

Text Mining to uncover Prehistoric Pastness in Museums

Haley Anne Schwartz^{1,2,*†}, Paula Jardón Giner^{2,3,†} and Xavier Rubio Campillo^{1,2,†}

¹*Departament de Didàctiques Aplicades, Universitat de Barcelona, Pg. de la Vall d'Hebron, 171, 08035 Barcelona, Spain*

²*DIDPATRI Grup de Recerca, Pg. de la Vall d'Hebron, 171, 08035 Barcelona, Spain*

³*Departament de Didàctica de les Ciències Experimentals i Socials, Universitat de València, Av. de Blasco Ibáñez, 13, El Pla del Real, 46010 València, Valencia, Spain*

Abstract

This paper is a presentation of a current work in progress, specifically the exploratory phase for determining a methodological framework, clear objectives, and establishing preliminary results to guide the future direction of the project. The paper sees the application of text analysis to a corpus body of texts with a focus on highlighting heritage and intersectional data present within these texts. The approach of text analysis allows for a quantitative analysis of modern perceptions of the past, narratives given to the past by modern people, and the resulting context elements of the past are placed in stemming from modern influences. With a focus on how prehistory is presented to modern people, in the specific context of museums, it is necessary to trace the contents of texts depicting the past in these museums. The overall goal of this paper is to have a deeper understanding of the impact modern narratives attributed to the past has on the prehistoric past in an educational context. Specifically, looking at narratives focused on the process of neolithization as discussed in museums. Additionally, preliminary explorations give insight into the benefits of the methodology and how to best establish next steps to propel future research.

Keywords

Text Mining, Topic Modelling, Museums, Prehistory, Digital Humanities

1. Introduction

Museums are integral tools through which the past is understood and interpreted using tangible scientific evidence. These institutions provide modern people with both physical and experiential elements for developing interpretations and relationships with the past [33] and educational opportunities [19]. As far as the physical, museums make accessible surviving material culture through exhibits and displays, granting visitors direct visual and physical access to remnants of the past [7]. Museums use storytelling and display techniques that cultivate an experience for visitors to further build connections [23]. The contextualization of the past in the form of digestible interpretations for visitors are linked to the place and time of origin for the material culture used to aid in the storytelling and educational process of the archaeological information available [21]. These include, but are not limited to photography, audiovisual


CHR 2024: Computational Humanities Research Conference, December 4–6, 2024, Aarhus, Denmark

*Corresponding author.

†These authors contributed equally.

✉ hschwartz@ub.edu (H. A. Schwartz); paula.jardon@uv.es (P. J. Giner); xrubio@ub.edu (X. R. Campillo)

ORCID [0009-0001-8231-3160](https://orcid.org/0009-0001-8231-3160) (H. A. Schwartz); [0000-0003-1542-7683](https://orcid.org/0000-0003-1542-7683) (P. J. Giner); [0000-0003-4428-4335](https://orcid.org/0000-0003-4428-4335) (X. R. Campillo)

 © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

supplementation, digital reconstructions, and texts. These modalities consumed by visitors are available and ripe for analysis.

The following paper details the exploratory stage for an in-progress research project utilizing quantitative textual analysis applied to a corpus of archaeological museum texts. This stage of the analysis was used to formulate the best methodological approach to carry out text mining for this corpus, set parameters, and solidify next steps in this project. This research is focused on the types of narratives linked with production societies within museums, specifically museums along the east of the Iberian Peninsula. The interest is in how prehistory-producing societies are taught in museums by analyzing associated texts.

As this is the first stage of the project all results are beneficial to evaluate the data, any need for additional data, and exploring the research questions through quantitative text analysis. The proposed questions are: 1) Through the texts produced in these museums is it possible to determine the key issues from those societies and periods? 2) Do the narratives fit with the current accepted research surrounding these societies and periods? 3) Are any problems from past societies related to problems of the present? and 4) How is the information treated and presented within these museums? These early results are indicative for next steps in this research project.

