DHQ: Digital Humanities Quarterly
Preview
2025
Volume 19 Number 2

# Unveiling the Critical Nexus of Data Preprocessing and Transparent Documentation for Result Quality and Reproducibility in Digital History

Clodomir Santana <clodomir_at_ieee_dot_org>, Tadeusz Manteuffel Institute of History, Polish Academy of Sciences, Poland and University of California, Davis, USA https://orcid.org/0000-0001-7869-7184
Demival Vasques Filho <demival_dot_vasques_at_uni_dot_lu>, Centre for Contemporary and Digital History, University of Luxembourg, Luxembourg https://orcid.org/0000-0002-4552-0427
Michał Bojanowski <mbojanowski_at_kozminski_dot_edu_dot_pl>, Department of Quantitative Methods and Information Technology, Kozminski University, Poland and Department of Anthropology, Autonomous University of Barcelona, Spain https://orcid.org/0000-0001-7503-852X
Agata Błoch <abloch_at_ihpan_dot_edu_dot_pl>, Tadeusz Manteuffel Institute of History, Polish Academy of Sciences, Poland https://orcid.org/0000-0002-2070-2750

## Abstract

This study underscores the importance of adequate data preprocessing, transparency, and documentation in digital history research, showcasing how these often overlooked practices impact research quality and reproducibility. We present a topic modelling case study involving over 160,000 records of official correspondence of the Atlantic Portuguese Empire from 1640 to 1822 to illustrate how these practices, associated with standardised formats and metadata conventions, facilitate the sharing and reproduction of experiments. First, we evaluate the impact of data cleaning and preprocessing on model performance. Second, concerning model selection, we compare the performance of latent Dirichlet allocation (LDA), latent semantic indexing (LSI), and Gibbs sampling algorithm for a Dirichlet mixture model (GSDMM). Besides stressing the underestimated significance of data preprocessing and transparent documentation to strengthen research robustness and contribute to a reproducibility culture, we also demonstrate the potential of topic modelling in digital historical studies, specifically in the context of the Atlantic Portuguese Empire.

## 1. Introduction: Bringing Preprocessing and Documentation to the Spotlight

The tools of digital humanities have considerably broadened the perspectives of historical scholarship. However, new opportunities do not come without challenges. In this paper we discuss one of these new challenges historians are facing: dealing with huge amounts of data that require methodological practices previously unfamiliar to these scholars. 1

Historical research prior to the 1970s and 1980s is often referred to as “pre-digital” quantitative history. The term “pre-digital” encompasses two important aspects within historical studies: Firstly, it refers to the period before the widespread introduction of computer software for quantitative data analysis. Secondly, it represents a time before the increasing accessibility of digitised archival collections and their storage in machine-readable formats [Burrows 2023] [Kesner 1982]. One of the first widely known experiments in digital analysis of historical data can be found in the pioneering work of William O. Aydelotte in 1963 [Aydelotte 1963]. Yet, digital (automated) data extraction and preprocessing took much longer to become part of the digital historian toolbox and only became widely used in the 21st century. [Petterson et al. 2016]. 2

In Padgett and Ansell’s study on the rise of the Medici family in 15th century Florence, a network analysis was carried 3

out in which a basic dataset was created by manually extracting data from various published and unpublished sources [Padgett and Ansell 1993]. Different types of relationships were manually coded, from kinship and patronage to personal and financial ties. Charles Wetherell, Andrejs Plakans, and Barry Wellman also used manual extraction of historical data for network analysis [Wetherell, Plakans and Wellman 1994]. The primary data source came from nominal censuses covering the period from 1795 to 1850, whose extraction was crucial in defining both the extent and structure of the personal communities of peasants residing in Pinkenhof in the Russian Baltic Province of Livland. Timothy Brook's approach was similar to the previously discussed manual data extraction methods [Brook 1981]. In researching Chinese trade networks during the 16th century under the Ming dynasty, he extracted and listed relevant information to create a table of place names and the corresponding administrative level and modern Chinese nomenclature.

4 With the growing interest in open access archives, many libraries in the Western hemisphere have introduced complex methods such as Optical Character Recognition (OCR) to digitise their collections [Salmi 2021]. This trend started with the initiative of governments and private companies in the 1980s who wanted to make their documentation more accessible [Kesner 1982]. An important observation that Richard Kesner made at the time was that archivists have traditionally struggled to agree on standardising metadata, as it was impossible to use a uniform format for different archival sources [Kesner 1982]. The lack of a standardised format for historical records is a key reason why this article highlights the importance of appropriate preprocessing in digital history.

5 When considering the profound impact that digital tools have on data-driven historical research, two different approaches can be identified: the traditional method of manual extraction followed by computer analysis, and the fully digital approach, where computational tools are used from the outset. When it comes to analysing networks, many historians opt for manual extraction on the one hand and then use tools such as Python, Gephi, or Nodegoat to map data, visualise relationships, and construct their arguments on the other [Haggerty and Haggerty 2011]; [Verbruggen, Blomme and D'haenick 2020]; [Atunes 2021]; [Ye 2022]. On the other hand, automated extraction can be performed from either digital databases and repositories with previously carefully prepared metadata [Wittek and Ravenek 2011]; [Petz and Pfeffer 2021]; [Geraerts and Vasques Filho 2024], born-digital sources [Linkevicius de Andrade and Vasques Filho 2022], or using Natural Language Processing (NLP) methods to extract entities directly from traditional historical sources [Elwert 2020]; [Zbíral amd Shaw 2022].

6 Having a paper document and creating a dataset with proper metadata are challenging tasks that require several steps. For instance, Elo describes how he processed the documents with OCR software and then used a Python programme to extract data such as institutions, parties and individuals [Elo 2020]. In another example, using data made available in Extensible Markup Language (XML) transcriptions of baptisms and marriages, Geraerts and Vasques Filho investigated religious choice in the 18th century [Geraerts and Vasques Filho 2024]. With Python scripts, they extracted the content of interest from the transcriptions and organised it in comma-separated values (CSV) files. These data were then ingested into a specifically designed graph database — modelled to connect people, events, and places — in such a way to facilitate quantitative and qualitative network analysis.

7 Computer-assisted historical research thus offers a wide range of possibilities for the analysis of manually or automatically extracted data. These methods include social network analysis [Atunes 2021]; [Petz and Pfeffer 2021]; [Ye 2022]; [Geraerts and Vasques Filho 2024]; [Linkevicius de Andrade and Vasques Filho 2022], quantitative analysis [Haggerty and Haggerty 2011], semantic text modelling [Zbíral amd Shaw 2022], topic modelling and random indexing [Wittek and Ravenek 2011]; [Fridlund and Brauer 2013]; [Ravenek, van den Heuvel and Gerritsen 2017], spatial methods [Verbruggen, Blomme and D'haenick 2020], or even a mix of them [Ravenek, van den Heuvel and Gerritsen 2017]; [Curran_2018]; [Elwert 2020]. That said, collected data is usually not ready for direct processing during the data analysis stage of the research – preprocessing is required in between. Jentsch and Porada illustrated the steps of digital research, showing the pipeline from unstructured text to structured data, that begins with data collection, moving to the OCR process, NLP methods, data access, and ends with data interpretation [Jentsch and Porada 2012, 90].

8 The lack of standardised metadata for historical and archival sources makes the preprocessing stage (with its proper documentation) even more relevant in historical research; as mentioned, this is one of the driving forces behind this discussion. Nevertheless, scholars in the field of digital history often prioritise data collection, interpretation, and

especially visualisation, often overlooking the critical intermediate steps. Referring to the early example of computational analysis of voting data in the House of Commons, Aydelotte summarised the preprocessing stage with a single statement: “I am confident that the information currently stored on my cards is largely accurate” [Aydelotte 1963, 139]. Still, repeating this experiment or adapting it to our needs seems to be an arduous task or even impossible. Reproducibility involves much more than just repetition and verification, but also an understanding of the meticulous steps that digital historians take to arrive at their results. In examining Padgett and Ansell’s exploration of the rise of the Medici family, to take yet another example, questions naturally arise about the methods they use to organise both published and unpublished sources, as well as their approach to standardising data. These concerns exist not only in cases where the data was extracted manually, but also in automated processes.

Such examples underscore a broader concern in the field: ensuring that research can be replicated or validated by others in order to maintain credibility [Peng 2011]. While this practice is becoming increasingly important [Stodden et al. 2014] [Peng and Hicks 2021], the field of digital history presents a particular challenge. As previously mentioned, digital historians often prioritise data collection and visualisation over adequate documentation of the preprocessing stage required for reproducibility. 9

Based on the illustrative workflow in Figure 1, showing the conventional process in digital historical research, we observe that steps 1 and 2 (experimental design and data collection) and steps 4 and 5 (data analysis and interpretation of results) are disproportionately emphasised. In contrast, step 3, the preprocessing phase, is often not reported with comparable detail. Our earlier works also share that characteristic and did not give much attention to the importance of the data preparatory decisions [Błoch, Vasques Filho and Bojanowski 2021] [Błoch, Vasques Filho and Bojanowski 2022]. 10

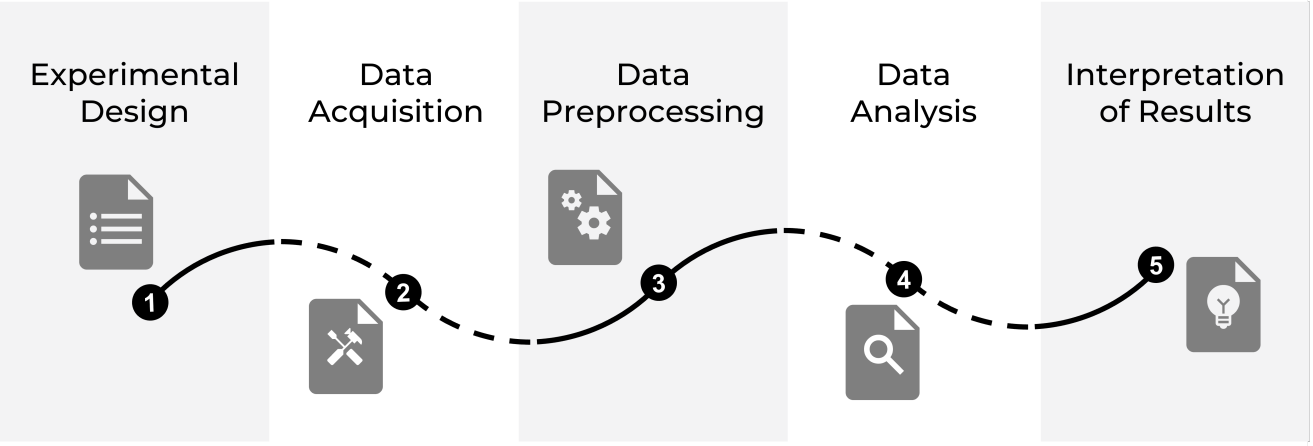


**Figure 1.** A typical workflow in digital history.

Thus, our goal is, in the first step, to review the preprocessing stage of a typical digital history project by discussing its elements and their relevance and providing examples of decisions made in the literature. In the second step, we illustrate the impact of preprocessing decisions on the substantive quality of results with a case study of a topic model for classification of a corpus of around 168,000 historical documents – in particular, how different choices of text preparation (filtering) and model selection affect the number of compositions of identified document clusters (topics). 11

To this end, the remainder of the article is structured as follows. In section 2, we review the stages of the typical workflow of digital history and the place the preprocessing takes in it – types of preprocessing decisions usually made, their impact illustrated with examples. In section 3, we discuss motivations, approaches, and solutions regarding the proper documentation as well as metadata creation and standardisation. Section 4 describes the case study in which we show how substantively impactful the preprocessing decisions can be. Finally, the paper concludes with a discussion in section 5. 12

## Data Preprocessing as the Foundation for High-Quality Results

Preprocessing is an iterative process that includes activities related to data preparation, such as cleaning and normalising data, spelling corrections, reducing imperfections, solving problems with missing data, and, in the case of textual data, also lemmatising/stemming and sometimes even removing stop words or tags [Kunilovskaya and Plum 2021] [García et al. 2016]. This particular stage is of great importance, even if it is not very fascinating and takes up a lot of time. Kunilovskaya and Plum claim that this stage is believed to take up to 80% of the entire research process. Nevertheless, they also refer to a report based on a survey in which the average time spent by 2,300 respondents worldwide in this critical phase was 45% [Kunilovskaya and Plum 2021]. It is also worth noting that this is an iterative process due to the eventual need to perform some steps more than once. For example, after executing the preprocessing pipeline and analysing the processed data, we noticed that a few terms were not removed because they were not present in the stop words list, for example. In this case, we modified this step to include these terms and performed the cleaning again. 13

These general preprocessing challenges become even more complex in the context of digital historical research due to the discrepancies between historical sources, archival practices and the nature of the documents in question. Determining whether the documents are digitised or physical, their linguistic diversity and their categorisation (e.g., administrative, religious or private) is crucial [Matthew and Bannister 2020] [Broadwell et al. 2020]. Each of these elements requires different approaches and an accurate assessment of the advantages and disadvantages of dealing with historical data. 14

In light of these complexities, thorough documentation of the preprocessing phase significantly increases data quality and fosters reproducibility. It is important that researchers provide insights into best practices for the manual extraction of data from both unprinted [Atunes 2021] and printed materials [Ye 2022], including analyses of small and large corpora [Elo 2020] [Elwert 2020]. In addition, sharing methods for standardising and cleaning data from existing databases and digital sources [Petz and Pfeffer 2021] [Geraerts and Vasques Filho 2024] together with identifying the most useful metadata for specific research purposes [Zbíral amd Shaw 2022] [Ravenek, van den Heuvel and Gerritsen 2017] is a promising way to advance the field of digital history. The question of how researchers engage with sources in different non-English languages [Wittek and Ravenek 2011] [Błoch, Vasques Filho and Bojanowski 2022] can also provide important insights for the broader community of digital historians focusing on non-mainstream materials and incorporating more marginalised archival collections. 15

An exemplary guide to solid preprocessing documentation designed to ensure reproducibility under an open access license can be found in the work done by Jiménez-Badillo et al. [Jiménez–Badillo et al. 2020]. The authors meticulously explain their methodology, focusing in particular on geographical text analysis, and describing in detail the individual steps and the challenges encountered in the process. It also provides knowledge about digital research and bridges the gap between data collection and data analysis by explaining what should be done in between. 16

Thus, effective documentation of the preprocessing phase – including detailed explanation of the nature and characteristics of the sources (raw data) – is essential to improve result quality and reproducibility. Regarding the latter, the objective is to enable other researchers who follow the described methodology to arrive at similar conclusions, or build upon previous results, when working with the same data. When it comes to results quality, as preprocessing is an iterative process, researchers can improve the results by optimising the approaches performed in this phase by maintaining a good documentation. If the preprocessing steps are insufficiently described, each iteration can generate more difficulties and uncertainties to digital historians. 17

The preprocessing pipeline (Figure 2) is employed to enhance the quality of the raw data [Lee, Liong and Jemain 2017] [Fan et al. 2021] [Engel et al. 2013], after data acquisition (or collection). It usually involves multiple steps addressing issues such as scattering, baseline changes, peak shifts, noise, missing values, and various other artefacts [Mishra et al. 2020]. This process is the foundation for robust and valid data analysis. The preprocessing pipeline should take into account the specificities of the data and the requirements of the data analysis models, help unveil relevant underlying structures and, when necessary, accurately predict properties of interest [Mishra et al. 2020] [Engel et al. 2013]. 18

Extensive research has yielded a variety of preprocessing techniques [Mishra et al. 2020] [Fan et al. 2021]. Within the 19

spectrum of these proposed methodologies, some are tailored towards the automated identification of preprocessing strategies [Mishra et al. 2019] [Mishra et al. 2020], while others address specific tasks, such as data transformations and standardisation [Cao, Williams and Williams 1999] [Bazyer 1995] [Shanker, Hu and Hung 1996]. Synergistically employing multiple preprocessing techniques can effectively mitigate artefacts that persist when relying solely on a singular technique. Figure 2 depicts an example of a preprocessing pipeline.

