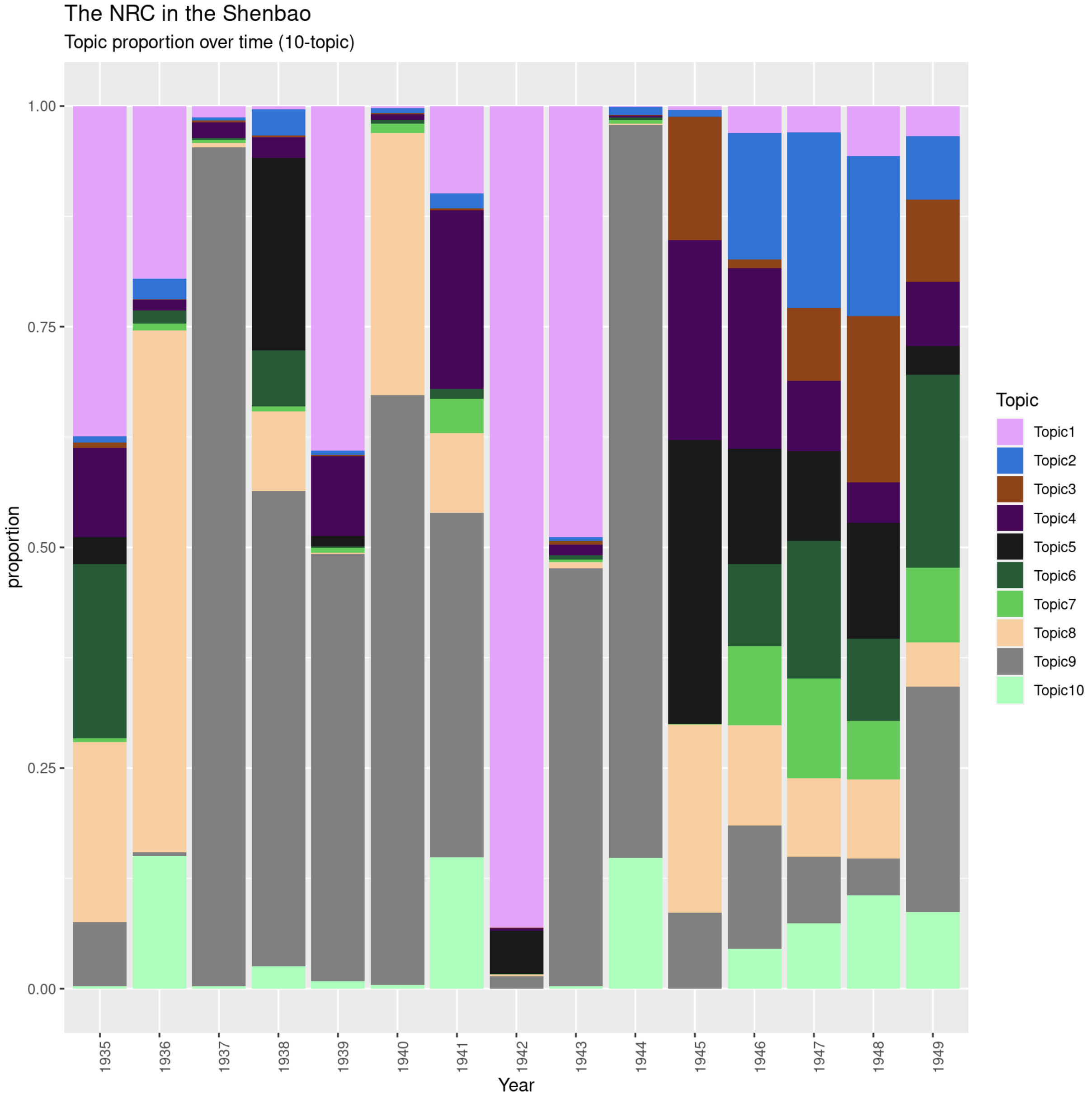


Figure 16: Topic proportion over time (1872-1892).

Figure 16 shows the proportion or prevalence of a specific topic over time. The x-axis represents the timeline, spanning the years mentioned, while the y-axis indicates the proportion or frequency of the topic. The graph provides insights into how the importance or relevance of the topic has evolved or fluctuated during this period. For example, topic 10 (Xian officials) and topic 12 (Chinese army) both experienced a slight increase over the period).

V DISCUSSION

HistText significantly contributes to the field of computational humanities by addressing the challenges of data mining in large-scale historical digital corpora, with a particular focus (though not exclusively) on Chinese sources. The vast volume of documents and words in the Modern China Textual Database necessitates efficient techniques for extracting insights from such extensive repositories. HistText meets this need with its user-friendly interface, advanced text analytics, and powerful visualization capabilities.