2. Background

2.1. Archaeological Heritage Discourse and Text Analysis

In the context of archaeological heritage, textual sources are an important source highlighting shifting narratives, contexts, and interpretations linked to tangible and intangible archaeological heritage – including landscapes, surviving sites, museums, etc. Discourse studies and continually evolving quantitative methods continue to highlight the benefits of textual analysis applied to archaeological heritage data [29, 4, 27, 31]. Museums contain a surplus of textual sources discussing tangible and intangible archaeological heritage.

In previous research, this method has been used to discern how archaeologists discuss social issues over time [27], temporal shifts within academic articles about archaeological heritage landscapes [schwartz2023text] and tracing geographical dispersion of archaeological research in regions using historical archive data and newspaper collections [22]. Applying textual analysis tools on museum texts aids in uncovering subjectivity and biases within museums in how archaeological knowledge is presented to visitors.

2.2. Museums as facilitators of Memory and Relationships with the Past

Data from surveys and other research provides insight into how modern people cultivate relationships with the past through perceptions, interpretations, and context in which people are exposed to the past [2]. These interpretations are correlated with individual experiences and cultural constructs illustrating that the past is seen by modern people not through the lens of what the past was but through the lens of the present [2]. Our ability to connect with the past is relegated to individual and group perceptions, stemming from the context in which the past is interpreted and presented . There is a cycle in the interpretation, presentation, perception,

and reinterpretation of the past that is dependent upon the era, new discoveries, new research, and shifting cultural structures.

Looking at museums as agents facilitating that cycle, it is important to look to the existing presentation of the past within these museums. In order to determine what constitutes the narratives currently presented to people about the past, to see the effects of archaeological education with this modality [11]. It is necessary to make clear that archaeological knowledge – in whichever context – is a produced knowledge linked to era, bias, interpretation, and existing research and discoveries [9]. Any and all educational pursuits are carried out subjectively, which is further reflected in museums – that span countries, cultures, and contexts. Furthermore, the dissemination of archaeological knowledge – in various stages of production or reproduction – is heavily influenced through social methods rather than more passive techniques [9]. The more the presentation of the past is considered in the museum context, the more insight we gain into the specific educational methods and narratives consumed by visitors.

2.3. The Context: Archaeological Museums and Neolithization in the Iberian Peninsula

Texts appearing within museums linked to the development of early farming societies and metal age societies– including discussions on the neolithization process – were utilized in this study. This project aims to gain a better understanding of the words and concepts linked to this part of the past. Specifically, the way in which narratives of neolithization and early producing societies are presented to the public in museums. Within the Iberian Peninsula, there are regional-specific variations and circumstances in which the overall development of these production societies occurred [1]. There are differences in which neolithization impacted specific regions and the people within, in conjunction with clear neolithic traits not dependent on a region (i.e., farming practices, economic structures, technological advancements, etc.) [12]. Additionally, changes to social structures and interactions between others which impacted the role of reciprocity, growth of social inequalities, and the development of social networks [1], will be explored in future research.

Within the Iberian Peninsula, new research and discoveries focused on the region [14, 17, 18] brings attention to the museums displaying these processes and. As the production and reproduction of this type of archaeological knowledge evolves, how quickly and accurately does this knowledge evolve within the museums. Analyzing museum texts allows one to see what contexts and narratives surrounding production societies and neolithization in the Iberian Peninsula is presented.

3. Materials and Methods

3.1. Materials

The corpus used in this project was built from a collection of texts in Català, compiled from permanent exhibitions appearing in eight museums over a period of thirty-years along the east of the Iberian Peninsula. These texts – with an overall sum of twenty-five thousand words – were extrapolated from panels that describe the different themes of the museum. They are

explanatory texts of general Neolithic culture characterizations (i.e., farming, work techniques, religion, etc.). They exclude display cases as the sole interest is the main narratives presented to visitors. The texts are general descriptors of Neolithic elements, with a select few descriptors of specific sites. R Statistical Software (v4.2.2)[30] was used to carry out all processing, analysis, and visualizations of the texts.