Usually, cleaning is the first preprocessing step. It aims to transform raw data into a clean, consistent, and reliable dataset that is accurate, amenable to subsequent analyses, modelling, and interpretation [Rahm and Do 2000]. In this step, we address issues such as identifying and treating missing values, outliers, duplicates, and any inconsistencies in the data. 20

Next, one might require data reduction and scaling depending on the data characteristics (e.g., high dimensional data) or problem tackled (e.g., classification and regression tasks). Data reduction aims to reduce the volume of data without losing important information [Lelewer and Hirschberg 1987]. It focuses on relevant data and eliminating noise or redundancy. In contrast, data scaling involves transforming the features of a dataset so that they fall within a similar scale or range Cao et al., 2016; [Alshdaifat et al. 2021]. Scaling the data can mitigate the risk of, for example, feature dominance due to their larger magnitudes. 21

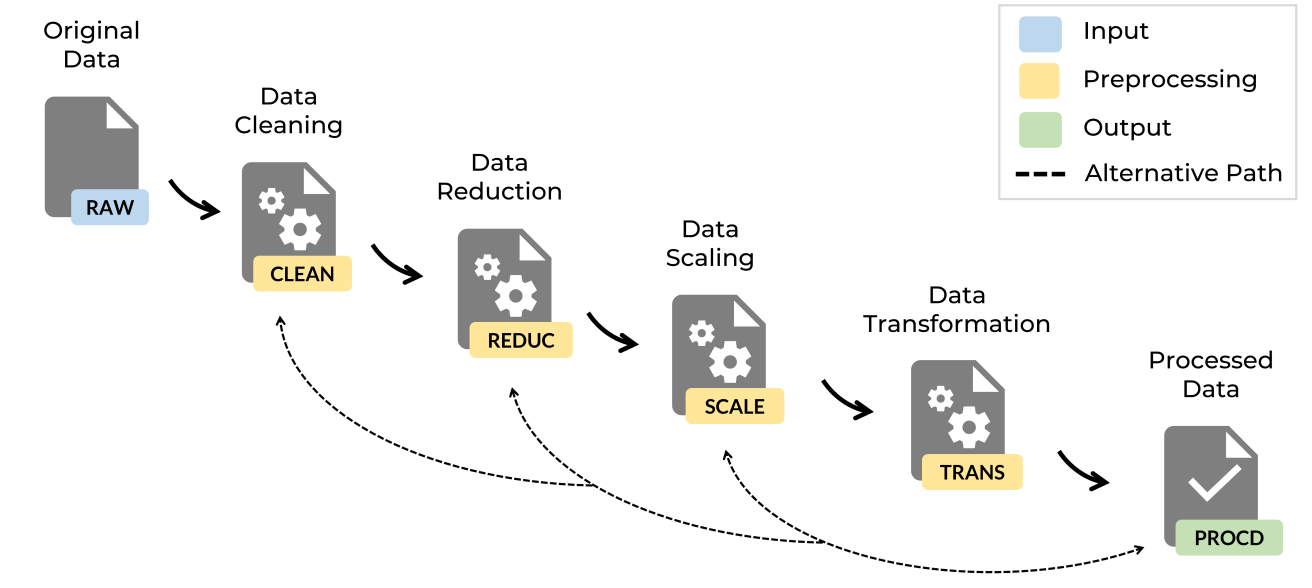


**Figure 2.** Example of typical steps in a preprocessing pipeline. Note that some of the depicted steps are optional in some cases. The section of the steps performed depends on the data characteristics and the goal of the analysis. Also, the dashed arrows go in both directions to indicate the iterative nature of the preprocessing.

Lastly, data transformation procedures might be necessary to ensure the compatibility of the data with the methods used for modelling and analysis [Fan et al. 2021]. For example, some methods struggle to handle categorical data, so an encoding strategy would ensure compatibility, leading to better and more reliable results. 22

As mentioned, the preprocessing pipeline might include other steps depending on the problem at hand, methods employed and analysis goals. Here, we aim to illustrate typical steps in a preprocessing pipeline and their intended outcome. We stress the importance of maintaining a record of all the changes made during the data preprocessing. This documentation will be valuable for reproducibility and explaining any data preprocessing steps in research or reports. 23

Furthermore, it is good practice to perform exploratory data analysis (EDA) [Tukey 1977] after performing the preprocessing steps or even between them. With the EDA, we check summary statistics, distributions, and relationships among variables to ensure the data look as expected. Lastly, let us highlight again that preprocessing is not a one-time task. As new data become available or the data collection process evolves, it is recommended to revisit and update the pipeline. 24

## Benefits of Preprocessing

The notion of data preprocessing affecting the outcome of tools such as machine learning models is widely known and accepted [Alshdaifat et al. 2021]; [Zelaya 2019]. Despite the challenges of assessing this effect, research from several domains has shown the positive impact of proper selection and application of preprocessing tools on the quality of the results [Zhu and Gao 2016] [Rinnan, Van Den Berg and Engelsen 2009] [Lee, Liong and Jemain 2017] [Joo, Choi, and Park 2000] [Alshdaifat et al. 2021]. The magnitude of these gains depends on several factors [Zhu and Gao 2016], such as the dataset characteristics (e.g. size, dimensionality, and missing values) and the problem tackled (e.g. classification, regression, and clustering), but we can find examples of 25% improvement when comparing the model results of an optimal preprocessing pipeline to the global model [Rinnan, Van Den Berg and Engelsen 2009]. Also, enhancements in the learning rate and terminal error were reported for neural network models when using adequate preprocessing [Joo, Choi, and Park 2000]. 25

Besides the benefits on the analysis performance, preprocessing contributes to a lower risk of analysis errors and biased results [Kamiran and Calders 2012] [Zelaya 2019]. Both data reduction (either by size or dimensionality) and scaling (with clarified model coefficients) enhance interpretation and facilitates understanding, especially for non-experts [Bazjanac and Kiviniemi 2007]. These tasks simplify the dataset, making it more manageable and accessible, and allow analysts to focus on relevant patterns, trends, or insights [Monroe et al. 2013]. 26

Finally, depending on the dataset size and the computational resources available, an adequate preprocessing pipeline can contribute to better computational efficiency. Data reduction techniques allow algorithms to run faster, making the overall process more efficient, and help cut storage costs by minimising the volume of data that must be stored. 27

## Challenges of Preprocessing

Several approaches to identifying the highest performance methods and their optimal combination were also developed to overcome issues due to inappropriate preprocessing [Xu et al. 2020] [Zelaya 2019] [Zhu and Gao 2016]. Nevertheless, the choice of an optimal preprocessing method is not trivial, as it depends on various aspects related to the properties of the data, goals of the analysis and the tools used to achieve these objectives [Engel et al. 2013]. 28

Furthermore, distinct preprocessing methodologies may mitigate specific artefacts while leaving some other effects behind [Rinnan, Van Den Berg and Engelsen 2009]. Hence, achieving optimal preprocessing requires integrating multiple techniques [Engel et al. 2013]. In such scenarios, determining the appropriate combination of methods and the order of their application further augments the intricacy of the problem [Engel et al. 2013]. 29

Thus, no preprocessing technique can be considered a gold standard that could be blindly applied to any data, producing satisfactory results irrespective of the data nature [Mishra et al. 2020]. Despite the efforts toward automating this process and developing more effective tools, method selection is often required to explore the available options [Engel et al. 2013]. This time-consuming exploration frequently results in a sub-optimal solution, as users can only directly explore a limited set of preprocessing techniques and their combinations [Mishra et al. 2020]. 30

For these reasons, the field remains open to advancements and novel approaches [Lee, Liong and Jemain 2017]. It still needs guidelines on the optimal selection and sequencing of preprocessing techniques, including determining when a singular technique suffices and how various techniques can be effectively integrated. 31

# The Role of Comprehensive Documentation in Ensuring Research Reproducibility

The lack of a golden preprocessing standard amplifies the relevance of a proper documentation in the digital humanities; but documentation is not limited to this stage only. Throughout this section, the term “documentation” refers to the systematic recording, detailing, and archiving of all aspects related to the research process. This process includes documenting the research design, methodologies, procedures, data collection techniques, analytical methods, results, interpretations, and conclusions. The primary purpose of documentation is to ensure transparency and reproducibility of 32

the research findings [Gorgolewski and Poldrack 2016]. It provides a comprehensive record that allows other researchers to understand, evaluate, replicate, and build upon the study.

From the reproducibility perspective, the meticulous documentation of every step in data preprocessing is paramount since they transform the raw data into a format suitable for analysis [Gibert, Sànchez–Marrè and Izquierdo 2016]; each step can significantly impact the outcomes and conclusions of the research. Moreover, detailing the parameters and configurations of the methods employed in the study is another crucial element for improving reproducibility and transparency [Haibe-Kains et al. 2020]. Parameters such as scaling factors, normalisation methods, missing data imputation techniques, and feature selection criteria can profoundly impact the data characteristics and the subsequent analytical results. By documenting these parameters and the rationale for their selection, researchers also provide insights into the preprocessing decisions, justify the chosen methodologies, and enable critical evaluation and validation of the research process. 33

Besides the parameters and configurations of the methods used, it is also essential to document the software used. Different software packages may have distinct algorithms, functionalities, and default settings that can influence the preprocessing outcomes [Chirigati et al. 2016]. Through the specification of the software version, modules, and any custom scripts or plugins, researchers enable others to replicate the research, ensuring consistency across studies. Moreover, by leveraging these documents, researchers can identify and address sources of bias, ensure consistency in data handling, and mitigate the risk of errors. 34

Although the documentation process is often associated with the procedures related to the experimental setup, it is far beyond that. It should be conducted from the first steps of the research workflow (i.e., experimental design, as seen in Figure 1). For example, creating well-documented research data and methodologies during the experimental design enhances research reproducibility and helps experts evaluate the research methodology, data quality, interpretations, and limitations. Also, it allows researchers from different disciplines to better understand and build upon existing data and findings. 35

Well-documented research is a valuable educational resource for training the next generation of scientists, researchers, and professionals. It is an example of good research practices in describing not only the research methodology and results, but also data creation and analysis. Clear documentation enables educators to incorporate real-world examples, case studies, and research findings into curricula, fostering critical thinking, research skills, and academic excellence. 36

In the literature, we can find examples of works addressing the standardisation of the documentation process. For example, the guiding FAIR (Findable, Accessible, Interoperable, Reusable) data principles [Jacobsen et al. 2020] are designed to enhance the management, sharing, and reuse of research data and digital resources. Developed by the scientific community, these principles aim to address challenges related to data discovery, accessibility, documentation, integration, and sustainability across diverse disciplines and data infrastructures [Lamprecht et al. 2020]. The FAIR principles can be summarised as follows: 37

- Findable: Data and resources should be easy to find for humans and machines. This result can be achieved by assigning persistent identifiers (such as DOIs or URLs), using standardised metadata, providing comprehensive documentation, and ensuring that data is indexed and searchable through relevant repositories, catalogues, or platforms.
- Accessible: Once located, data and resources can be retrieved and accessed with minimal impediments. Accessibility concerns providing open access, ensuring appropriate authentication and authorisation mechanisms, and offering data in standardised, machine-readable formats. Accessibility ensures that data is available for analysis, validation, and reuse.
- Interoperable: Data and resources should be interoperable, meaning they can be integrated, combined, and analysed across different systems, platforms, and disciplines. Interoperability involves using standardised formats, terminologies, ontologies, and protocols and adhering to established data models and conventions. This principle enables researchers to combine data from multiple sources, perform cross-disciplinary analyses, and derive new insights.

- Reusable: Data and resources should be reusable, which can be used for various purposes beyond their initial context or research objectives. The steps related to this principle include providing clear licensing, rights, and usage permissions, ensuring data quality and provenance, and offering comprehensive documentation, methodologies, and code. Reusability encourages transparency, reproducibility, and collaboration, allowing researchers to validate findings, build upon existing knowledge, and contribute to collective scientific progress.

38 There are also some domain-specific approaches, such as the data documentation initiative (DDI) – an international standard for documenting social, behavioural, economic, and health science data [Rasmussen and Blank 2007]. The DDI specification provides a comprehensive framework for describing the various aspects of data, including its collection, processing, distribution, and use. It aims to facilitate the interoperability, preservation, and sharing of research data by promoting standardised metadata documentation. This standard can be beneficial when describing and documenting data sources from the conceptualisation to the archival stages.