By simplifying the data mining process, HistText lowers the entry barrier for researchers, enabling those without extensive computational skills to use advanced text analysis techniques effectively. We contend that such accessibility fosters interdisciplinary collaboration and encourages scholars from various fields to engage with historical texts using computational methods. 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 analyse historical documents comprehensively and systematically. Additionally, HistText's data visualization features help identify patterns, trends, and connections within the corpus, allowing researchers to formulate new research questions, generate hypotheses, and deepen their understanding of historical contexts. The application also paves the way for sophisticated methods such as network analysis, topic modelling, and sentiment analysis.

Despite its considerable contributions and potential, HistText has limitations that must be acknowledged. A primary limitation is its focus on full-text analysis, which excludes the incorporation of images. Textual analysis offers valuable insights, yet historical documents often contain visual elements that provide context and meaning. Incorporating image analysis capabilities would enhance the application's scope, enabling researchers to glean insights from both textual and visual components of historical documents.

Another challenge for HistText involves access and copyright issues. Many digital historical documents are under copyright restrictions, rendering them inaccessible for analysis or limiting the availability of certain text collections. Overcoming these barriers demands collaboration with libraries, archives, and other institutions to ensure legal access to texts and adherence to copyright regulations. Additionally, forming partnerships with institutions specializing in digitization and preservation of historical documents will allow HistText to broaden its collection and enhance researcher access.

The ENP-China project is particularly concerned with HistText's transferability to other texts and research communities beyond the study of modern China. Although HistText primarily focuses on Chinese historical texts and China-related English-language sources, its methodologies can be adapted to analyse texts from various regions and periods. Expanding language support and incorporating diverse corpora would make HistText a versatile tool for researchers across different cultural and linguistic backgrounds. HistText is open source and available on GitLab, facilitating such expansion.

Looking forward, the future development and enhancement of HistText hold several promising avenues. Integrating more advanced machine learning algorithms and natural language processing techniques could improve the accuracy and efficiency of text analysis. By leveraging newer, large-scale models, HistText could handle complex linguistic patterns, automate information extraction, and identify meaningful patterns within historical texts. Ultimately, our aspiration is for historians and scholars to seamlessly integrate HistText into their toolbox and to establish a symbiotic relationship with traditional methods, allowing them to create innovative research solutions tailored to the demands of the ever-expanding corpora of digitized source materials.

ACKNOWLEDGEMENTS

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). Apart from the authors of the present paper, Pierre Magistry and Jeremy Auguste have greatly

contributed to the development of the HistText application.

References

Cécile Armand. Bonding minds, bridging nations: Sino-American alumni networks in the Era of Exclusion (1882-1936). In Christian Henriot, editor, *Modern China in Flux: Networks, Mobility, and Transformation*. De Gruyter, Berlin, 2024.

Matthias Arnold and Henrike Rudolph. Network Data in the Early Chinese Periodicals Online Database (ECPO). *Journal of Historical Network Research*, 5(1), September 2021. ISSN 2535-8863. doi: 10.25517/jhnr.v5i1.118. URL http://jhnr.uni.lu/index.php/jhnr/article/view/118. Number: 1.

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(1): 147–170, February 2023. ISSN 2524-7840. doi: 10.1007/s42803-022-00055-6. URL https://doi.org/10.1007/s42803-022-00055-6.

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, October 2008. ISSN 1742-5468. doi: 10.1088/1742-5468/2008/10/P10008. URL https://dx.doi.org/10.1088/1742-5468/2008/10/P10008.

Baptiste Blouin. *Event Extraction from Facsimiles of Ancient Documents for History Studies*. Doctoral dissertation, Aix-Marseille University, 2022.

Baptiste Blouin and Pierre Magistry. Contextual Characters with Segmentation Representation for Named Entity Recognition in Chinese. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 2–12, Hanoi, Vietnam, October 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.paclic-1.1.

Baptiste Blouin, Benoit Favre, Jeremy Auguste, and Christian Henriot. Transferring Modern Named Entity Recognition to the Historical Domain: How to Take the Step? In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 152–162, NIT Silchar, India, December 2021. NLP Association of India (NLPAI). URL https://aclanthology.org/2021.nlp4dh-1.18.

Baptiste Blouin, Benoit Favre, and Jeremy Auguste. Simulation d'erreurs d'OCR dans les systèmes de TAL pour le traitement de données anachroniques (Simulation of OCR errors in NLP systems for processing anachronistic data). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier TAL et Humanités Numériques (TAL-HN)*, pages 78–87, Avignon, France, June 2022. ATALA. URL https://aclanthology.org/2022.jeptalnrecital-humanum.9.

