Unraveling the Threads of Thrace: A Text Mining Expedition in Pliny's Natural History

Kristiyan Simeonov
Department of Classics, Sofia University

Abstract: *This endeavor aims to create an innovative information extraction algorithm for Pliny's "Natural History." We used the state-of-the-art Python NLP library SpaCy and the Latin language models in LatinCy to develop a modern solution. The algorithm accepts a single lemma or a list of lemmas as input, producing a CSV dataset containing citations, context, and lemma variants. This facilitates efficient linguistic analysis of Pliny's work, initially focusing on Moesia and Thrace. We curated datasets on ethnonyms, places, mountains, and waterways. Using Streamlit and Matplotlib, we improved user interaction and visualization, aiding researchers in exploring ancient Thrace in Pliny's writings.*

Keywords: nlp, data science, Roman literature, data mining, lemma
Ключови думи: nlp, наука за данните, римска литература, извличане на данни, лема

Kristiyan S. Simeonov is an appointed Researcher R1 at the Department of Classics, Sofia University, holding a bachelor's degree in Classics and a double MA degree in Digital Humanities and Cybersecurity Management. He is a collaborator for CLaDA-BG, SUMMIT, and DiGi Thrace project.
Email: sergeevs@uni-sofia.bg,
ORCID: 0009-0007-3641-8432

INTRODUCTION[1]

The objective of our work is to develop an algorithm for data mining instances and mentions of Thrace in classical Latin literature. We stopped on the work of Plinius *Naturalis Historia* or Natural history, because of its encyclopedic nature. The work of Pliny is a massive source of information containing around half a million tokens.

Finding a suitable version of the text to turn into a dataset for mining was the first stage in our ongoing research. We opted after a prolonged search to use the Tesserae project data made available by Kyle P. Johnson in their GitHub repository[2].

Algorithm for Data Mining Naturalis Historia

An algorithm was written to convert the data available in the Tesserae project repository in the TESS file format into a CSV dataset. A dataset was constructed from the available textual data. When this first step of our project was

[1] The present paper is a result of the activities within the "Measuring Ancient Thrace" project no. КП-06-Н50/3 from 30.11.2020, financed by BNSF.
[2] *Burns* 2019: https://github.com/cltk/latin_text_tesserae

completed, a decision was made not only to be able to search a word, phrase, or a sentence in the work of Pliny, but with the ability to simultaneously search all the forms of a word, simply by typing its lemma. So instead of searching the list of words: *culina, culinae, culinarum, culinis, culinas*, etc. we can simply use *culina* as input to get all the forms of the word. To create such a data mining algorithm, we utilize one of the LatinCy models in the NLP library SpaCy[3]. But only getting a list of uses of a lemma in dataset didn't provide us with exhausting enough information, so we opted to enrich the algorithm with a chapter lookup function. This function would allow us to see which form of the word at hand is used and where it was used.

ENHANCEMENTS FOR EFFICIENT SEARCHING

To further enhance the data mining algorithm, we created a contextualization feature, which would also give the words surrounding the lemma we are searching for. An algorithm was created, based on the following equation:

context(i)=' '.join(tokens[j].text | j in range (max(0,i−5),min(len(doc),i+6)))

In which *i* is the index of the token at hand, the variable *tokens* simply represent the list of tokens in the document. To ensure that the context doesn't go out of bounds we use max(range(0, *i*-5)) and min(len(doc), *i* + 6)). The equation concatenates the text of selected tokens with a space between them, providing context around the token at index *i* with' '.join. By considering the surrounding tokens within a specified range, we aim to provide a nuanced perspective for exploring the dataset.

To optimize the time consumption of the algorithm we created a possibility to search for a list of lemmas in the dataset *e.g.*, input: *thracia, moesia, dacia*. With the computational power available in the LatinCy model, we managed to create an algorithm that not only searches for a word in a dataset, but provides tokenization, contextualization, book, and chapter lookup as well a multiple word search.