3.2. Methods

The texts were all collected using the same process of OCR with manual and automatic inference. The individual texts were photographed and converted using OCR, with the texts manually verified when necessary. During this process, it was decided to record all texts as a single document per museum. The reasoning is each museum splits the content differently. Therefore, a side-by-side comparison of each text (i.e., panel) would not be useful and would include strong biases (i.e., museums where text is split in several panels would have a higher weight than other ones where text is concentrated on a few large panels). Following the text collection, Topic modelling was applied as a type of ‘distant reading’ to trace linguistic patterns regarding the storytelling process between different museums presentation of archaeological knowledge. R was used to explore content, word appearance, word frequency, and context. All texts were loaded into, processed, and the corpus built within R using a number of packages: tidytext (v0.3.4)[32], tm (v0.7.9)[6] and topicmodels (v0.2.13)[8].

Topic modelling was selected for its machine learning prowess in the sorting and classification of documents in a corpus, assigning topics, and highlighting temporal distribution of topics [5]. The process of topic modelling itself is a Bayesian analytic approach which is capable of identifying semantic structures within documents that make up a corpus and then reorder the entities within a corpus (words, documents, topics) based on probability distribution between the entities of a corpus (topics linked with words, documents linked with topics) [3, 26]. There are a range of methods for carrying out topic modelling with Latent Dirichlet allocation (LDA) chosen due to how the algorithm searches for topics, assigns topics, and having unstructured topics which fits best with both the small size of the corpus and the datatype [26]. Previous research highlighted the benefits of applying LDA to archaeological heritage texts [27, 31].

Previous research was consulted to determine a preprocessing chain to help remove bias and noise as much as possible. A metrics test was run to determine the appropriate number of topics per the corpus, parameters of utilizing stopwords and applying lemmatization, and training an LDA model was decided upon using previous research as guidelines [26, 27, 31]. Once the parameters were set, metrics tests were run, and the model trained – the processed corpus was analyzed using the previously mentioned R packages in conjunction with the additional package ldatuning (v1.0.2) [24] to run the LDA algorithm.

4. Preliminary Explorations

This first exploration of the data used text mining techniques using R to analyze word frequency and occurrence. The focus was to find the overall most frequently appearing terms throughout the corpus and then the most frequently occurring terms for each text per museum. Looking first at the frequency of term appearance overall, as seen in Figure 1, the entirety of the terms