39 The DDI has facilitated the development of various tools and resources designed to support the implementation, management, and use of its metadata standards. These range from metadata editors and converters to validation tools and repositories [Blank and Rasmussen 2004]. For example, the DDI Codebook [DDI Initiative 2024a] is an XML-based structured documentation format that facilitates the creation of codebooks, data dictionaries, and questionnaires using the metadata standards. It enables researchers to document survey instruments, variables, question text, response categories, and other relevant information in a standardised format. The DDI covers all major content areas but is generally limited to descriptive narrative.

40 Another example of a tool for documenting DDI metadata is the DDI Lifecycle [DDI Initiative 2024c]. The latter is a comprehensive metadata standard that supports the documentation of the entire research data lifecycle, from conceptualisation and data collection to data processing, analysis, dissemination, and archiving. The DDI Lifecycle provides a structured framework for capturing detailed metadata about data variables, attributes, methodologies, and documentation.

41 Besides these two examples, various metadata editors and tools have been developed to facilitate the creation, editing, validation, and management of DDI metadata. These tools provide graphical user interfaces, templates, wizards, and validation checks to assist researchers, data professionals, and organisations in implementing DDI standards effectively. Some tools enable metadata transformation, conversion, and interoperability between different formats and standards. Tools such as structured data transformation language (SDTL) and extended knowledge organisation system (XKOS) support the conversion of metadata to and from DDI XML, DDI Codebook, SDMX, RDF, and other relevant formats, ensuring compatibility and consistency across data systems and platforms [DDI Initiative 2024b].

42 The DDI also complies with FAIR principles. The DDI Alliance and other organisations offer training, workshops, webinars, and educational resources to support the adoption, implementation, and best practices of DDI standards. These resources provide guidance, tutorials, case studies, and examples to assist users in understanding and applying DDI metadata standards effectively. Also, local initiatives, such as the European Open Science Cloud (EOSC) [Budroni, Claude-Burgelman, and Schouppe 2019] and the German National Research Data Infrastructure (NFDI) [Klingner 2023], are proposed to create a unified and accessible ecosystem for research data and digital resources. Both projects aim to accelerate the transition towards open science by providing researchers and institutions access to a wide range of research data, tools, services, and infrastructures.

43 Historical research can benefit from applying the FAIR framework or the DDI initiative. For example, the findable principle from FAIR could be used to standardize the metadata created to describe historical sources (e.g., author, period, location). This standardization will aid the creation of searchable databases for archival materials and digitized records. These steps also benefit research using digital methods since a standard data representation reduces the amount of data preparation for applying digital methods.

44 The accessible principle is another key concept from FAIR, which is important for historical research, particularly for its reproducibility. Accessible data means storing historical datasets in open repositories with clear access protocols. It also

requires ensuring long-term digital preservation to prevent data loss. This step helps to guarantee that the research is reproducible and verifiable and that data remains valid for future scholars.

Moreover, historians who adopt a common standardized data format to represent their data and metadata comply with the interoperable principle of FAIR. In this case, instead of using some proprietary or OS-specific data format, researchers should opt, for example, for XML, CSV, and JSON as their data format. Lastly, the reusability principle will ensure that historians build upon previous research by providing clear documentation on data sources, methodology, and provenance. 45

DDI can also be helpful for historical research in the context of standardized metadata creation. Historical research dealing with large-scale structured data can use DDI to organize the datasets, making them easier to analyse. Furthermore, consistent data description is vital for comparing sources from different times and regions. Lastly, the DDI initiative contributes to the enhancement of digital archives by supporting the integration of data into repositories. 46

Publications are another form of documenting the research. Although the amount of information that can be included in publications is limited, authors should be encouraged to provide as much information as possible on the methodology, data gathering, and pre- and post-processing methods. Including this information in the main manuscript is discouraged to avoid overextension and overcomplicated texts. However, this information can be provided as supplemental material, for example. Git repositories are also an excellent alternative for digital historians to share their code analysis of the data. For data, there are free data repositories that can also be linked to the population, allowing for easy access to the data. 47

Moreover, specialized scientific journals focusing on datasets are growing in popularity. These journals can be an excellent opportunity for publishing datasets and documenting the methodology for their creation. Documenting dataset creation is essential for ensuring research transparency, reproducibility, and data integrity. It helps other researchers understand how the data was collected, processed, and analysed, allowing for validation and reuse while minimizing errors and biases. It also provides guidance for researchers looking for solutions to creating similar datasets. Lastly, this kind of publication venue represents a vital cultural shift towards acknowledging the importance of data documentation and changing the mentality that results are the only part of the research process worth publishing. 48

## Case Study: Unveiling Historical Narratives through Topic Modelling

After advocating the relevance of preprocessing and proper documentation to research, particularly digital history, we present a case study. We apply topic modelling methodologies to unravel latent themes and subjects inherent to an extensive corpus comprising over 168,000 official correspondence of the Atlantic Portuguese Empire. The dataset was sourced from the Historical Overseas Archives of Lisbon between 1640 and 1822. 49

The data utilized was sourced from the digital repository of the *Instituto de Investigação Científica Tropical* (ACTD-IICT). This archive has published PDF files with entries from a few collections from the Portuguese Overseas Archives.[1] The dataset created contains three columns corresponding to the document identification number (internal ID used by the archive to classify the documents), the source file, which indicates from which PDF the document was extracted, and the last column is the texts extracted. Due to the nature of the source files (i.e., PDFs with machine-readable text), OCR and similar methods were not used. More information on the characteristics of this dataset can be found in the work of Błoch et al. [Błoch, Vasques Filho and Bojanowski 2022]. 50

We divide our study into two main phases. First, we explore the impact of data cleaning on the model's performance and other preprocessing decisions. Among these models, we assessed latent Dirichlet allocation (LDA) [Blei, Ng and Jordan 2003], latent semantic indexing (LSI) [Rosario 2000], and Gibbs sampling algorithm for a Dirichlet mixture model (GSDMM) [Yin and Wang 2014]. Next, we discuss the importance of understanding the characteristics of the data to select a suitable model. We also comment on issues with metrics used to assess the model's performance and determine the appropriate number of topics. 51

We selected Python version 3.11.5 to build our experimental setup and analysis, and the experiments we performed in a 52

12th Gen Intel(R) Core(TM) i7-12700H 2.30 GHz computer with 40.0 GB of RAM, HD of 1TB, NVIDIA GeForce RTX 3050, running Windows 11 Pro version 22H2. The supplementary material details the code, the results generated, and the libraries' specifications. To facilitate reproducibility, besides the code, we also specify the versions of each external library used.

Listing these specifications and using open-source multiplatform tools allows us to improve FAIR's findability and interoperability principles. To be fully compliant with FAIR, we intend to target accessibility and reusability by releasing a version of our dataset on our project website. Together with the code and documentation provided, this dataset will allow researchers to test, replicate and validate our methodology, and that data can be used for other research related to the Atlantic Portuguese Empire. 53

## Impact of Preprocessing

Although we aim to automatically leverage topic modelling models to extract significant themes from the corpus, we must first preprocess the dataset before applying these models to achieve better results. Our preprocessing pipeline consists of data cleaning and transformation, while data scaling and reduction were unnecessary. We refer to "data cleaning," as to a series of procedures to remove unnecessary text elements. These elements include special characters, parentheses and brackets, location and people's names, and dates. 54

First, we remove the text's initial and last part of the text in each record, as it only contains metadata (date, original location of the letter, and archival information) which is useful for other types of analysis but not for topic modelling — metadata do not contain semantic information relevant for our purposes. To illustrate our point, we can take the word Lisbon as an example. Several documents are sent from/to Lisbon; due to the high frequency of this work, the models could create a topic grouping all documents referencing it. In this scenario, the matter discussed in those documents could be related to military, religious, administrative or any other subject. Hence, the group based on the location is not helpful. Figure 3 shows an example text with the removed parts highlighted in yellow. 55

> "216. 1655, Abril, 24, Lisboa CONSULTA do Conselho Ultramarino ao rei D. João IV sobre a nomeação de pessoas para cargo de governador de São Tomé, por terminar o triénio de Cristóvão de Barros Rego, sendo candidatos Carlos de Nápoles, Inácio Gago da Câmara e António da Fonseca de Ornelas. AHU-São Tomé, cx. 2, doc. 115. AHU_CU_070, Cx. 2, D. 216"

**Figure 3.** Example of the text highlighting the filtered parts containing dates, locations, and the document archival information. The text can be translated as "216. 1655, Abril, 24, Lisbon CONSULTATION of the Overseas Council to King D. João IV on the appointment of people to the position of governor of São Tomé, at the end of the three-year term of Cristóvão de Barros Rego, with candidates being Carlos de Nápoles, Inácio Gago da Câmara, and António da Fonseca de Ornelas. AHU-São Tomé, cx. 2, doc. 115. AHU_CU_070, Box 2, D. 216".

Next, we employ regular expressions to remove special characters such as quotation marks, brackets, parentheses, and a string of white space characters. We also leverage a Named Entity Recognition (NER) model [Nadeau and Sekine 2007]; [Błoch, Vasques Filho and Bojanowski 2022] to remove names of people and locations present in the body of the text. As we explained before, grouping the documents by location is irrelevant to knowing the actual topic discussed in them. Analogously, we remove people's names as, besides the lack of meaningful information, grouping by names could lead to misinterpretations due to homonyms (e.g., grouping all documents referencing "Maria" would aggregate both commoners and a queen). 56

Together with the NER model, we could employ other methods to identify and remove people and locations from the text. For instance, we could filter the text by creating a list of location names from Portugal and the colonies and using libraries available on the web that provide a list of people's common names per location and language. The main drawback of this approach is the continuous effort of feeding these lists with locations and people not previously captured. Thus, we opted for the NER model as, instead of searching for specific words characterised as location names, the model learns from examples of the patterns in the text that indicate that a noun is a location name. The 57

model's performance depends on the quantity and quality of the examples provided during its training phase.

Several NER libraries and pre-trained models are available for our use, provided they work with texts in Portuguese. As we are using Python and due to the availability of resources, we chose to experiment with the spaCy library [Srinivasa-Desikan 2018] and its pretrained models. Nevertheless, the amount of pre-trained models and the best-performing ones are usually for texts in English. In our case, although pre-trained models in Portuguese are available, they are based on a modern version of Portuguese, usually from news and social media data sources. Hence, their accuracy with historical texts is impacted. Figure 4 shows an example of entities recognised in one sample of our dataset by the pre-trained NER Portuguese model [spaCy 2024]. 58

Although this model can reach accuracy levels up to 89% because it was trained using news sources of modern Portuguese, the accuracy in our data can be much lower for documents from the 17th until the beginning of the 19th century [Błoch, Vasques Filho and Bojanowski 2022]. As seen in Figure 4, one of the problems of this model is the long names that are identified as multiple entities. Another drawback of using a pre-trained model is the limitation of the type of entities it can recognise. We can see in Figure 4 that titles and occupations (e.g., king, prince, queen, viceroy, duke, general) were not tagged as this model was not trained to identify them. 59

Due to these entity identification issues related to the specificities of our data, we trained a custom model by providing examples from our corpus. The training process is described in our previous work [Błoch, Vasques Filho and Bojanowski 2022], and this new model achieved an overall accuracy of 93,1%. The main drawback of a custom NER model is the time required to select, annotate, and train it. Still, compared to general pre-trained models, it has several advantages, such as enhanced performance in the problem tackled and the definition of additional entities we wish to identify. For example, besides identifying locations, our custom model can distinguish between masculine and feminine names and identify titles, occupations and different types of organisations (e.g., military and religious). 60

Figure 5 presents the entities recognised in the text shown in Figure 4. We can see in Figure 5 a better identification of people's names when compared to the pre-trained model. We can also observe that titles and occupations were tagged in the model trained in our data. Having a better NER model in our preprocessing pipeline means that fewer irrelevant entities will reach the topic modelling, making the topics less noisy. Omitting the usage of NER to clean the text or even not disclosing the usage of a custom model impacts the results that can be achieved in the topic modelling stage. Figure 6 shows examples of the impact of not using NER models to filter out irrelevant text elements on the LDA performance. We can see in Figure 6(B) the presence of location names such as "baía", "espírito_santo" and "maranhao". Not all documents in those groups are from this location, which can lead to a false interpretation that the other keywords in those groups are related to those places. 61

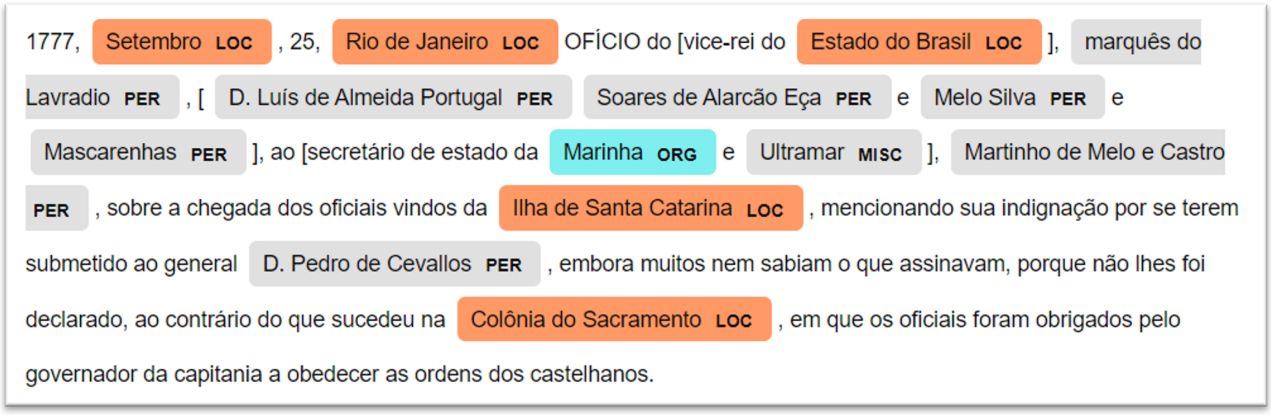


**Figure 4.** Example of entities tagged by the pre-trained model. The text can be translated as "1777, September 25, Rio de Janeiro OFFICIAL LETTER of the [Viceroy of the State of Brazil], Marquis of Lavradio, [D. Luís de Almeida Portugal Soares de Alarcão Eça e Melo Silva e Mascarenhas], to the [secretary of state for the Navy and Overseas], Martinho de Melo e Castro, about the arrival of the officers from Santa Catarina Island, mentioning his indignation at having submitted to General D. Pedro de Cevallos, although many did not even know what they were signing, because it was not declared to them, unlike what happened in the Colony of Sacramento, in which the governor of the captaincy forced the officers to obey the orders of the Castilians." We can see in this example that the Marquis of Lavradio's name was recognised as two different entities.