DATASET CONSTRUCTION AND ENRICHMENT

With the algorithm described in the first part of the paper we managed to create datasets of variable sizes concerning different aspects of the research at hand like ethnonyms - mainly names of tribes like the *triballi, getae, sapei* etc., places *abdera, maronea*, rivers - *strymon*, mountains: *rhodopa* and *haemus*, we also searched for prominent leaders in Thrace like Sitalces and Theres without a result. From the different prompts a curious find was documented: when we use the word *getae* which is plural and is used to name the tribe, no results were found, but when we modify the query and search for *geta*, the model yielded relevant results associated with the tribe.

A total of 73 preliminary entries about Thrace were found and collected into a dataset. Each section of the dataset represents a lemma, the context in which it appears, and the book/chapter from Pliny's work where it can be found.

The dataset provides insights into the cultural, historical and geographic landscape of ancient Thrace including entities, such as cites (e.g. *"abdera"*, *"cherronesum"*), rivers (e.g. *hebrus*), mountains (e.g., *"haemus"*), and ethnonyms (e.g., *"bessi"*, *"dardani"*). We may be inclined to think that the incorporation of diverse geographical entities, shows Pliny's comprehensive investigation of the province. The dataset's specific interest on Thrace and contextualized references in Pliny the Elder's work make it an important asset for researchers and specialists interested in finding more in-depth information regarding the province.

DISCUSSION AND INTERPRETATION

The lemma *abdera* appears in the context of geographical descriptions in book 4 chapter 27 and book 6 chapter 70. Similarly, *dardani* is mentioned in various contexts across different books, suggesting it might be a significant term concerning the larger context of Thrace. The

[3] *Burns, Bernhardt, Geelhaar, Koch*: https://huggingface.co/latincy/la_core_web_lg

lemma *thracia* appears frequently in the dataset and it is mentioned in a wide range of topics, including geography, climate, agriculture, and natural history. For instance, one excerpt mentions the fertility of Thrace because of cold and heat[4], while another discusses the Maronian wine originating from the coastal part of Thrace[5]. Pliny's Natural History is renowned for its extensive coverage across diverse fields, including history, botany, geography, and medicine. Within the dataset, references to the Strymon River shed light on the utilization of the tribulus plant, notably in the feeding of horses (*foliis tribuli equos saginant*)[6]. The historical practice of Thracians feeding tribulus to their horses offers intriguing insights into ancient customs. Furthermore, the usage of this plant in certain Asian cultures as a remedy for kidney stones[7] may underscore a shared understanding of medicinal plants. This observation aligns with scholarly consensus regarding the medical focus of Book XXII of *Naturalis Historia*[8].

This dataset could prove to be a valuable resource for studying Pliny's Natural History, particularly for understanding how he describes and categorizes the natural world. A visualization tool was created with Streamlit[9] and matplotlib[10] to give the end user of the dataset a method of exploring it, without any technical know-how.

STREAMLIT VISUALIZATION TOOL: ENHANCING DATASET EXPLORATION

Incorporating a Streamlit visualization tool has been instrumental in transforming raw dataset information into an interactive and user-friendly format. This tool provides a streamlined interface for users to explore and interpret the dataset generated through our algorithmic approach to mining instances of Thrace in Pliny's Naturalis Historia.

The tool's primary functionality includes the selection of CSV files, visualization of lemma frequency, and exploration of chapter-wise lemma mentions. Utilizing Plotly Express, the tool generates interactive and visually appealing plots[11], adding depth to the statistical insights it offers.

Users can select specific CSV files, such as "allData.csv", "places.csv", "ethnonyms.csv", and more, allowing for a focused exploration of different aspects of the dataset concerning Thrace. The tool provides basic statistics on lemma frequency and token count, giving users a quick overview of the dataset's composition. Interactive plots, including a bar chart illustrating lemma frequency and a pie chart depicting the distribution of lemma frequencies as can be seen on **Fig. 1**, enable users to have insights into usage of specific lemmas within the dataset. Chapter-wise lemma mentions are visualized through a bar chart, offering a detailed view of lemma occurrences across different sections of Pliny the Elder's work.