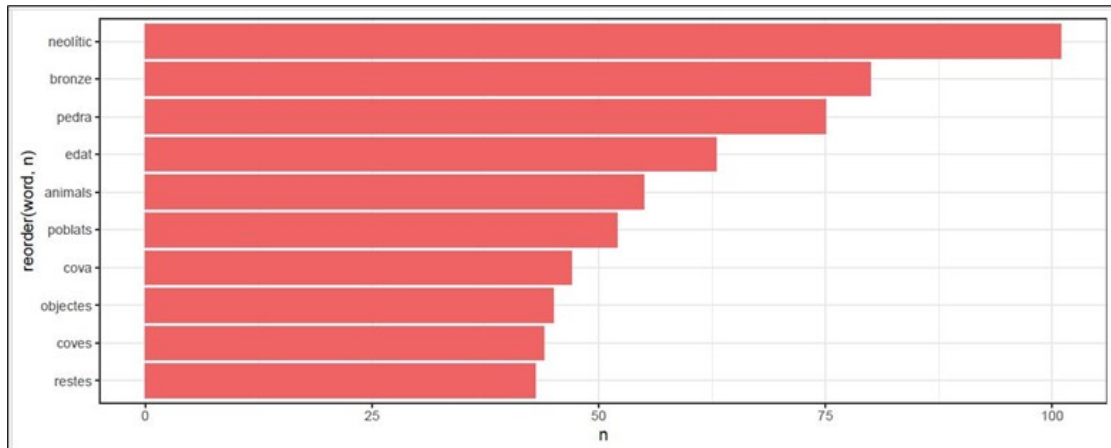


Figure 1: Overall Term-Frequency

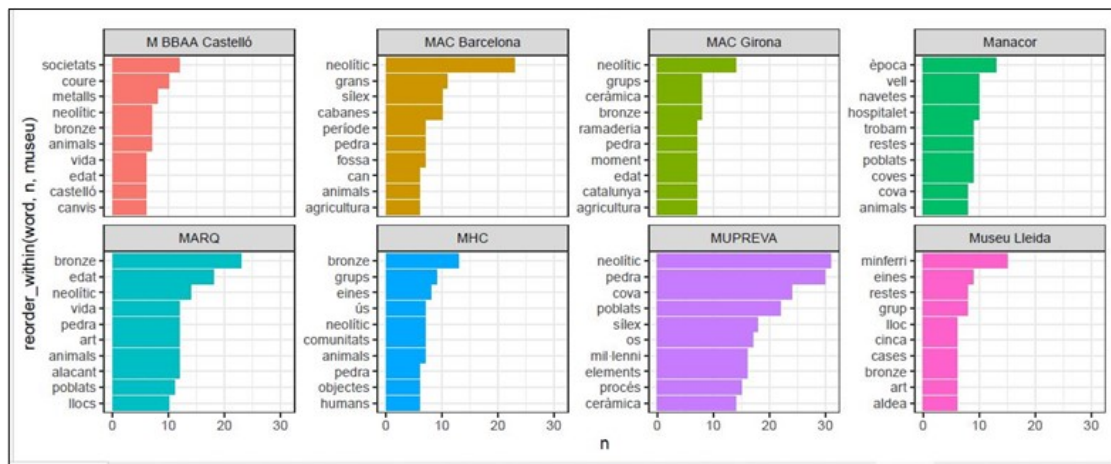


Figure 2: Term-Frequency per Museum

present are of importance. The terms all are indicative and relative to the overall process of neolithization. As the terms are in Català, translations include: neolitic(*neolithic*), pedra(*stone*), poblats(*villages*), cova/coves(*cave(s)*), restes(*remains*), edat(*age*), silex(*flint*), objectes(*objects*).

Now looking at the frequency of word appearance per text per museum, differences in the narratives per museum are discernible within Figure 2. For starters, in the entries for M BBAA Castelló (*Museo de Bellas Artes de Castellón*), key terms include: coure(*copper*), vida(*life*), canvis(*changes*), and all centered around the province Castelló – also a term – located in Valencia. Whereas, in the entries for MAC Barcelona (*Museu d'Arqueologia de Catalunya, Barcelona*), key terms are: sílex(*flint*), cabanes(*huts*), and fossa(*moats*).The last example are the entries for Museu Lleida(*Museu de Lleida*) with the terms: restes(*remains*), eines(*tools*), lloc(*site*) all pertaining to Minferri – an additional term – which is an Early Bronze Age settlement located in Lleida. Just from this intervention, there is a delineation between narratives in museums – linked to location and elements of value in each location.