62

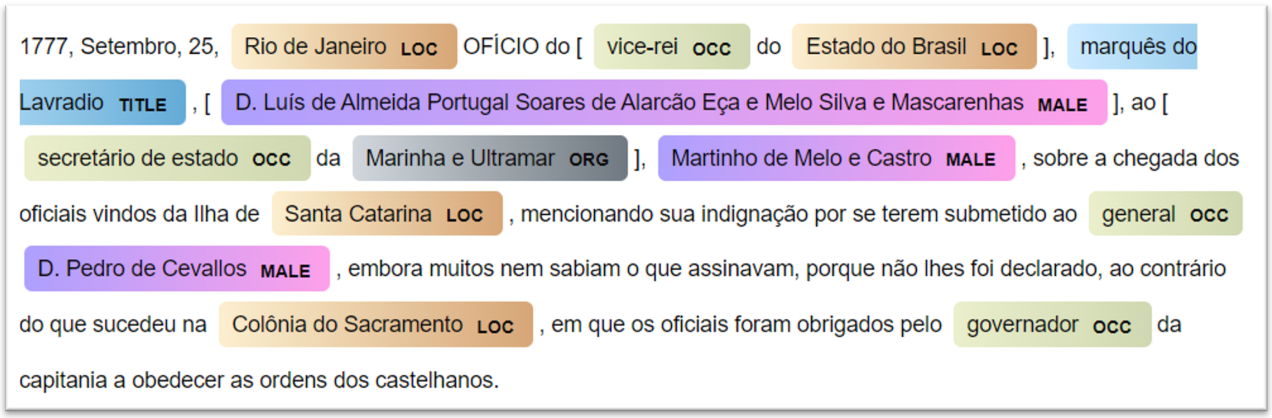


**Figure 5.** Example of entities tagged by the custom-trained model. The text can be translated as "1777, September 25, Rio de Janeiro OFFICIAL LETTER of the [Viceroy of the State of Brazil], Marquis ofLavradio, [D. Luís de Almeida Portugal Soares de Alarcão Eça e Melo Silva e Mascarenhas], to the [secretary of state for the Navy and Overseas], Martinho de Melo e Castro, about the arrival of the officers from Santa Catarina Island, mentioning his indignation at having submitted to General D. Pedro de Cevallos, although many did not even know what they were signing, because it was not declared to them, unlike what happened in the Colony of Sacramento, in which the governor of the captaincy forced the officers to obey the orders of the Castilians." Note the correct identification of people's names and recognition of more categories: gender, title, and occupation.

63

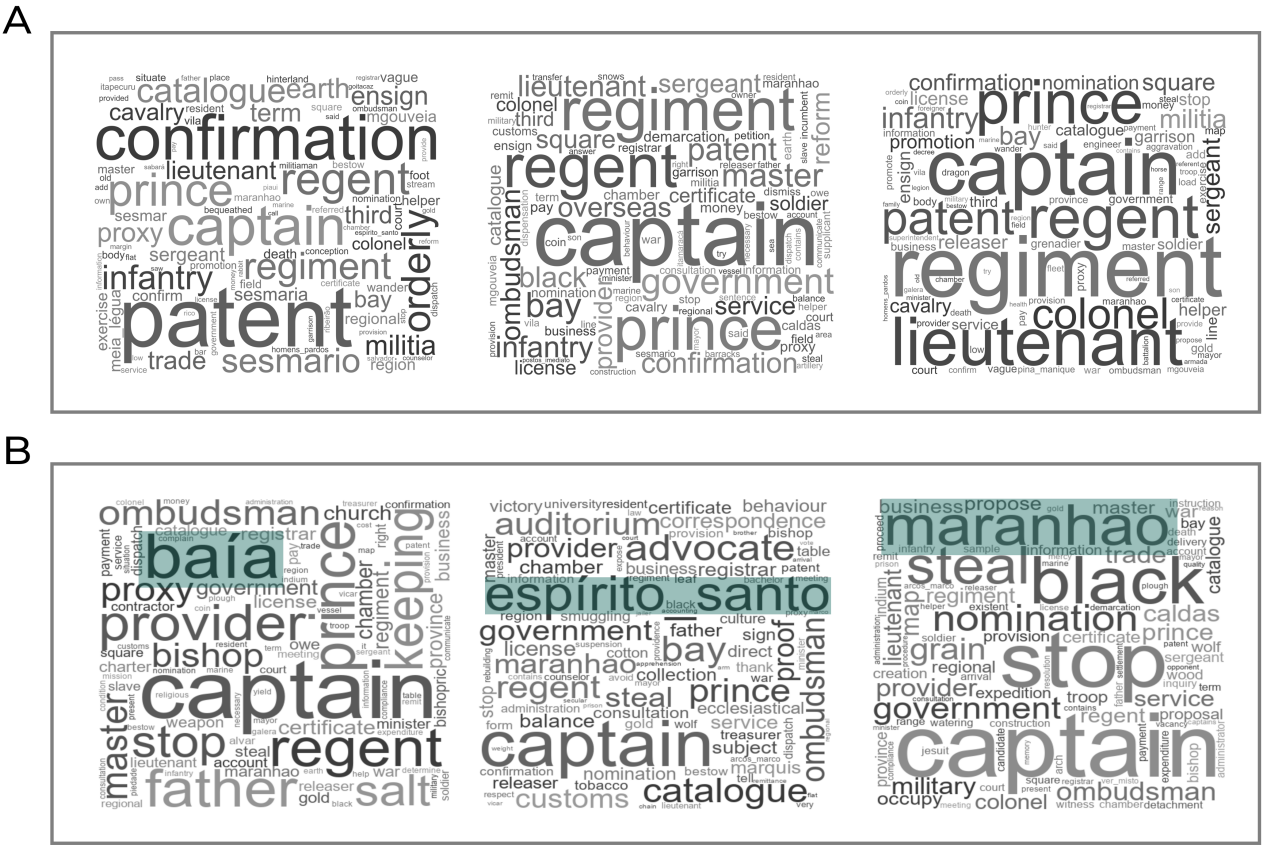


**Figure 6.** Examples of the impact of using NER to filter specific entities on the LDA's performance. (A) results with entities removed and (B) results without removing them. Note in (B) the presence of words irrelevant to characterise the topic, such as "baía", "maranhao" and "espírito_santo" which refers to locations.

The results in Figure 6 reveal key topics linked to colonial administration, trade networks, the transatlantic slave trade, religious missions, military conflicts, and cultural exchanges. Without the noise of the location names, the themes can help historians reconstruct governance strategies, economic dependencies, forced labor systems, religious influence, and hybridized cultural practices across former Portuguese territories. 64

In the last step of our cleaning procedure, we remove stopwords, which are common words often excluded from text-processing applications, such as search engines and NLP tasks. These words are excluded because they typically do not contribute much to a text's overall meaning or interpretation. They appear so frequently in texts that they do not provide value when analysing or understanding topics' content. The complete list of stopwords is in the supplementary material. Among the words in this list, we have weekdays, months, interjects, adverbs, and other words not filtered in the previous steps. 65

Figure 7 illustrates a few examples of the topics identified by the LDA model when the stopwords are removed (A) and when they are not (B). We can see in Figure 7 that examples of topics obtained without removing stopwords are dominated by these words that contribute little to the understanding of the main subject of the groups. Figure 7 (A) highlights themes related to military and administrative structures within the Portuguese Empire. Words such as "captain", r"egiment", "patent", and c"onfirmation" indicate a focus on military organization, ranks, and official appointments. These terms could be a reflection of how Portugal managed its colonial territories through a hierarchical system of governance and defense, essential for maintaining control over vast regions. The presence of terms like "prince" and "regent" further indicates connections to royal authority, underscoring the centralized nature of decision-making within the Empire. 66

In contrast, the noise illustrated in Figure 7 (B) introduced by keeping the document type clouds the analysis and can lead to different interpretations. Terms such as "attachment", "secretary", and "letter" suggest a focus on record-keeping 67

and official correspondence and emphasize the bureaucratic aspect of administration. These terms indicate a focus on record-keeping and official correspondence rather than the actual content of the documents.

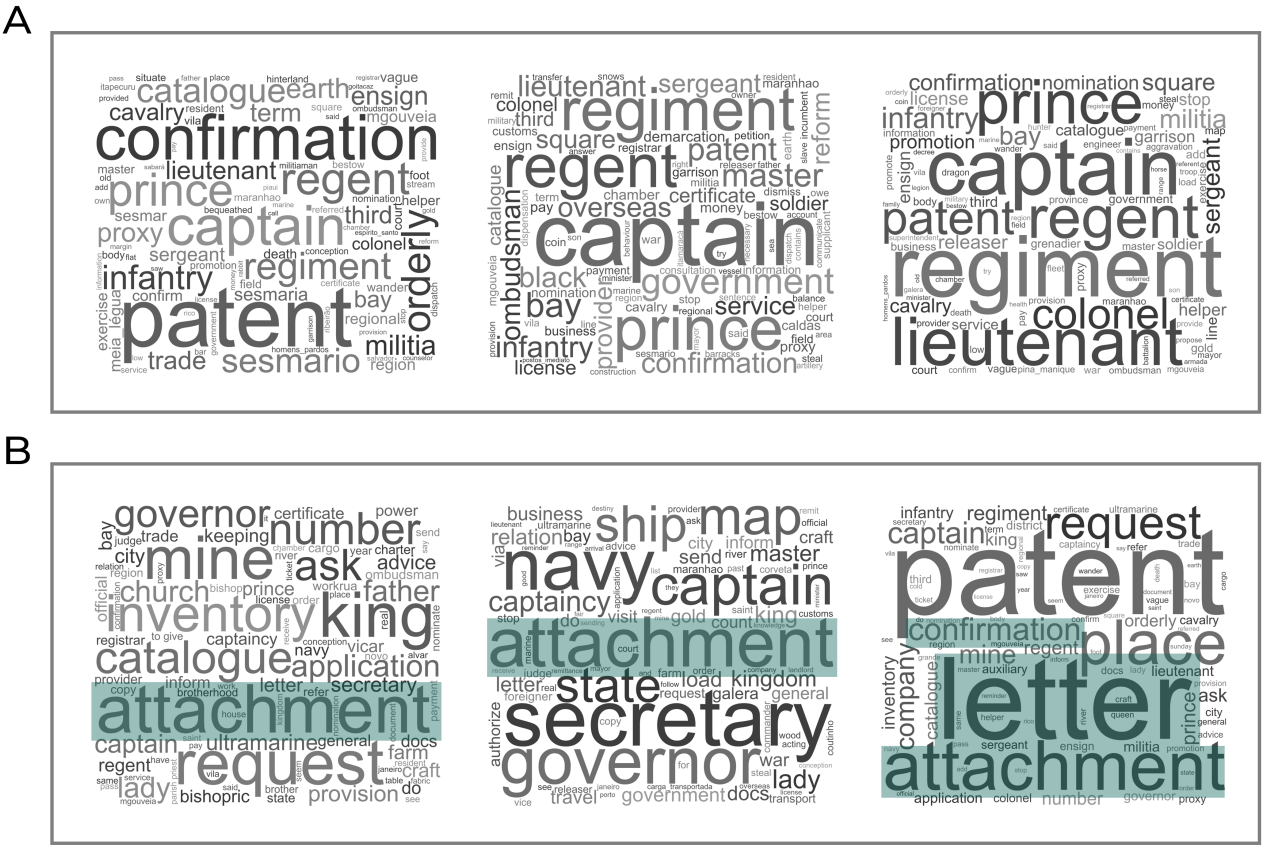


**Figure 7.** Examples of the impact of filtering stopwords on the LDA's performance. (A) results with stopwords removed and (B) results without removing them. Note in (B) the presence of words irrelevant to characterise the topic, such as "letter", "attachment", and "confirmation".

These are chunks of *n* consecutive items, such as words, in a sequence of text or speech. They help analyse and 69 68
model the structure and patterns within language, facilitating tasks like prediction, text generation, and information retrieval. In this work, we leverage the Gensim library [Srinivasa-Desikan 2018] to generate the set of unigrams and bigrams. The creation of n-grams and lemmatisation can sometimes be classified as feature extraction or feature engineering, depending on the research [Poulston, Stevenson and Bontcheva 2015] . Regardless of the classification, it is well-established that topic models can benefit from these procedures [Nikolenko, Koltcov and Koltsova 2017] [Skorkovská 2012][Poulston, Stevenson and Bontcheva 2015].

As mentioned, a few steps in our preprocessing pipeline are classified as data transformation. These procedures 70
transform the original data into other structures or change the word form. For example, we apply lemmatisation to group the different inflected forms of a word, reducing them to their base or root form — called the lemma [Korenius et al. 2004] — to normalise and simplify the text for analysis and understanding. Another example of data transformation applied is the creation of n-grams [Sidorov et al. 2014].

The main benefits of the usage of n-grams are related to capturing collocations and improving topic coherence. N- 71
grams, especially bigrams and trigrams, allow the model to identify and group together frequently appearing phrases. Also, single words might have multiple meanings, and by employing n-grams, the model can differentiate between these meanings and create more focused topics. This representation can be particularly valuable if your topics are defined by specific terminology, such as "Espírito Santo" or "Rio de Janeiro", both location names. However, without using n-grams, the model would interpret these terms in isolation, leading to inaccurate evaluations of their meaning. For example, "Espírito" and "Santo" (i.e., from Portuguese "Holy" and "Spirit") could be linked to documents discussing religious

matters. While “Rio” and “Janeiro” (i.e. from Portuguese “River” and “January”) could be interpreted as a location and a date.