To facilitate a more in-depth exploration, the tool includes an expander feature. Users can click to view detailed context information for each lemma, including the lemma itself, the corresponding book/chapter, and the context in which the lemma appears as can be seen on **Fig. 2.**

The Streamlit data visualization tool enables us to transform data into an interactive format, encouraging users, researchers, and enthusiasts to engage in a more insightful exploration of Pliny the Elder's *Naturalis Historia*.

CONCLUSION

Our study offers a thorough method for locating references to Thrace in classical Latin literature, with an emphasis on Pliny the Elder's *Naturalis Historia*. A feature-rich dataset and the created algorithm have produced insightful information about the geographical, historical, and cultural features of ancient Thrace.

[4] Plin. Nat. 17.5.
[5] Plin. Nat. 14.16.
[6] Plin. Nat. 22.12.
[7] *Kamboj, Aggarwal, Singla, Puri* 2011: 154.
[8] *Doody* 2010: 9.
[9] Streamlit: https://streamlit.io/
[10] *Hunter* 2007: 90-95.
[11] *Plotly* 2015: https://plotly.com/python/plotly-express/

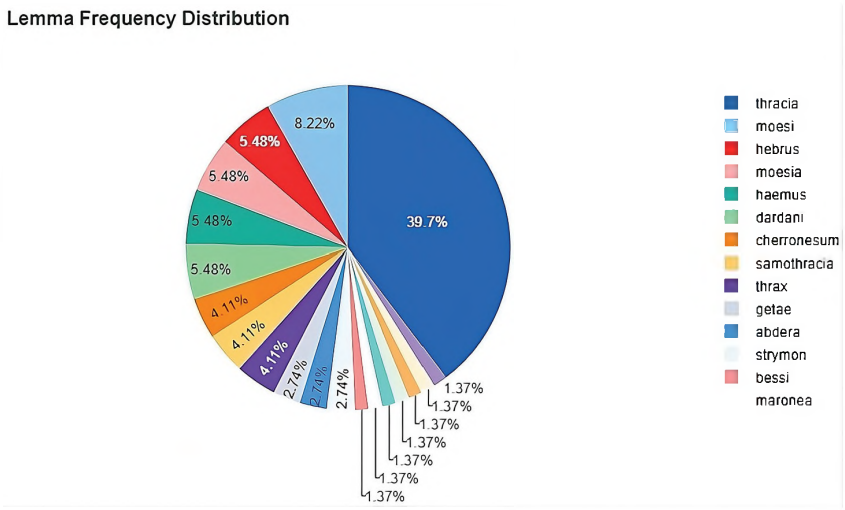

Figure 1. Plotly Express pie chart representing the lemma frequency distribution in the whole dataset made from Pliny's *Naturalis Historia.* (Author: K. Simeonov).

Lemma: haemus

Book/Chapter: plin. nat. 31.19

Context: arborum alimenta consumebant sicut in haemo obsidente gallos cassandro cum valli

Lemma: hebrus

Book/Chapter: plin. nat. 4.19

Context: sapaeos odomantos odrysarum gens fundit hebrum accolentibus carbiletis pyrogeris drugeris caenicis

Lemma: hebrus

Book/Chapter: plin. nat. 4.19

Context: minores rhodopae subditi inter quos hebrus amnis oppidum sub rhodope poneropolis

Lemma: hebrus

Book/Chapter: plin. nat. 4.20

Context: gygemeros meritus melamphyllos flumina in hebrum cadentia bargus syrmus macedoniae

Figure 2. Streamlit visualization of the output from the Python text mining algorithm showing lemma, book/chapter references from *Naturalis Historia,* and the context in which the lemmas were used. (Author: K. Simeonov).