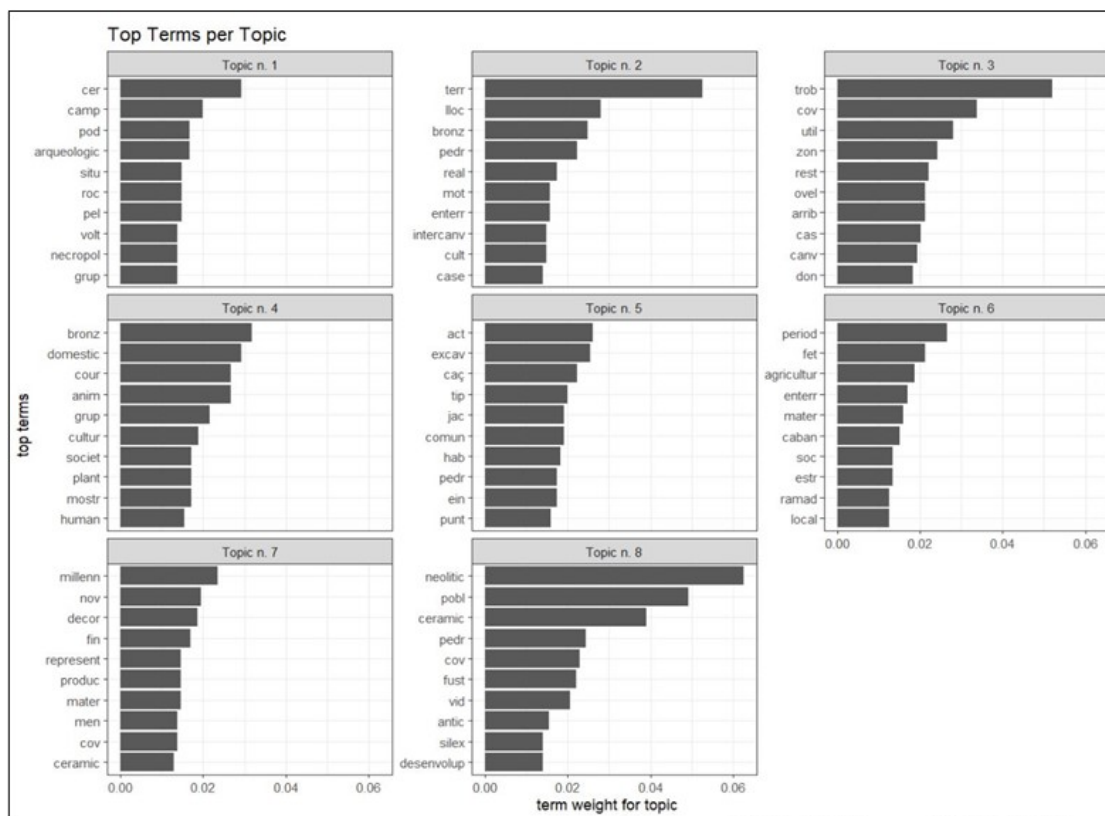


Figure 3: Top Ten Terms per Topic

The next step was running topic modelling. In this iteration, the model is made up of nine topics composed of the ten most frequently occurring terms per topic. These preliminary explorations made visible the types of topics within the corpus and the trends in each topic, as illustrated in Figure 3.

Inference based on the terms give insight to the general theme of the topic itself. For example, in Topic 4 terms include: *lloc(place)*, *bronz(bronze)*, *vid(life)*, *cult(worship)*, and *art* – which indicates the context of art and culture within the bronze age. Whereas in topic 8 key terms are: *cov(cave)*, *cultur(culture)*, *décor(decoration)*, and *bronz(bronze)* – leading to infer this topic pertains to cave art during this period. Additionally, it is through interpreting these topics that names are developed as a result. Proposed names per topic are presented in Table 1.

Focusing on the themes within each of the topics highlight contexts within the text, of what terms were occurring in proximity together. These early results provided the researchers of the project the chance to consider potential changes to the dataset, adjustments to the methodological approach, and other ways of analyzing the contents of the texts.

Table 1: Interpreted Topic Names

Topic Number	Topic Name	Terms
Topic 1	ritus funeraris	necropol, funer, case
Topic 2	agricultura i ramaderia	domestic, societ, agric, cour
Topic 3	materials	pedr, ceramic, silex, neolithic
Topic 4	art del bronze	bronz, lloc, art, cult, conserve
Topic 5	tipus d'estructures arqueològiques	fet, enterr, caracter, period, estr
Topic 6	paleolític	paleolithic, punt, represent, comun, product
Topic 7	edat del ferro	epoc, ferr, cultur, anim, pobl, trob
Topic 8	art rupestre	object, cultur, pres, bronz, cov, decor
Topic 9	treball de camp en arqueologia	excav, trev, tip, jac, mater, fin