Our pipeline’s last data transformation method is TF-IDF (term frequency-inverse document frequency) — a technique for weighing the importance of words in a document collection. As the name suggests, it comprises two main components: term frequency (TF) and inverse document frequency (IDF). The first gauges how often a word appears in a document relative to its total word count. Words that appear frequently are more important for that particular document. In contrast, IDF estimates how often a word appears across all documents in the collection. The idea behind IDF is that words appearing in many documents are less informative because they don’t distinguish the specific content of a document. 72

By combining these factors, TF-IDF gives higher weights to frequent words within a document but rare across the entire collection. This weighting scheme helps identify the keywords that best characterise the specific topics covered in that document. TF-IDF helps further filter out irrelevant words and emphasise the most informative terms to uncover the underlying themes, leading to more accurate and meaningful topic discovery. 73

We must stress that the adopted preprocessing pipeline was designed to meet the needs of our data, tools, and project goals. Hence, we are not presenting it as a standard process for any topic modelling task on the historical corpus. Instead, by introducing this case study, we aim to provide examples of the preprocessing pipelines, modelling decisions and the impact of each step on the model’s result. 74

Besides preprocessing, other factors such as model selection and parametrisation also strongly impact results. In the next section, we will explore how we selected the models and parameters and assessed their performance. 75

## Model Selection, Parametrisation, and Evaluation

Model selection is challenging since it depends on the characteristics of the data, the problem’s needs, and the resources available. For instance, the length of the text is an important aspect that should be considered while selecting a topic model. Traditional approaches, such as the LDA model, might perform poorly in a corpus of short texts compared to techniques developed specifically for these texts. 76

Figure 8 depicts the distribution of text lengths in our corpus. The original distribution is long-tailed, but for visualisation purposes, we limited the range of the x−axis to the interval [0,200] in (A) and [0,1250] in (B). The truncated data still represent more than 95% of the dataset’s entries. We can see that the average text size in our dataset is 59 words or 361 characters, comparable to datasets of Twitter messages. 77

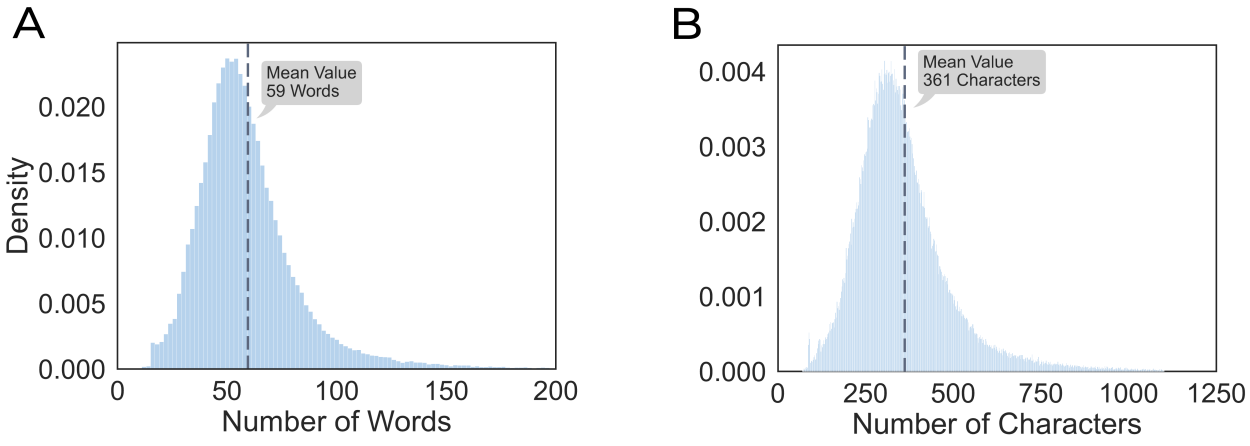


**Figure 8.** Distribution of the text size in the dataset by number of words (A) and number of characters (B). The average text size in the data is 59 words or 361 characters, comparable to the traditional 280-character limit size of a tweet.

Because our texts are short on average, they were ready for topic modelling right after the preprocessing stage mentioned before. That said, for a corpus of long texts, such as book chapters, it can be required to split the text into chunks that will feed the models separately. The primary reasons behind this practice are twofold. First, topic modelling algorithms can be computationally expensive with long texts, and splitting the text into smaller chunks helps decrease the computational complexity of processing the texts. Second, reducing the text size can positively impact the topic coherence (i.e., how well the words within a topic relate to each other semantically). This impact is linked to the possible jumps between different ideas or themes in long texts.

It is necessary to note that the solution to the text length problem is more complex than splitting the text into the smallest chunk size possible, and there is no one-size-fits-all convention. Using very small chunks of text might not provide enough context for accurate topic identification. Hence, when dealing with a corpora of lengthy texts, it is fundamental to perform preliminary results to assess the impact of the text chunk sizes on the topic modelling performance.

In our case, this assessment was unnecessary due to the short nature of our text sizes. However, to test the hypotheses that text size can impact the model's performance and that traditional approaches, such as LDA and LSI, might not perform as well as other models developed to work with short texts, such as the GSDMM, we will perform experiments comparing the performance of these three models on our data using the same preprocessing pipeline. Before presenting the results, we will discuss the model's evaluation methods.

Different approaches to evaluating the quality of the solutions generated can be based on analysing the relevant terms in each topic, distance maps, and metrics. Figure 9 shows examples of these three categories. So far, we have presented examples of topics using word clouds (i.e. representing the most relevant terms in each topic) and visually analysing them. In this approach, we can employ two primary quality evaluation criteria: the number of words overlapping in different groups and the possibility to identify a group's dominant subject. Hence, good solutions are described as well-defined groups with little to no overlapping terms.

Examples of results for LSI, GSDMM, and LDA can be seen, respectively, in Figures 10, 11, and 12. We defined that the models should return 20 groups to assess their capabilities of splitting the corpus into a relatively high number of topics. This approach aimed to obtain several groups with more specific topics rather than a few groups with broader topics. Notably, we also investigated the relationship between the number of topics and the quality of the results, which we will discuss later.

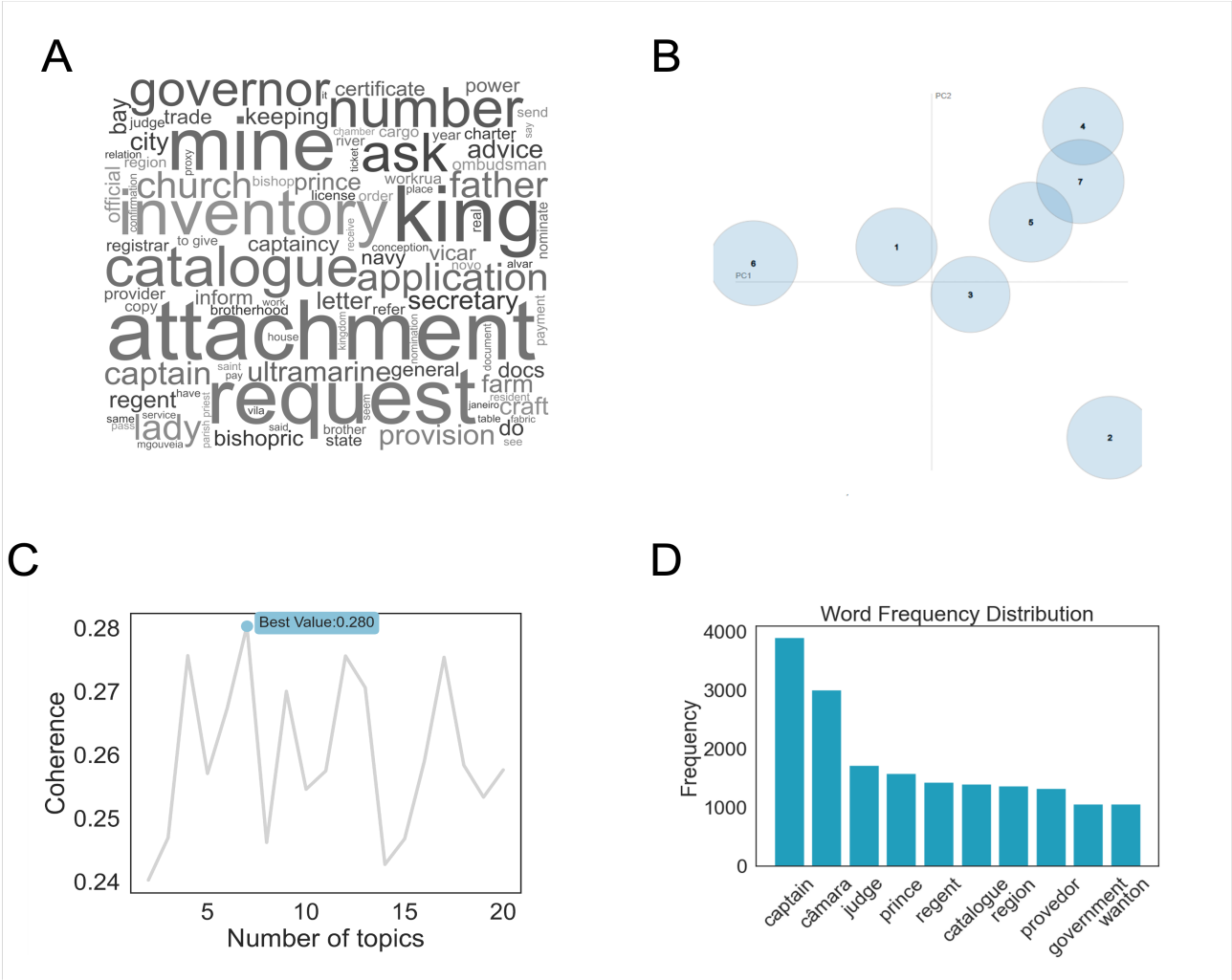


**Figure 9.** Common forms of assessing the topic modelling results. (A) top relevant terms in each topic, (B) distance map, and (C) metrics. Here, we present the word cloud to represent the most relevant works in a topic, (B) depicts the plot of a distance map based on multidimensional scaling, (C) shows an example of the Coherence metrics being applied to assess the results of LDA for a different number of topics, and (D) frequency distribution of the words in a topic.

Utilising our evaluation criteria for the word clouds, we can see in Figure 10 that the LSI results were the only model that failed to meet both conditions. There are a large number of overlapping terms across groups, such as “catalogue”, “confirmation”, “bay”, “patent”, and “consultation”. Also, it is difficult to determine the relationship between the terms inside each group. However, Figure 11 shows that the GSDMM presented the best results with minimal overlapping relevant words and relatively cohesive topics, while the LDA model had the second-best groups. Though a few terms overlap between different groups in the LDA’s results, Figure 12 indicates that the main issue is the difficulty of interpreting each group’s main topic. 85 84

Another approach is to use metrics to quantify the performance of the models. The literature offers various metrics to evaluate and support the model selection process [Meaney et al. 2023]. However, it is essential to notice that a significant percentage of these metrics require a ground truth (i.e., a dataset with the text and their expected groups), which is not always possible. For example, the dataset used in this work does not have the expected topical classification, which restricts the metrics that can be applied. 86

In our experiments, we used the model coherence metric [Meaney et al. 2023] to measure the interpretability and semantic coherence of the topics produced. This metric outputs values from 0 to 1, in which a higher coherence score typically indicates that words characterising a topic co-occur frequently across the documents in the corpus. Figure 13 shows an application of the coherence metric to evaluate LDA models with different numbers of topics. The idea behind the experiment depicted in Figure 13 is to use the metric to choose the best number of topics. In this approach, the best 87

number of topics will be the one that maximises the metric (e.g., the highest value for the coherence metric). It is also possible to use multiple model evaluation strategies and, for example, combine the use of metrics with the inspections of the top relevant words per topic.

Figures 14 and 15 show the characteristics of the groups found for seven (best coherence score) and 20 topics (maximum number tested). We can see in Figure 14 better separations of topics compared to Figure 15. The overlapping of words is also visible in the samples of topics provided in each figure. These samples represent groups in distinct quadrants in the plot; the amount of overlapping of essential words (i.e., large ones) is considerably higher in Figure 15 than in Figure 14.

As seen in Figure 13, these model evaluation metrics can also be employed to select the best parametrisation for a given model. In this case, an exhaustive or heuristic search algorithm can find the best or near-best parameters to optimise a given metric. For example, a hyperparameter tuning approach based on a simple exhaustive search can be used to find the best configuration that maximises the model coherence score.

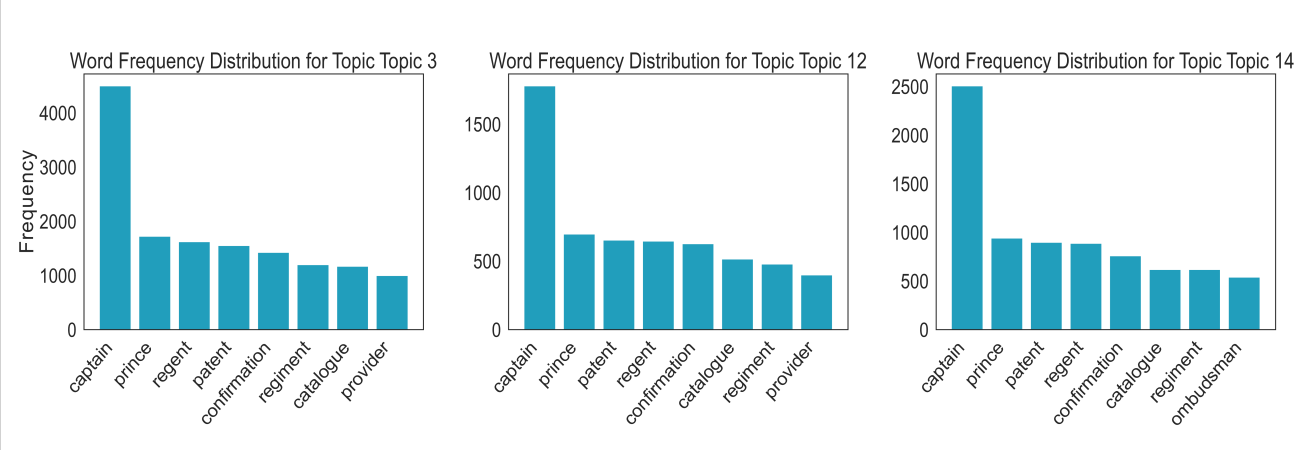


**Figure 10.** Example of the top 10 most frequent words in three groups returned by the LSI algorithm. The complete list is available in the repository. Observe the amount of overlap between different groups, which is a sign of improper splitting which is an indication of not well-defined groups.