We enhanced the dataset with a range of features, such as contextualization, chapter lookup functions, and lemma-based searches, by algorithmically transforming the Tesserae Project data into a CSV format. The dataset offers a thorough examination of place names, rivers, mountains, and ethnonyms connected to Thrace, consisting of 73 preliminary entries. The proposed context formula, which considers the index of the current token and its surrounding context, has proven to be a valuable tool for exploring the dataset with a nuanced perspective. The Streamlit tool enhances the usability of the dataset, ensuring that it can be leveraged by a broader audience for research and educational purposes.

Our findings demonstrate the significance of considering different word forms in data mining endeavors, as highlighted by intriguing nuances in language usage. Despite not aiming to present an exhaustive list, our work highlights the potential for further research and exploration in this domain.

FUTURE PERSPECTIVES AND PLANS

Looking forward, our plans involve enhancing the algorithm to be able to identify named entities and further optimization would be made to reduce its search speed. We recognize the importance of creating a more comprehensive primary dataset containing the whole classical Latin literature, thus enabling us to do a more complete search process. With this enhancement we would make the algorithm a versatile tool applicable to multiple areas of authors and interests.

BIBLIOGRAPHY:

Primary Sources

Naturalis Historia. Pliny the Elder. Karl Friedrich Theodor Mayhoff. Lipsiae. Teubner. 1906.

Secondary Sources

Doody 2010: *Doody*, Aude. Pliny's Encyclopedia: The Reception of the Natural History. Cambridge: Cambridge University Press.

Hunter 2007: *Hunter*, John D. Matplotlib: A 2D Graphics Environment. – Computing in Science & Engineering, vol. 9/3, 90-95.

Kamboj, Aggarwal, Singla, Puri 2011: *Kamboj*, Parul, Milan *Aggarwal*, Sugam *Singla*, Sanjeev *Puri*. Effect of Aqueous Extract of Tribulus Terrestris on Oxalate-Induced Oxidative Stress in Rats. – Indian Journal of Nephrology, No. 21/3, 154–159.

Online resources

Burns 2019: *Burns*, Patrick J. Tesserae Project, Classical Language Toolkit. https://github.com/cltk/latin_text_tesserae (accessed 07.06.2024).

Burns, Bernhardt, Geelhaar, Koch: *Burns*, Patrick J., Nora *Bernhardt*, Tim *Geelhaar*, Vincent *Koch*. spaCy. la_core_web_lg, version 3.7.2. https://huggingface.co/latincy/la_core_web_lg (accessed 06.06.2024).

Plotly 2015: Plotly Technologies Inc. Collaborative data science Publisher. Montréal, QC. https://plotly.com/python/plotly-express/ (accessed 08.06.2024).

Streamlit: Streamlit. The Fastest Way to Build Custom ML Tools. https://streamlit.io/ (accessed 07.06.2024).

Разплитане на нишките на Тракия: Извличане на данни от „Естествена история“ на Плиний Стари

Кристиян Симеонов

Целта на изследването е да се разработи алгоритъм за извличане на информация за Тракия в текстове на класическата римска литература, като се фокусира върху „Естествена история“ на Плиний Стари. Алгоритъмът, който може да се адаптира за всяко произведение на класически латински език, преобразува данните от формата TESS в CSV набор от данни, което дава възможност за търсене по лема и разпознаване на контекста на съответната глава от произведението с помощта на NLP моделите LatinCy от библиотеката SpaCy. Добавената функция за контекстуализация показва съседните на търсената лема думи. Алгоритъмът е оптимизиран за търсене по няколко леми, поддържа функция за токенизация, контекстуализация и търсене на книги/глави от произведението. Наборът от данни, създаден с помощта на алгоритъма, включва 73 записа, свързани с етноними, топоними, реки и планини в Тракия, което дава възможност за културни, исторически и географски наблюдения. Инструмент за визуализация, разработен с програмната библиотека Streamlit, осигурява лесен достъп до данните, като чрез интерактивни графики позволява на потребителите да изследват честотата на лемите, споменаванията на глави и контекста. Изследването предоставя метод за намиране на споменавания на Тракия в труда на Плиний, като в бъдеще се планира да се подобри функцията за разпознаване на записите и да се създаде цялостен първичен набор от данни за класическата римска литература.