5. Discussion

Museums are one example of a center for archaeological heritage discourse in which individuals can experience material and immaterial culture which helps develop one's relationship with the past [25, 16]. Focusing on archaeological museums specifically, they are a respected place in which visitors experience, perceive, and learn about archaeological knowledge in a public setting. Before continuing further, it is important to note the many criticisms and controversies that are necessary in museum discourse today regarding problems and unethical museum-practices of the past – and in some cases – the present [10]. Regarding the scope of this paper as the exploratory phase, these points will not be discussed further presently but are integral to consider and include in the future.

Museums are integral with cultivating memory of the past seen through the interpretations used to create narratives presented to visitors [15]. Yet, narratives surrounding the past are not stagnant and as new information becomes available, new technologies allow for more complex study [20]. It is necessary to trace shifts in changing narratives and interpretations, looking at how often museums update their presentations and exhibits [15]. Archaeological heritage discourse sheds light on the narratives, contexts, and perceptions of the dominating views of the time – and that extends to discourse in museums. Texts existing within museums should be utilized to the same degree as other archaeological heritage textual sources to gain insight into the untapped data within. These first results are valuable for reviewing the current methodology process and determining how the project can best progress.

As far as assessing how text analysis can illuminate key issues from the past as it is connected to early production societies, the present themes and word frequency illuminate the focal point of the narratives and interpretations presented to visitors. If museums are determined to provide visitors with insight into all facets of the past, terms and themes related to said issues should be visible through text mining. At this stage of the project, any answers to research questions are preliminary and are more valuable in determining next steps. Again, the primary focus has been on the efficacy of the methods, potential changes to the data, and what adjustments need to be made for parameters.

5.1. Challenges

Originally, during the preprocessing of the texts before building the corpus, there was a limit to the cleaning of the text. In part, this is due to the text in Català and the need to see how the language is recognized within R. There was no general list of stopwords for Català, though there is one in English. Nor a personalized list of stopwords, something that would aid in removing noise [28]. Early explorations indicated the necessity and additions to the preprocessing chain that include: general stopwords (i.e., *s'hi(in)*, *d'un(of one)*, and *s'han(they have)*), lemmatization, and removing special characters. While these changes were employed and explored, more intervention into the preprocessing chain in the context of working with a non-English language is needed.

5.2. Exploratory Results

The preliminary figures illustrate how with the strengthening of the methodological approach, the research questions can be answered through a quantitative analysis. With the goal of exploring the extent current exhibitions offering this archaeological knowledge consider all advances of knowledge over the last three decades.

With Figure 1, focusing on early production societies and neolithization, the terms we can expect about these peoples and societies in general are present (i.e., *pedra(stone)*, *poblats(villages)*, *ceràmica(ceramics)*, *silex(flint)*). For Figure 2, it is further possible to discern the relationship between topics and museums. For example, looking at MBAA Castelló, in this area neolithization is in relation to animals, with habitat more often in caves which is reflective of the mountain region. Where, MAC Barcelona sees terms that fit more describing the neolithic cultures in general within Catalunya, with more specialized concepts used by archaeologists. This is reflective of the museum having an academic and technological discourse.

Regarding these texts reflecting narratives in accepted research, there are discrepancies. For example, within these texts there is no mention of Mesolithic peoples, nor any mention of the transformations of neolithic landscapes. There is also a focus on specific neolithic activities that take precedence over others. There is a larger appearance of farming compared to other neolithic activities, such as herding. Additionally, there is no mention of the impact of neolithic on prehistoric landscapes. This reflects that some issues are the importance of purely economic activities and technology beyond other types of traits that define neolithic (i.e., social structures, sexual division of labor, environmental dynamics, etc.). Furthermore, with the goal of being able to interpret and understand the treatment and presentation of these texts, additional analysis can consult linked metadata or comparing the evolution or stagnation of the information.