Baptiste Blouin, Cécile Armand, Christian Henriot, and Hen-Hsen Huang. Unlocking Historical Chinese: A Study on Word Segmentation in Transitional Chinese Texts. 2023.

Justin Brooks. COCO Annotator, 2019. URL https://github.com/jsbroks/coco-annotator/.

Guillaume Chiron, 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 *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries*, JCDL '17, pages 249–252, Toronto, Ontario, Canada, June 2017. IEEE Press. ISBN 978-1-5386-3861-3.

W. South Coblin. A Brief History of Mandarin. *Journal of the American Oriental Society*, 120(4):537–552, 2000. ISSN 0003-0279. doi: 10.2307/606615. URL https://www.jstor.org/stable/606615. Publisher: American Oriental Society.

Daniel J. Cohen, Michael Frisch, Patrick Gallagher, Steven Mintz, Kirsten Sword, Amy Murrell Taylor, William G. Thomas, and William J. Turkel. Interchange: The Promise of Digital History. *The Journal of American History*, 95(2):452–491, 2008. ISSN 0021-8723. doi: 10.2307/25095630. URL https://www.jstor.org/stable/25095630. Publisher: [Oxford University Press, Organization of American Historians].

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*. Modern Language Association, New York , N.Y., 2013.

Ryan Cordell. Viral Textuality in Nineteenth-Century Us Newspaper Exchanges. In Veronica Alfano and Andrew Stauffer, editors, *Virtual Victorians*, pages 29–56. Palgrave Macmillan US, New York, 2015. ISBN 978-1-349-48530-7 978-1-137-39329-6. doi: 10.1057/9781137393296_3. URL http://link.springer.com/10.1057/9781137393296_3.

Jack Dougherty and Kristen Nawrotzki, editors. *Writing history in the digital age*. Digital Humanities. University of Michigan Press, Michigan, 2016. ISBN 978-0-472-05206-6. OCLC: 1080392618.

Maud Ehrmann, Estelle Bunout, and Marten Düring. Survey of digitized newspaper interfaces (dataset and notebooks), August 2019. URL https://zenodo.org/record/3369875.

Maud Ehrmann, Matteo Romanello, Simon Clematide, and Alex Flückiger. Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers. page 38, October 2020a.

Maud Ehrmann, Matteo Romanello, Simon Clematide, Phillip Benjamin Ströbel, and Raphaël Barman. Language Resources for Historical Newspapers: The Impresso Collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 958–968., Marseille, 2020b. European Language Resources Association (ELRA).

Andreas Fickers and Juliane Tatarinov, editors. *Digital history and hermeneutics: between theory and practice*. Studies in digital history and hermeneutics. De Gruyter Oldenbourg, Berlin, 2022. ISBN 978-3-11-072407-3. URL https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=3307515. OCLC: 1337590285.

Michael J. Galgano, J. Chris Arndt, and Raymond M. Hyser. *Doing history: research and writing in the digital age*. Wadsworth Cengage Learning, Boston, MA, 2nd ed edition, 2013. ISBN 978-1-133-58788-0. OCLC: 759910902.

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*, 2024.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python. 2020. doi: 10.5281/zenodo.1212303.

Elisabeth Kaske. *The politics of language in Chinese education, 1895-1919*. Sinica Leidensia. Brill, Leiden, 2008. ISBN 978-90-04-16367-6. OCLC: 171268385.

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(2):368–385, June 2019. ISSN 2055-7671. doi: 10.1093/llc/fqy048. URL https://doi.org/10.1093/llc/fqy048.

Sanna Kumpulainen and Elina Late. Struggling with digitized historical newspapers: Contextual barriers to information interaction in history research activities. *Journal of the Association for Information Science and Technology*, 73(7):1012–1024, 2022. ISSN 2330-1643. doi: 10.1002/asi.24608. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.24608. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.24608.

Thomas Lansdall-Welfare, Saatviga Sudhahar, James Thompson, Justin Lewis, FindMyPast Newspaper Team, and Nello Cristianini. Content analysis of 150 years of British periodicals. *Proceedings of the National Academy of Sciences*, 114(4):E457–E465, January 2017. ISSN 0027-8424, 1091-6490.

Pierre Magistry. Language(s) of the Shun-pao . Taipei, 2019. URL https://analytics.huma-num.fr/Pierre.Magistry/Shun-pao/.