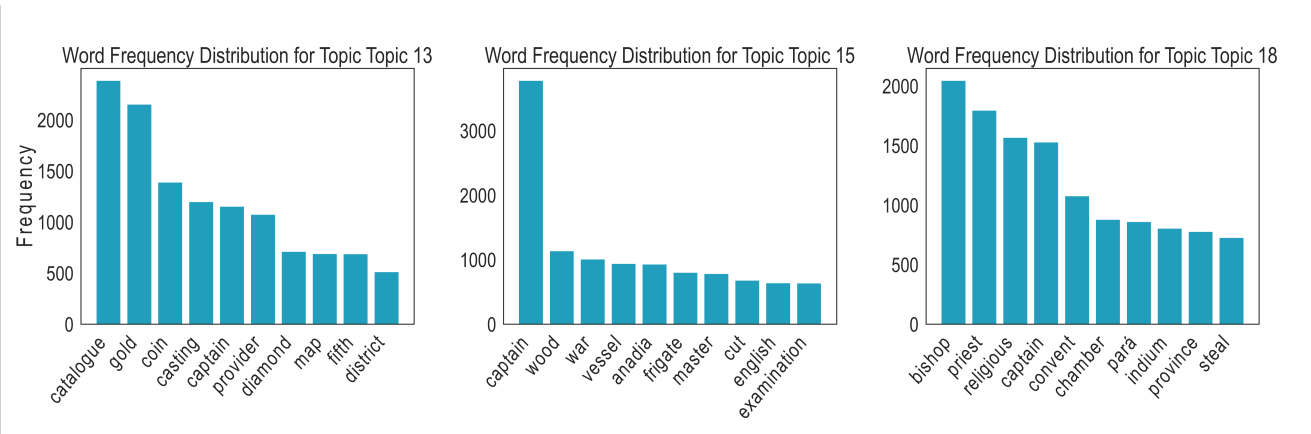


**Figure 11.** Example of the top 10 most frequent words in three groups returned by the GSDMM algorithm. The complete list is available in the repository. Here, we can see a low overlap between words in different groups, and it is easier to interpret most groups' general meaning. For example, group 18 seems to have religious topics, group 15 is connected to military issues, and 13 has words related to commerce and trading.

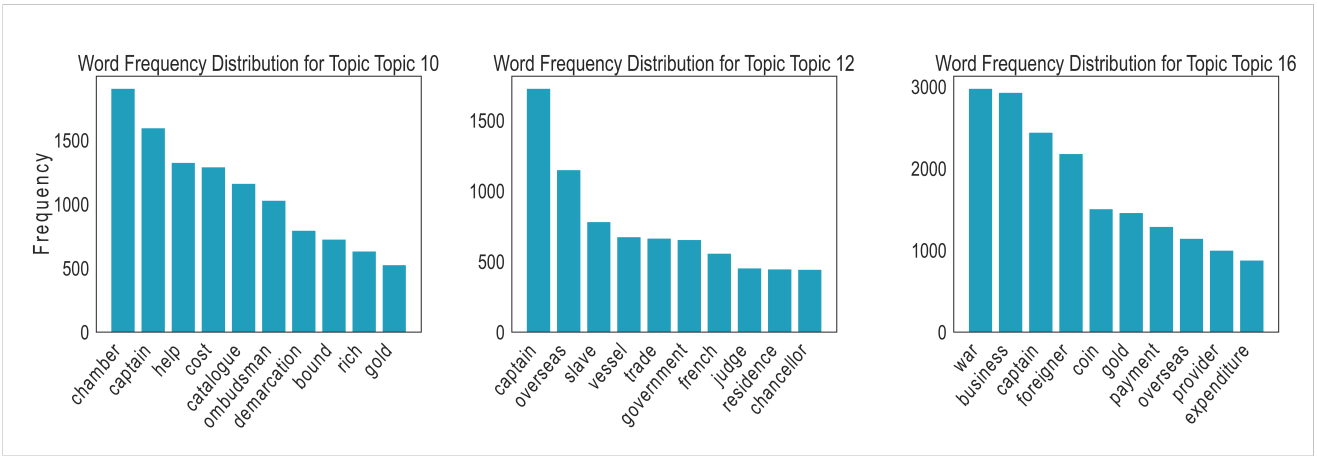


**Figure 12.** Example of the top 10 most frequent words in three groups returned by the LDA algorithm. The complete list is available in the repository. Although there is not much overlap between the words in these groups, extracting the general meaning of the groups is still difficult.

92

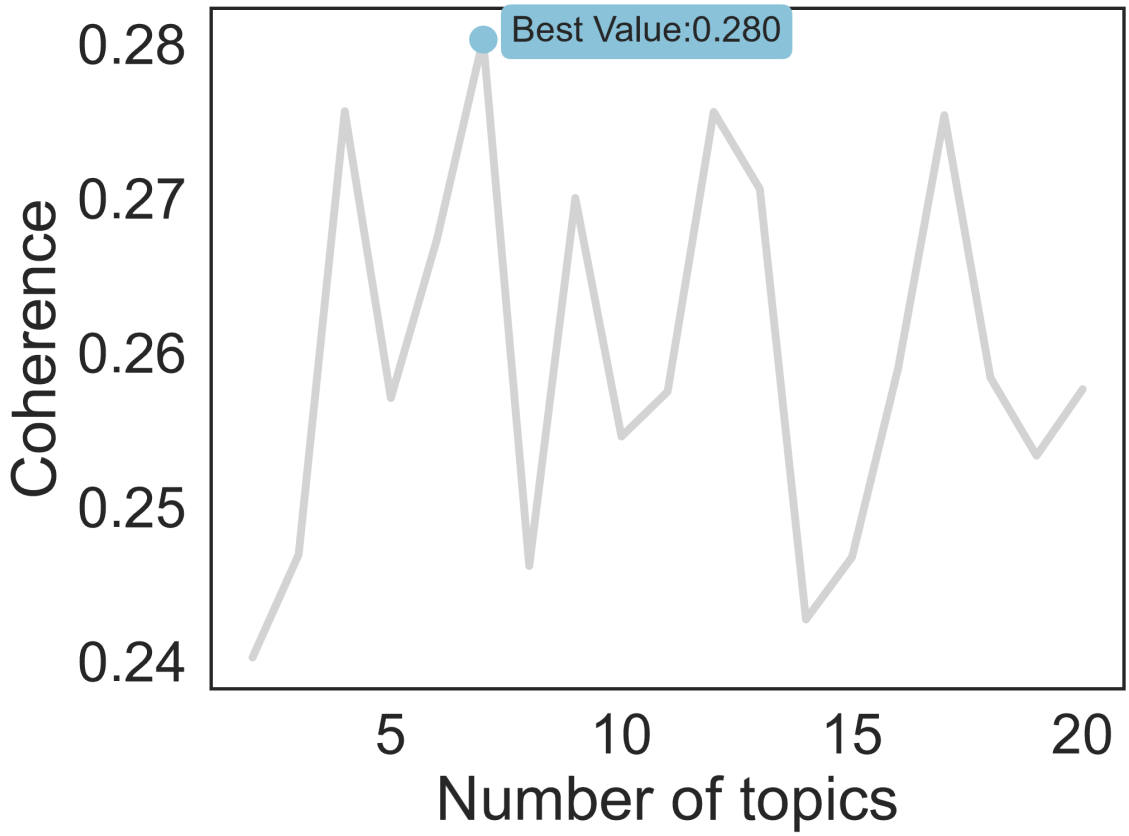


**Figure 13.** Example of application of the coherence metric to evaluate the performance of the LDA with different parametrisation. In this case, we analysed the impact of the number of topics on the model performance. According to this metric, the best performance was achieved for seven topics with a coherence score of 0.280.

It is worth noting, however, that this approach has potential issues. First, some metrics do not have a limited range for 94 93 their output value, so interpreting and comparing the results to other research can be challenging. Second, the interpretation of the metric results is challenging even when it is restricted to an interval. For example, a 0.5 coherence score might be good for one dataset and not enough for another. Lastly, different models have different optimal numbers of topics that maximise the coherence score, as shown in Figure 16. For all these reasons, it is not recommended to

rely on a single method for model evaluation.

Although metrics can guide defining the number of topics and identifying overlapping groups, determining the criteria for a satisfactory topical classification for a set of documents must also satisfy the interpretability requirement. Ultimately, the division of the groups should be done in a fashion that allows researchers to understand and determine the main topic for each group. Ensuring the interpretability of the results is an iterative process that relies on the synergy of automation tools and human expertise. This process shapes the preprocessing pipeline, and the selection of methods used for extracting information from the data. Based on our experiments, the GSDMM was the algorithm that provided better results, with clearer topical groups compared to the LSA and LDI. The GSDMM's superiority over the LDA and LSI can be attributed to its design to work with short texts such as the ones in our dataset. LDA and LSI could struggle to find better topics due to the small text size, which provided reduced context about the main topic of the text. 95

Lastly, the examples provided in this section are not meant to serve as a universal preprocessing guideline as such a thing does not exist. It also does not show all the possible points to be considered when documenting research. Our goal is to provide an example of how we designed our preprocessing methodology to best suit the requirements of our problem (e.g., extracting information from a large historical corpus) and the specificities of the methods we planned to use (e.g., topic modelling algorithms). We also wanted to showcase how the lack of transparency in documenting the preprocessing can make research nearly impossible to reproduce and lead to different conclusions. For example, if we were to say only that we used the NER model to remove names of people and locations before applying topic modelling, the results would differ significantly from those published, given the proven accuracy limitations of general-purpose NER models for historical texts in Portuguese. 96

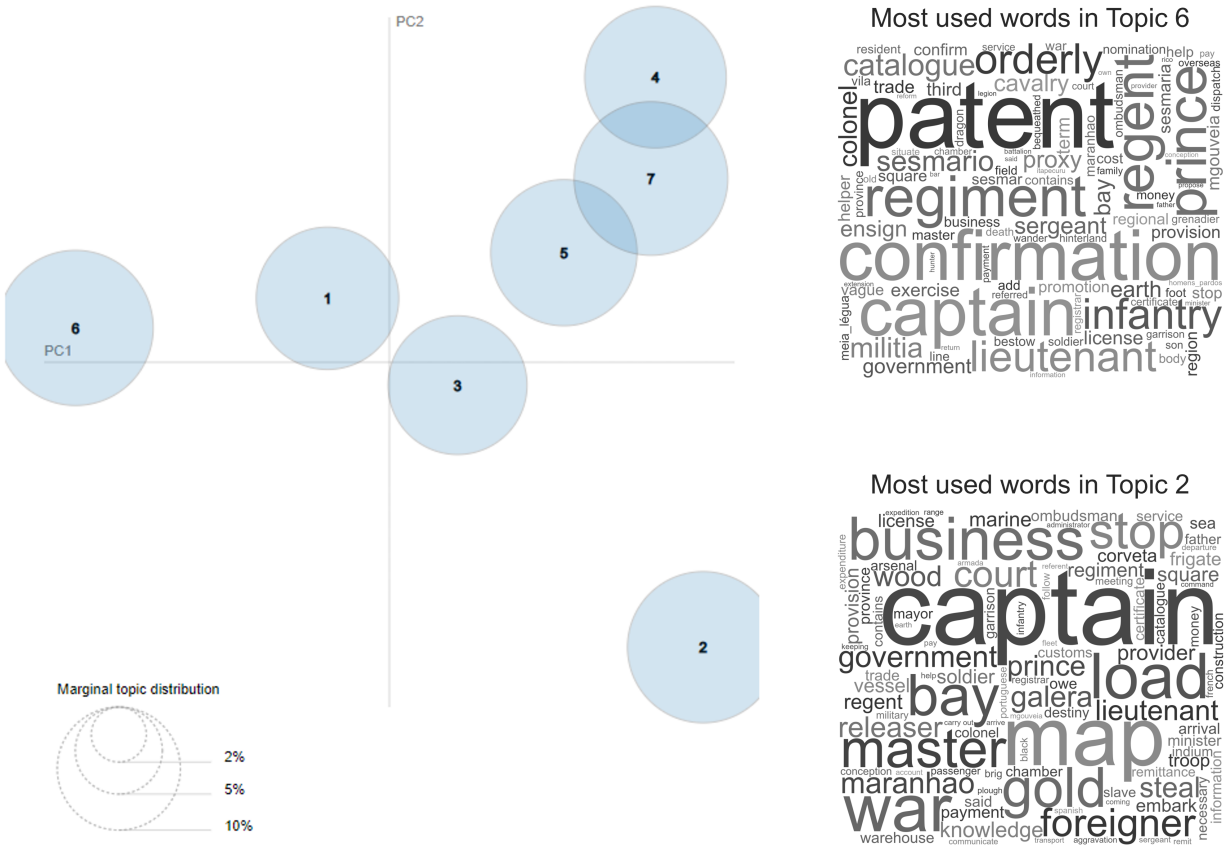


**Figure 14.** Intertopic Distance Map (IDM) for the LDA model with seven groups (maximum coherence score). The word clouds depict examples of the most prominent words in two topics distant from each other in the IDM plot.

