Mapping India in Pliny the Elder's *Natural History*

Dawn, Lizao Zhuang (r0914937)

Abstract

this is an abstract

Contents

List of Figures

List of Tables

1 Introduction

1.1 *Natural History* and its complexity

Pliny the Elder’s *Natural History* is widely recognized as the earliest encyclopedia in the world, manifesting a pioneering effort in comprehensively cataloging the vast array

of human knowledge from that era.

The work is thematically divided into 37 books, covering a diverse range of subjects including astronomy, geography, zoology, botany, medicine, and more. Pliny meticulously consulted a wide range of Greek and Roman references, totaling approximately 2,000 volumes[1], and interwove his own literary interpretation or comments to the narratives.

Despite the carefully designed knowledge-ordering framework (Lao 2016), scholars have observed a paradoxical complexity in *Natural History*, evident in its linguistic style, narrative approach, and use of references. The work compiles inconsistent toponyms from Greek and Latin, includes digressions in descriptions (Roller 2022), exhibits changes in vocabularies and sentence structures (Pinkster 2005). However, it is precisely this complexity that makes the work more fascinating and not only a valuable source to the knowledge and worldview of the ancient world, but also a gateway into Pliny's conceptualization, imagination, and even the prevailing imperial ideology.

The complexity and interconnectivity of the general structure of *Natural History* is further highlighted in different aspects by refreshing approaches. In terms of content organization of the work, Healy (1999) vandicated Pliny's original contribution in unveiling the technology and science engagement of the Rome Empire from the description about natural phenomena and scientific experiment to the development of scientific language in Latin, taking the historical, political and liguistic context into consideration. And Naas (2002) discussed how Pliny formulated the the diversed materials into his encyclopaedic structure, revealing the work's multifaceted nature as an epistemological, ideological, and moral project. By analysing Pliny's employment of the historical exemplum in the work, Schultze (2011) argues how the specific literary device directed and teased the readers and established a profound connection between human beings and the entire spectrum of nature in *Natural Hisroty*.

In addition to the close reading methods used in the prior analyses of the context and references in *Natural History*, Rydberg-Cox (2021) employs network analysis method with different metrics to map the interrelationships between Pliny's sources and the topics discussed in the work. Furthermore, Fantoli (2022) presents a comparative study of book 2 of *Natural History* and book 7 of Seneca's work *Natural Questions*, both centered on astronomy, utilizing statistical analysis to identify Pliny's unique stylistic features based on variations in their discourse distribution, and proved the encyclopedic authorial intent shown in *Natural History* with correspondence and tree analysis. These two studies also demonstrate how distant reading methodologies offer novel insights into the understanding of ancient treatises.

[1] *Natural History* 1.5.1 (https://topostext.org/work/148)

1.2 Spatial perstpective in *Natural History*

As pointed out by Beagon (2011), differenciating from his pressedors, Pliny showed a “terrestrial curiosity” in *Natural History*, emphasizing a recognition of the physical, material world. In this regard, the vision of geography plays a pivotal role in distributing information, knowledge, and events throughout *Natural History*.

Drawing from the long-established topographical and ethnographic traditions, Pliny seamlessly connects volumes dedicated to geography (books 3-6) with broader elements, activities, and cultural, historical, and societal contexts(Roller 2022), exemplified in his portrayal of exotic plants, communities’ habitats, imperial expeditions, and trade ventures. In other words, geographical names occured in each book of *Natural History* served as signposts guiding readers through diverse lands, shedding light on how Pliny and his contemporaries perceived and conceptualized the world around them.

A normalized frequency of place name occurence in the work is calculated as the ratio of counts of the occurrences of place names in each book to the word lengths of the book (Table 1). The bar chart (Figure 1) depicted the comparison of distribution of place names in the books of *Natural History*. The observation is in line with content structure of *Natural History*, that books 3-6 centered around the themes of “Geography and ethnography”, contains the most mentions of location names, and place names are also frequently referred in books about agriculture and horticulture (book 12-14), aquatic life (book 31), and mining and mineralogy (book 34-37).