Pierre Magistry. Le(s)? chinois du Shun-pao . In *Journées GDR LIFT 2021*, Grenoble, France, December 2021. GDR LIFT. URL https://hal.science/hal-04059911.

Ian Milligan. *History in the age of abundance?: how the web is transforming historical research*. McGill-Queen's University Press, Montreal, 2019. ISBN 978-0-7735-5821-2. OCLC: 1080220047.

Eva Pfanzelter, Sarah Oberbichler, Jani Marjanen, Pierre-Carl Langlais, and Stefan Hechl. Digital interfaces of historical newspapers: opportunities, restrictions and recommendations. *Journal of Data Mining and Digital Humanities*, HistoInformatics, January 2021. doi: 10.46298/jdmdh.6121. URL https://hal.science/hal-02480654.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

Roy Rosenzweig. *Clio wired: the future of the past in the digital age*. Columbia University Press, New York, 2011. ISBN 978-0-231-15086-6. OCLC: 595738931.

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(1):63–88, 2017. ISSN 0003-0279. doi: 10.7817/jameroriesoci.137.1.0063. URL https://www.jstor.org/stable/10.7817/jameroriesoci.137.1.0063. Publisher: American Oriental Society.

Jing Tsu. *Kingdom of Characters: The Language Revolution That Made China Modern*. Riverhead Books, New York, 2022. ISBN 978-0-7352-1472-9.

Joe H. Ward. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236–244, March 1963. ISSN 0162-1459. doi: 10.1080/01621459.1963.10500845. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500845. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1963.10500845.

Jeffrey Weng. What Is Mandarin? The Social Project of Language Standardization in Early Republican China. *The Journal of Asian Studies*, 77(3):611–633, 2018. ISSN 0021-9118. URL https://www.jstor.org/stable/26572530. Publisher: [Cambridge University Press, Association for Asian Studies].

A RESOURCES

We have released the public version of the code on Gitlab. The current version of HistText 1.6 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.

B APPENDIX: HISTTEXT FUNCTIONS

Control Functions	
accepts_date_queries	Check if a corpus accepts date queries
get_default_ner_model	Get the name of the default NER model for a given corpus
get_error_status	Retrieve the error status of a response.
get_server_status	Get the status of the server
list_corpora	List available collections in SolR
Query Functions	
search_documents	Search for documents
search_documents_ex	Extended Search for documents
search_concordance	KWIC Search In ENP Corpora
search_concordance_ex	Extended KWIC Search In ENP Corpora
search_concordance_on_df	KWIC search in a custom dataframe
get_documents	Retrieve document from ID
count_documents	Get the number of articles matching a query, by date
count_search_documents	Count the number of documents that can be returned by a query
view_document	View a single document in RStudio
Data extraction functions	
ner_on_corpus	Apply Named Entity Recognition on a corpus
ner_on_df	Apply Named Entity Recognition on the specified column of a dataframe
run_ner	Apply Named Entity Recognition on a string
run_qa	Apply Question-Answering on a string
qa_on_corpus	Apply Named Entity Recognition on a corpus
qa_on_df	Apply Named Entity Recognition on the specified column of a dataframe
extract_regexps_from_subcorpus	apply a collection of Regexps to a collection of documents
Advanced functions	
list_search_fields	List possible search fields for a given corpus
get_search_fields_content	Retrieve the content associated with each search field
list_filter_fields	List possible filter fields for a given corpus
list_ner_models	List available NER models on the server
list_possible_filters	List possible filter values for a given filter field
list_precomputed_corpora	List corpora with precomputed annotations
list_precomputed_fields	List fields of a given corpus that have precomputed annotations
list_qa_models	List available NER models on the server
load_pdf_as_df	Load the text from a PDF into a data frame
proquest_view	Display an entry from ProQuest Corpus
Chinese-specific functions	
list_cws_models	List available CWS models on the server
run_cws	Apply Chinese Word Segmentation on a string
get_default_cws_model	Get the name of the default CWS model for a given corpus
cws_on_corpus	Apply Chinese Word Segmentation on a corpus
cws_on_df	Apply Chinese Word Segmentation on the specified column of a dataframe
sinograms_to_py	sinograms() to pinyin conversion
wade_to_py	wade-giles to pinyin conversion
Graph functions	
get_padagraph_url	Send a tidygraph to padagraph and return the URL
in_padagraph	Send a tidygraph to padagraph and displays it
load_in_padagraph	Load and send a previously saved graph object into padagraph
save_graph	Save a tidygraph into a file
Server functions	
query_server_get	GET a resource from the server
query_server_post	POST a file to the server
set_config_file	Sets the config file in order to specify the server URL to use (+ other needed information).