97

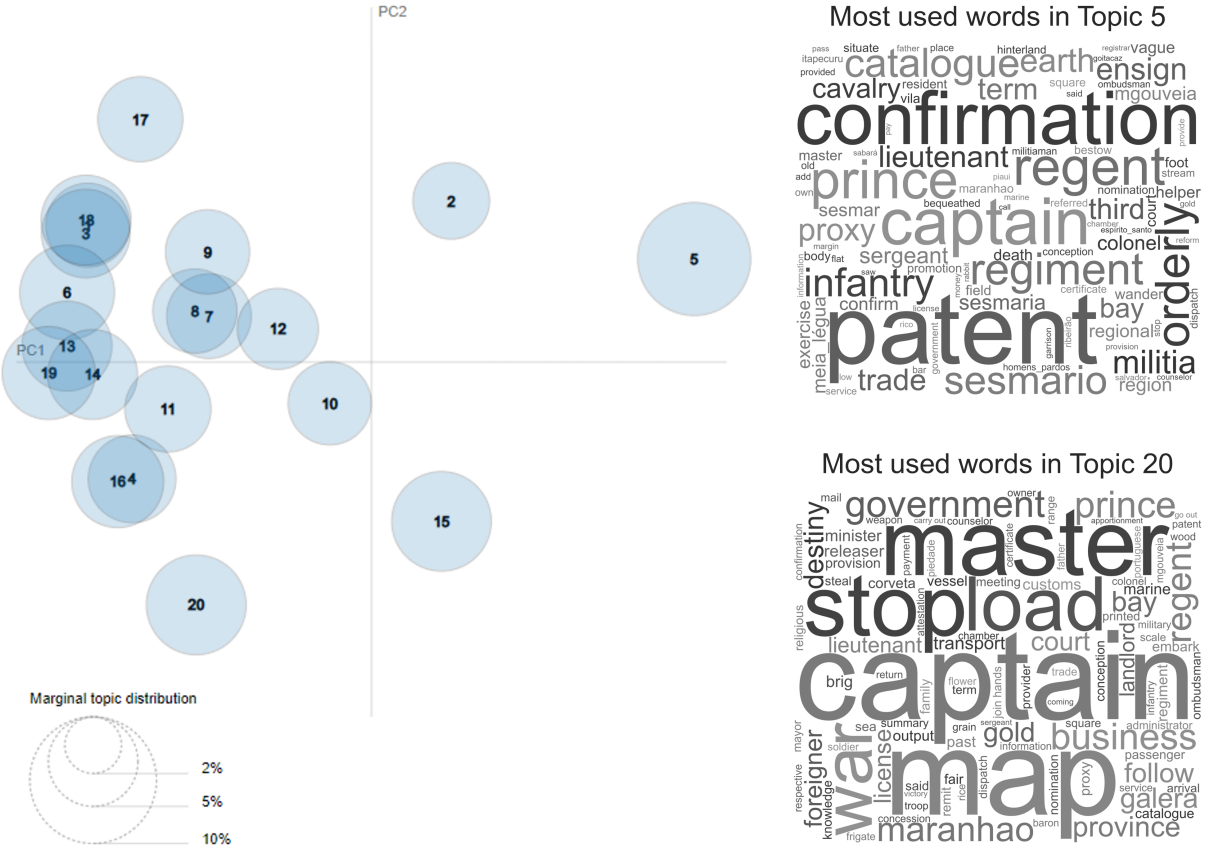


**Figure 15.** Intertopic Distance Map (IDM) for the LDA model with twenty groups (maximum number tested). The word clouds depict examples of the most prominent words in two topics distant from each other in the IDM plot. Note the high number of cluster overlaps in the IMD and words in the word clouds.

98

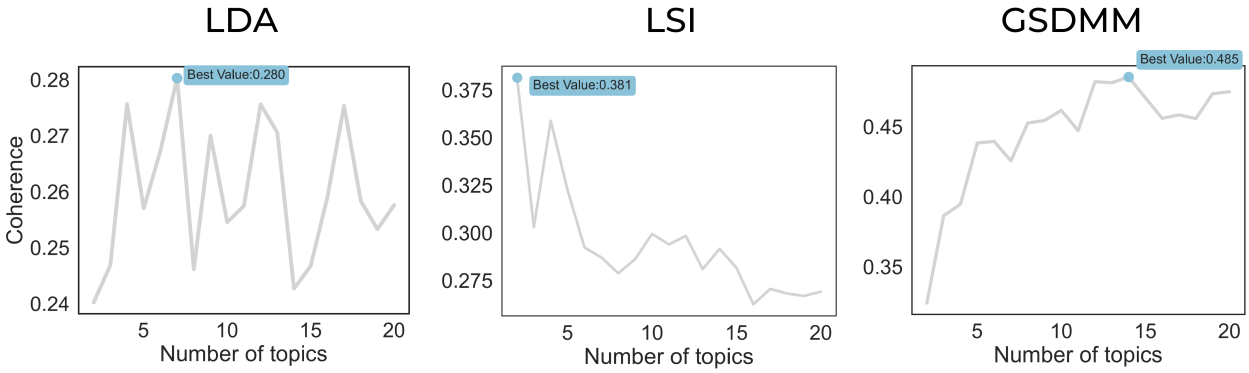


**Figure 16.** Best coherence score for the LDA, LSI, and GSDMM. Note the difference not only in the metric value but also in their corresponding number of topics

99

## Conclusion: Reproducibility Challenges in Digital History

100

At a time when large amounts of historical data are increasingly being digitised, the quality and reliability of the results depend on the detailed preprocessing of this digital information. From archival records to datasets of digitised manuscripts, best practices of careful data preparation to produce optimal historical analyses are critical. In digital historical research, the effectiveness of subsequent analyses and interpretations is closely linked to the careful preparation and refinement of historical data, making data preprocessing a critical foundation for high-quality outcomes

in digital historical research.

In this paper, we have explored the critical role preprocessing and proper documentation have in digital humanities research, using topic modelling as a case study. These often-neglected tasks compose the pipeline of digital research, profoundly affecting its reproducibility and transparency: Through our experiments with several topic models, we have demonstrated how variations in preprocessing techniques and the absence of comprehensive documentation can impair the reliability of research outcomes. By doing so, we discuss how important it is for researchers in the digital humanities to carefully consider and document all the preprocessing steps, including the reasons for each decision, the specific tools, libraries and parameters that were used, and any changes that were made to the original data. Clear and comprehensive documentation enables other researchers to understand, criticize and build on one's own work. The lack of a golden standard due to the richness and diversity of historical sources makes these tasks even more valuable – distinct datasets and research objectives have their own particularities and need to be approached in different ways. 101

Improving on methods, standards related to preprocessing have both narrow technical and a broader methodological/epistemological side. On the technical side, we hope that more aware and transparent pre-processing practices should lead to developing and adoption of more and more widely spread standards. Not only among researchers, but also affecting archival practices. In particular, we imagine that archives and archivists can prioritize the creation of FAIR-compliant datasets, enriching digital historical research. On the broader methodological side, more transparent preprocessing decisions facilitate the critical discussion of the decisions and can further disciplinary debates about biases, assumptions and limitations embedded in those decisions, as well as in the computational tools employed. Surfacing such phenomena will ensure a more thoughtful and responsible historical scholarship. 102

To close, it is also useful to realize that the recommendations provided in our article coincide with changes in how history and digital history are practiced. The main difference between traditional and digital history is the shift from individual engagement with the sources to a fully collaborative and interdisciplinary approach. It is no longer a solitary initiative but requires teamwork and the integration of different tools and methods. The decisions made in the preprocessing phase, before the data is visualized, have a significant impact on the results. For this reason, digital history is not only about the final result, but also about building a shared knowledge of the basic steps between the source collection and its interpretation that strengthens the commitment to reproducibility. While every historian is trained in methodologies of working with source materials, these are not usually shared. Contemporary humanities moved toward "intersubjectivity" and transparency of the research process improving on reproducibility and results transparency in digital history research must focus on collaboration instead of data preprocessing standards. 103

## Data Availability

A version of the dataset used in this research can be accessed on Zenodo (https://doi.org/10.5281/zenodo.15766967). 104

## Code Availability

Scripts and Notebooks in Python with our analyses and to reproduce the results in this paper were archived with Zenodo (https://doi.org/10.5281/zenodo.15090621). 105

## Acknowledgments

This research is supported by the Polish National Science Centre (Narodowe Centrum Nauki - NCN) grant number 2022/45/B/HS3/00473, project: "Imperial Commoners of Brazil and West Africa (1640-1822): global history from a correspondence network perspective." The funders had no role in study design, data collection and analysis, decision to publish, or manuscript preparation. 106

## Author Contributions Statement

CS performed the analysis. MB, DVF and AB gathered the data. CS, MB, DVF and AB designed the analysis. All authors discussed the results and contributed to the final manuscript. 107

# Competing Interests Statement

The authors declare no competing interests.



## Notes

[1] Arquivo Científico Tropical: https://actd.iict.pt/. Collection “Fundos e Coleções no Arquivo Histórico Ultramarino”: https://actd.iict.pt/community/actd:AHUF

## Works Cited


**Alshdaifat et al. 2021** Alshdaifat, E.A. et al. (2021) “The effect of preprocessing techniques, applied to numeric features, on classification algorithms’ performance”, *Data*, 6(2), p.11. https://doi.org/10.3390/data6020011

**Atunes 2021** Antunes, A.P. (2021). “Social Network Analysis in the History of Sciences: Visualising Sociability in Scientific Expeditions with Gephi”, *Publicaciones de la Asociación Argentina de Humanidades Digitales.* https://n2t.net/ark:/13683/ehed/GFz.

**Aydelotte 1963** Aydelotte, W.O. (1963) “Voting Patterns in the British House of Commons in the 1840s”, *Comparative Studies in Society and History*, 5(2), pp.134-163. http://doi.org/10.1017/S0010417500001596.

**Bazjanac and Kiviniemi 2007** Bazjanac, V. and Kiviniemi, A. (2007) “Reduction, simplification, translation and interpretation in the exchange of model data”, *Cib W*, 78, pp.163-168.

**Bazyer 1995** Baxter, M.J. (1995) “Standardization and transformation in principal component analysis, with applications to archaeometry”, *Journal of the Royal Statistical Society Series C: Applied Statistics*, 44(4), pp.513-527. https://doi.org/10.2307/2986142.

**Blank and Rasmussen 2004** Blank, G. and Rasmussen, K.B. (2004) “The data documentation initiative: the value and significance of a worldwide standard.” *Social Science Computer Review*, 22(3), pp.307-318. https://doi.org/10.1177/08944393042631.

**Blei, Ng and Jordan 2003** Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003) “Latent dirichlet allocation”, *Journal of machine Learning research*, 3(Jan), pp.993-1022.

**Błoch, Vasques Filho and Bojanowski 2021** Błoch, A., Vasques Filho, D. and Bojanowski, M. (2021) “Slaves, Freedmen, Mulattos, Pardos, and Indigenous Peoples: The Early Modern Social Networks of the Population of Color in the Atlantic Portuguese Empire.” In *The Digital Black Atlantic* (pp. 150-161). University of Minnesota Press. https://doi.org/10.5749/j.ctv1kchp41.16.

**Błoch, Vasques Filho and Bojanowski 2022** Błoch, A., Vasques Filho, D. and Bojanowski, M. (2022) “Networks from archives: Reconstructing networks of official correspondence in the early modern Portuguese empire”, *Social Networks*, 69, pp.123-135. https://doi.org/10.1016/j.socnet.2020.08.008.

**Błoch 2025** Błoch, A. et al. (2025) “Cracking the historical code: From unstructured correspondence corpora to computational analysis”. In *Models of data extraction and architecture in relational databases of early modern private political archives*. Edizioni Ca Foscari. http://doi.org/10.30687/978-88-6969-919-1/006.

**Broadwell et al. 2020** Broadwell, G.A. et al. (2020) “Ticha: Collaboration with Indigenous communities to build digital resources on Zapotec language and history”, *Digital Humanities Quarterly*, 14(4). Available at: https://dhq.digitalhumanities.org/vol/14/4/000529/000529.html.

**Brook 1981** Brook, T. (1981) “The merchant network in 16th Century China: A discussion and translation of Zhang Han's ‘On Merchants’”, *Journal of the Economic and Social History of the Orient/Journal de l'histoire economique et sociale de l'Orient*, pp.165-214. https://doi.org/10.1163/156852081X00086.

**Budroni, Claude-Burgelman, and Schouppe 2019** Budroni, P., Claude-Burgelman, J. and Schouppe, M. (2019) “Architectures of knowledge: the European open science cloud”. *ABI Technik*, 39(2), pp.130-141. https://doi.org/10.1515/abitech-2019-2006.

**Burrows 2023** Burrows, T., 2023. “Reproducibility, verifiability, and computational historical research”, *International Journal of Digital Humanities*, 5(2), pp.283-298. https://doi.org/10.1007/s42803-023-00068-9.

**Cao, Stojkovic and Obradovic 2016** Cao, X.H., Stojkovic, I. and Obradovic, Z. (2016) “A robust data scaling algorithm to improve classification accuracies in biomedical data”, *BMC Bioinformatics*, 17, pp.1-10. https://doi.org/10.1186/s12859-016-1236-x.

**Cao, Williams and Williams 1999** Cao, Y., Williams, D.D. and Williams, N.E. (1999) “Data transformation and standardization in the multivariate analysis of river water quality”, *Ecological Applications*, 9(2), pp.669-677. https://doi.org/10.2307/2641153.

**Chirigati et al. 2016** Chirigati, F. et al. (2016 June) “Reprozip: Computational reproducibility with ease”. In *SIGMOD '16: Proceedings of the 2016 international conference on management of data*, pp.2085-2088. https://doi.org/10.1145/2882903.289940

**Curran_2018** Curran, B. et al. (2018) “Look who’s talking: Two-mode networks as representations of a topic model of New Zealand parliamentary speeches”, *PloS ONE*, 13(6), p.e0199072. https://doi.org/10.1371/journal.pone.0199072.

**DDI Initiative 2024a** DDI Initiative. (2024a) Available at: https://ddialliance.org/ddi-codebook.

**DDI Initiative 2024b** DDI Initiative. (2024b) Available at: https://ddialliance.org/overview-of-current-products.

**DDI Initiative 2024c** DDI Initiative. (2024c) Available at: https://ddialliance.org/ddi-lifecycle.

**Elo 2020** Elo, K. (2020) “Utilizing historical network analysis on meta-data to model East German foreign intelligence cycle in the Baltic Sea Region 1975–89”. In *The Power of Networks*, pp.153-171. Routledge.

**Elwert 2020** Elwert, F. (2020) “Social and semantic network analysis in the study of religions”. In *The Power of Networks*, pp.172-186. Routledge.