5.3. Next Steps

Regarding the selected texts, at the present moment, these are the only texts and only museums utilized – but there is a potential to bring in texts from an additional four museums in the specified region. As far as narratives present within museums remaining up to date with current research, the use of a qualitative approach – ‘close reading’ – of these texts and recent articles could provide more details. Another option could be adding an additional corpus made up of

recent research to analyze separately and compare results. Comparing corpus to corpus, to see potential similarities and differences between the narratives presented in museums and those produced in peer-reviewed journals. Based on the early explorations of the corpus, the final analysis filtered the corpus for any term not present in at least 2 museums and the number of topics was set at nine following metrics testing. The number of term appearances can also be adjusted, along with additional metrics testing to further evaluate the number of topics. The end result of this stage is a future direction for how this project can further develop, solidify the code and methodological framework.

6. Conclusion

Our ability to interpret, perceive, and relate to the past is only possible resulting from the value and significance of surviving material and immaterial culture. The increasing accessibility of this material and immaterial culture present within museums is linked to what modern people find valuable and significant enough from the past, to hold this level of presence within modern culture – on display in the modern collective memory [13]. Much like value and significance assigned to remnants of the past can change, so does the knowledge and interpretation of the surviving past change. Text analysis is a valuable tool to trace shifts in the knowledge and interpretation of archaeological heritage – including texts surrounding archaeological heritage currently residing in museums. This project will continue applying the methodological approach to better understand and disseminate current patterns of archaeological knowledge, and their respective narratives and contexts as they exist within museums.

References

- [1] J. B. Aubán. “The social and symbolic context of Neolithization”. In: *SAGVNTVM Extra* 5 (2002), pp. 209–234.
- [2] A. W. Barker. “Exhibiting archaeology: archaeology and museums”. In: *Annual Review of Anthropology* 39.1 (2010), pp. 293–308.
- [3] D. M. Blei and J. D. Lafferty. “Topic models”. In: *Text mining*. Chapman and Hall/CRC, 2009, pp. 101–124.
- [4] A. Burkette and R. Skeates. “The Words that Archaeologists Choose: A Maltese Case Study in Artifact Terminology, Corpus Linguistics and Discourse Analysis”. In: *Journal of Mediterranean Archaeology* 35.1 (2022).
- [5] A. Daud, J. Li, L. Zhou, and F. Muhammad. “Knowledge discovery through directed probabilistic topic models: a survey”. In: *Frontiers of computer science in China* 4 (2010), pp. 280–301.
- [6] I. Feinerer and K. Hornik. *tm: Text Mining Package*. 2022. URL: <https://CRAN.R-project.org/package=tm>.
- [7] C. Goulding. “The museum environment and the visitor experience”. In: *European Journal of marketing* 34.3/4 (2000), pp. 261–278.