**Engel et al. 2013** Engel, J. et al. (2013) “Breaking with trends in pre-processing?”, *TrAC Trends in Analytical Chemistry*, 50, pp.96-106. https://doi.org/10.1016/j.trac.2013.04.015.

**Fan et al. 2021** Fan, C. et al. (2021) “A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data”, *Frontiers in energy research*, 9, p.652801. https://doi.org/10.3389/fenrg.2021.652801.

**Fridlund and Brauer 2013** Fridlund, M. and Brauer, R. (2013) “Historizing topic models: A distant reading of topic modeling texts within historical studies”. In *Cultural Research in the Context of “Digital Humanities”: Proceedings of International Conference 3-5 October 2013, St Petersburg* (pp.152-63). Herzen State Pedagogical University.

**García et al. 2016** García, S. et al. (2016) “Big data preprocessing: methods and prospects”, *Big data analytics*, 1, pp.1-22. https://doi.org/10.1186/s41044-016-0014-0.

**Geraerts and Vasques Filho 2024** Geraerts, J. and Vasques Filho, D. (2024) “Networks of confessional affiliation: Religious choice and the schism of Utrecht”, *Journal of Historical Network Research*, 10(1) pp.54-91. https://doi.org/10.25517/jhnr.v10i1.77.

**Gibert, Sànchez–Marrè and Izquierdo 2016** Gibert, K., Sànchez–Marrè, M. and Izquierdo, J. (2016) “A survey on pre-processing techniques: Relevant issues in the context of environmental data mining”. *AI Communications*, 29(6), pp.627-663. https://doi.org/10.3233/AIC-160710.

**Gorgolewski and Poldrack 2016** Gorgolewski, K.J. and Poldrack, R.A. (2016) “A practical guide for improving transparency and reproducibility in neuroimaging research”. *PLoS Biology*, 14(7), p.e1002506. https://doi.org/10.1371/journal.pbio.1002506.

**Haggerty and Haggerty 2011** Haggerty, J. and Haggerty, S. (2011) “The life cycle of a metropolitan business network: Liverpool 1750–1810”, *Explorations in Economic History*, 48(2), pp.189-206. https://doi.org/10.1016/j.eeh.2010.09.006.

**Haibe-Kains et al. 2020** Haibe-Kains, B. et al. (2020) “Transparency and reproducibility in artificial intelligence”. *Nature*, 586(7829), pp.E14-E16. https://doi.org/10.1038/s41586-020-2766-y

**Jacobsen et al. 2020** Jacobsen, A. et al. (2020) “FAIR principles: interpretations and implementation considerations”, *Data intelligence*, 2(1-2), pp.10-29. https://doi.org/10.1162/dint_r_00024

**Jentsch and Porada 2012** Jentsch, P. and Porada, S. (2021) “From text to data: Digitization, text analysis and corpus linguistics”, *Digital Methods in the Humanities. Challenges, Ideas, Perspectives*, pp.89-128. https://doi.org/10.2307/j.ctv2f9xskk.6

**Jiménez–Badillo et al. 2020** Jiménez–Badillo, D. et al. (2020) “Developing geographically oriented NLP approaches to sixteenth–century historical documents: Digging into early colonial Mexico”, *Digital Humanities Quarterly*, 14(4). Available at: https://dhq.digitalhumanities.org/vol/14/4/000490/000490.html.

**Joo, Choi, and Park 2000** Joo, D.S., Choi, D.J. and Park, H. (2000) “The effects of data preprocessing in the determination of coagulant dosing rate”, *Water Research*, 34(13), pp.3295-3302. https://doi.org/10.1016/S0043-

1354(00)00067-1.

**Kamiran and Calders 2012** Kamiran, F. and Calders, T. (2012) “Data preprocessing techniques for classification without discrimination”, *Knowledge and information systems*, 33(1), pp.1-33. https://doi.org/10.1007/s10115-011-0463-8.

**Kesner 1982** Kesner, R.M. (1982) “Historians in the information age: Putting the new technology to work”, *The Public Historian*, 4(3), pp.31-48. https://doi.org/10.2307/3377464.

**Klingner 2023** Klingner, C.M. et al. (2023)“ Research data management and data sharing for reproducible research — results of a community survey of the german national research data infrastructure initiative neuroscience”, *eNeuro*, 10(2). https://doi.org/10.1523/ENEURO.0215-22.2023.

**Korenius et al. 2004** Korenius, T. et al. (2004, November) “Stemming and lemmatization in the clustering of finnish text documents”. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, pp.625-633. https://doi.org/10.1145/1031171.103128.

**Kunilovskaya and Plum 2021** Kunilovskaya, M. and Plum, A. (2021, September) “Text preprocessing and its implications in a digital humanities project”. In Proceedings of the *Student Research Workshop Associated with RANLP 2021*, pp. 85-93. Available at: https://aclanthology.org/2021.ranlp-srw.13/.

**Lamprecht et al. 2020** Lamprecht, A.L. et al. (2020) “Towards FAIR principles for research software”, *Data Science*, 3(1), pp.37-59. https://doi.org/10.3233/DS-1900.

**Lee, Liong and Jemain 2017** Lee, L.C., Liong, C.Y. and Jemain, A.A. (2017) “A contemporary review on Data Preprocessing (DP) practice strategy in ATR-FTIR spectrum”, *Chemometrics and Intelligent Laboratory Systems*, 163, pp.64-75. https://doi.org/10.1016/j.chemolab.2017.02.008.

**Lelewer and Hirschberg 1987** Lelewer, D.A. and Hirschberg, D.S. (1987) “Data compression”, *ACM Computing Surveys (CSUR)*, 19(3), pp.261-296.

**Linkevicius de Andrade and Vasques Filho 2022** Linkevicius de Andrade, D. and Vasques Filho, D. (2022) “Moderation and authority-building process: the dynamics of knowledge creation on history subreddits”, *Internet Histories*, 6(4), pp.369-390.

**Matthew and Bannister 2020** Matthew, L. and Bannister, M. (2020) “The form of the content: The digital archive Nahuatl/Nawat in Central America”, *Digital Humanities Quarterly*, 14(4). Available at: https://dhq.digitalhumanities.org/vol/14/4/000491/000491.html.

**Meaney et al. 2023** Meaney, C. et al (2023) “Quality indices for topic model selection and evaluation: a literature review and case study”, *BMC Medical Informatics and Decision Making*, 23(1), p.132. https://doi.org/10.1186/s12911-023-02216-1.

**Mishra et al. 2019** Mishra, P. et al. (2019) “Automatic de-noising of close-range hyperspectral images with a wavelength-specific shearlet-based image noise reduction method”, *Sensors and Actuators B: Chemical*, 281, pp.1034-1044. https://doi.org/10.1016/j.snb.2018.11.034.

**Mishra et al. 2020** Mishra, P. et al. (2020) “New data preprocessing trends based on ensemble of multiple preprocessing techniques”, *TrAC Trends in Analytical Chemistry*, 132, p. 116045. https://doi.org/10.1016/j.trac.2020.116045.

**Monroe et al. 2013** Monroe, M. et al. (2013) “Temporal event sequence simplification”, *IEEE Transactions on Visualization and Computer Graphics*, 19(12), pp. 2227-2236. https://doi.org/10.1109/tvcg.2013.200.

**Nadeau and Sekine 2007** Nadeau, D. and Sekine, S. (2007) “A survey of named entity recognition and classification”, *Lingvisticae Investigationes*, 30(1), pp.3-26. http://dx.doi.org/10.1075/li.30.1.03nad.

**Nikolenko, Koltcov and Koltsova 2017** Nikolenko, S.I., Koltcov, S. and Koltsova, O., 2017. “Topic modelling for qualitative studies”, *Journal of Information Science*, 43(1), pp.88-102. http://dx.doi.org/10.1177/0165551515617393.

**Padgett and Ansell 1993** Padgett, J.F. and Ansell, C.K. (1993) “Robust action and the rise of the Medici, 1400-1434”, *American Journal of Sociology*, 98(6), pp.1259-1319. Available at: https://www.jstor.org/stable/2781822.

**Peng 2011** Peng, R.D. (2011) “Reproducible research in computational science”, *Science*, 334(6060), pp.1226-1227. https://doi.org/10.1126/science.1213847.

**Peng and Hicks 2021** Peng, R.D. and Hicks, S.C. (2021) “Reproducible research: A retrospective”, *Annual Review of Public Health*, 42, pp.79-93. https://doi.org/10.1146/annurev-publhealth-012420-105110.

**Petterson et al. 2016** Pettersson, E. et al. (2016, July) “HistSearch-Implementation and evaluation of a web-based tool for

automatic information extraction from historical text". In *HistoInformatics@ DH*, pp.25-36.

**Petz and Pfeffer 2021** Petz, C. and Pfeffer, J. (2021) "Configuration to conviction: Network structures of political judiciary in the Austrian Corporate State", *Social Networks*, 66, pp.185-201. https://doi.org/10.1016/j.socnet.2021.03.001.

**Poulston, Stevenson and Bontcheva 2015** Poulston, A., Stevenson, M. and Bontcheva, K. (2015) "Topic models and n–gram language models for author profiling". In *Proceedings of CLEF*, 1391, pp. 1-7.

**Rahm and Do 2000** Rahm, E. and Do, H.H. (2000) "Data cleaning: Problems and current approaches", *IEEE Data Engineering Bulletin*, 23(4), pp.3-13.

**Rasmussen and Blank 2007** Rasmussen, K.B. and Blank, G. (2007) "The data documentation initiative: A preservation standard for research", *Archival Science*, 7, pp.55-71. https://doi.org/10.1007/s10502-006-9036-0.

**Ravenek, van den Heuvel and Gerritsen 2017** Ravenek, W., van den Heuvel, C. and Gerritsen, G. (2017) "The ePistolarium: Origins and techniques", *CLARIN in the Low Countries*, London: Ubiquity Press, pp.317-323. https://doi.org/10.5334/bbi.26.

**Rinnan, Van Den Berg and Engelsen 2009** Rinnan, Å., Van Den Berg, F. and Engelsen, S.B. (2009)" Review of the most common pre-processing techniques for near-infrared spectra", *TrAC Trends in Analytical Chemistry*, 28(10), pp.1201-1222. https://doi.org/10.1016/j.trac.2009.07.007

**Rosario 2000** Rosario, B. (2000) "Latent semantic indexing: An overview", *Techn. rep. INFOSYS*, 240, pp.1-16.

**Salmi 2021** Salmi, H. (2021) *What is digital history?*, Medford, Massachusetts: Polity Press.

**Shanker, Hu and Hung 1996** Shanker, M., Hu, M.Y. and Hung, M.S. (1996) "Effect of data standardization on neural network training", *Omega*, 24(4), pp.385-397. https://doi.org/10.1016/0305-0483(96)00010-2

**Sidorov et al. 2014** Sidorov, G. et al. (2014) "Syntactic n-grams as machine learning features for natural language processing", *Expert Systems with Applications*, 41(3), pp.853-860. https://doi.org/10.1016/j.eswa.2013.08.015.

**Skorkovská 2012** Skorkovská, L. (2012) "Application of lemmatization and summarization methods in topic identification module for large scale language modeling data filtering". In *Text, Speech and Dialogue: 15th International Conference, TSD 2012, Brno, Czech Republic, September 3-7, 2012. Proceedings 15*, Berlin Heidelberg: Springer, pp.191-198.

**Srinivasa-Desikan 2018** Srinivasa-Desikan, B. (2018) *Natural language processing and computational linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*. Birmingham, UK: Packt Publishing Ltd.

**Stodden et al. 2014** Stodden, V., Leisch F. and Peng, R.D. (2014) *Implementing Reproducible Research.* CRC Press.

**Tukey 1977** Tukey, J.W. (1977) *Exploratory data analysis*. Reading, Massachusetts: Addison-Wesley.

**Verbruggen, Blomme and D'haenick 2020** Verbruggen, C., Blomme, H. and D'haeninck, T. (2020) "Mobility and movements in intellectual history: a social network approach". *The Power of Networks: Prospects of Historical Network Research*. New York, New York: Routledge, pp.125-150.

**Wetherell, Plakans and Wellman 1994** Wetherell, C., Plakans, A. and Wellman, B. (1994) "Social networks, kinship, and community in Eastern Europe", *The Journal of Interdisciplinary History*, 24(4), pp.639-663. https://doi.org/10.2307/205629.

**Wittek and Ravenek 2011** Wittek, P. and Ravenek, W. (2011) "Supporting the exploration of a corpus of 17th-Century scholarly correspondences by topic modeling". In *SDH 2011 Supporting Digital Humanities: Answering the unaskable*. University of Copenhagen.

**Xu et al. 2020** Xu, Y. et al. (2020) "Effects of data preprocessing methods on addressing location uncertainty in mobile signaling data", *Annals of the American Association of Geographers*, 111(2), pp.515-539. https://doi.org/10.1080/24694452.2020.1773232.

**Ye 2022** Ye, M.J. (2022) "A history from below: Translators in the publication network of four magazines issued by the China Book Company, 1913–1923", *Translation Studies*, 15(1), pp.37-53. https://doi.org/10.1080/14781700.2021.1950043

**Yin and Wang 2014** Yin, J. and Wang, J. (2014, August) "A dirichlet multinomial mixture model-based approach for short text clustering". In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 233-242. https://doi.org/10.1145/2623330.2623715.

**Zelaya 2019** Zelaya, C.V.G. (2019, April) "Towards explaining the effects of data preprocessing on machine learning". In

*2019 IEEE 35th international conference on data engineering (ICDE)*, pp.2086-2090. \.

**Zhu and Gao 2016** Zhu, C. and Gao, D. (2016) “Influence of data preprocessing”, *Journal of Computing Science and Engineering*, 10(2), pp.51-57. http://dx.doi.org/10.5626/JCSE.2016.10.2.51

**Zbíral amd Shaw 2022** Zbíral, D. and Shaw, R.L. (2022) “Hearing voices: reapproaching medieval inquisition records”, *Religions*, 13(12), p.1175. https://doi.org/10.3390/rel13121175.

**spaCy 2024** spaCy (2024) “pt_core_news_lg”, *Hugging Face*. Available at: https://huggingface.co/spacy/pt_core_news_lg.