- [8] B. Grün and K. Hornik. *topicmodels: Topic Models*. 2022. URL: <https://CRAN.R-project.org/package=topicmodels>.
- [9] Y. Hamilakis. “Archaeology and the politics of pedagogy”. In: *World archaeology* 36.2 (2004), pp. 287–309.
- [10] N. Harris. “Museums and controversy: Some introductory reflections”. In: *The Journal of American History* 82.3 (1995), pp. 1102–1110.
- [11] D. Henson. “Archaeology and education”. In: *Key concepts in public archaeology* (2017), pp. 43–59.
- [12] V.-P. Herva, K. Nordqvist, A. Lahelma, and J. Ikäheimo. “Cultivation of Perception and the Emergence of the Neolithic World”. In: *Norwegian archaeological review* 47.2 (2014), pp. 141–160.
- [13] T. Ireland, S. Brown, and J. Schofield. “Situating (in) significance”. In: *International Journal of Heritage Studies* 26.9 (2020), pp. 826–844.
- [14] N. Isern, J. Fort, A. F. Carvalho, J. F. Gibaja, and J. J. Ibañez. “The Neolithic transition in the Iberian Peninsula: data analysis and modeling”. In: *Journal of Archaeological Method and Theory* 21 (2014), pp. 447–460.
- [15] S. Jones. “Dialogues between past, present and future: Reflections on engaging the recent past”. In: *Archaeology, the Public and the Recent Past* (2013), pp. 163–176.
- [16] O. Knauss. “Museum Best Practices for Managing Controversy”. In: *National Coalition Against Censorship* (2019).
- [17] Í. G.-M. de Lagrán. “Recent data and approaches on the Neolithization of the Iberian Peninsula”. In: *European Journal of Archaeology* 18.3 (2015), pp. 429–453.
- [18] Í. G.-M. de Lagrán, E. Fernández-Domínguez, and M. A. Rojo-Guerra. “Solutions or illusions? An analysis of the available palaeogenetic evidence from the origins of the Neolithic in the Iberian Peninsula”. In: *Quaternary International* 470 (2018), pp. 353–368.
- [19] J. Lea. “Teaching the past in museums”. In: *Museums and Archaeology*. Routledge, 2022, pp. 473–484.
- [20] M. D. McCoy and T. N. Ladefoged. “New developments in the use of spatial technology in archaeology”. In: *Journal of Archaeological Research* 17 (2009), pp. 263–295.
- [21] S. Moser. “Representing archaeological knowledge in museums: Exhibiting human origins and strategies for change”. In: *Public Archaeology* 3.1 (2003), pp. 3–20.
- [22] P. Murrieta-Flores and I. Gregory. “Further frontiers in GIS: Extending spatial analysis to textual sources in archaeology”. In: *Open Archaeology* 1.1 (2015).
- [23] J. K. Nielsen. “Museum communication and storytelling: articulating understandings within the museum structure”. In: *Museum management and curatorship* 32.5 (2017), pp. 440–455.
- [24] M. Nikita. *ldatuning: Tuning of the Latent Dirichlet Allocation Models Parameters*. 2020. URL: <https://CRAN.R-project.org/package=ldatuning>.

- [25] B. L.-A. M. O'Mara. "Museums and controversy: you can't have one without the other." PhD thesis. 2007.
- [26] A. Onan, S. Korukoglu, and H. Bulut. "LDA-based topic modelling in text sentiment classification: An empirical analysis." In: *Int. J. Comput. Linguistics Appl.* 7.1 (2016), pp. 101–119.
- [27] G. Park, L.-Y. Wang, and B. Marwick. "How do archaeologists write about racism? Computational text analysis of 41 years of Society for American Archaeology annual meeting abstracts". In: *Antiquity* 96.387 (2022), pp. 696–709.
- [28] W. Parwita. "A document recommendation system of stemming and stopword removal impact: A web-based application". In: *Journal of Physics: Conference Series*. Vol. 1469. 1. IOP Publishing. 2020, p. 012050.
- [29] G. Plets, P. Huijnen, and D. van Oeveren. "Excavating archaeological texts: Applying digital humanities to the study of archaeological thought and banal nationalism". In: *Journal of Field Archaeology* 46.5 (2021), pp. 289–302.
- [30] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL: <https://www.R-project.org/>.
- [31] H. A. Schwartz. "Text Mining Analysis of Perception in Archaeological Landscapes: The Case of Stonehenge". In: *Proceedings of the 7th ACM SIGSPATIAL International Workshop on Geospatial Humanities*. 2023, pp. 40–43.
- [32] J. Silge and D. Robinson. "tidytext: Text Mining and Analysis Using Tidy Data Principles in R". In: *Joss* 1.3 (2016). DOI: 10.21105/joss.00037. URL: <http://dx.doi.org/10.21105/joss.00037>.
- [33] A. Witcomb. "Thinking about others through museums and heritage". In: *The Palgrave handbook of contemporary heritage research*. Springer, 2015, pp. 130–143.