Table 1: Normalized distribution of place names in *Natural History*

Book	Total_length	Place_count	Place_freq
1	2778	1	0.000360
2	30570	406	0.013281
3	18037	1007	0.055830
4	15434	1309	0.084813
5	18872	1112	0.058923
6	27891	1012	0.036284
7	21204	225	0.010611
8	24176	185	0.007652
9	19197	140	0.007293
10	20816	121	0.005813
11	27345	77	0.002816
12	13906	188	0.013519
13	13243	164	0.012384
14	15277	189	0.012372

Book	Total_length	Place_count	Place_freq
15	14552	135	0.009277
16	25442	180	0.007075
17	29387	82	0.002790
18	35850	222	0.006192
19	18822	146	0.007757
20	22743	21	0.000923
21	17896	95	0.005308
22	16491	24	0.001455
23	15764	17	0.001078
24	17491	56	0.003202
25	16734	85	0.005079
26	15448	35	0.002266
27	12444	40	0.003214
28	26476	28	0.001058
29	13976	31	0.002218
30	14395	23	0.001598
31	12204	222	0.018191
32	14635	76	0.005193
33	17946	113	0.006297
34	18972	193	0.010173
35	21283	277	0.013015
36	21295	357	0.016764
37	22255	282	0.012671

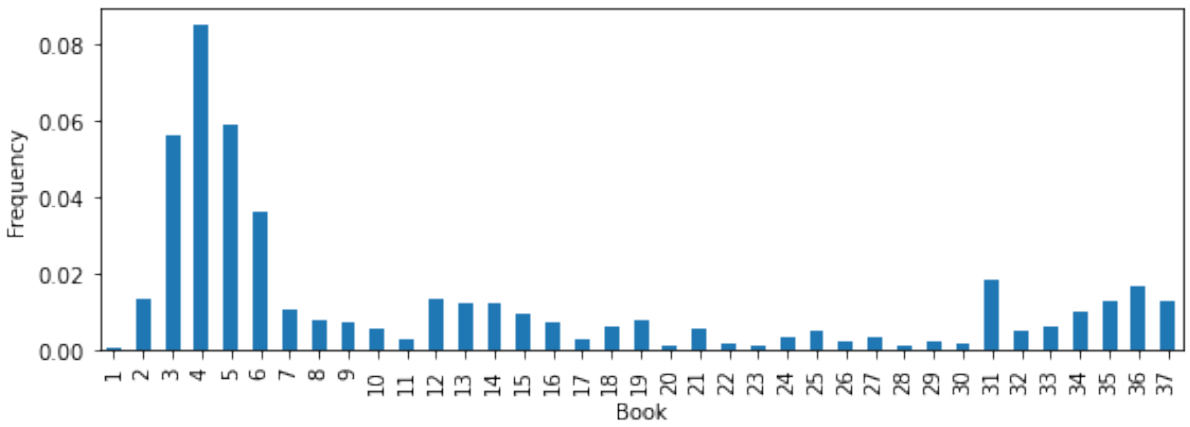

Figure 1: Normalized distribution of place names in *Natural History*

1.3 Text source for the study

Natural History is originally written in Latin. For the purpose of this study, an English translation conducted by Henry T. Riley (1816-1878) and John Bostock (1773-1846), which was first published in 1855, is utilized. The translated text is obtained in a digitized version from the TOPOSText project, having been sourced from the Perseus Project and governed by a Creative Commons Attribution-Share-Alike 3.0 U.S. License.

Annotations of people's name, places' name and geographical coordinates are available together with the text of *Natural History* (Book1-11, Book12-37) on TOPOSText project. This invaluable resource allows for the creation of a dataset that includes both the textual contents and geographical annotations, which can be utilized to investigate the distribution of place names in the entire text and examine the frequencies and patterns of geography-related content.

The extension of the extracted corpora and the workflow of the extraction will be further explained in the Methodology chapter (Section 3).

2 Research Question

2.1 Prominent mentioned places in *Natural History*

Based on the geographical annotations in *Natural History* provided by TOPOSText project, there are 2052 unique places mentioned in *Natural History*.

The top 20 most frequent place names mentioned (as 1% of total) in *Natural History* is shown in Table 2.

Table 2: Top 20 mentioned place names in Natural History

	ToposText_ID	Place_Name	Lat	Long	Count
1687	https://topostext.org...	Italy	40.6000	16.30000	292
2034	https://topostext.org...	Rome	41.8910	12.48600	269
52	https://topostext.org...	Egypt	27.1000	30.70000	261
82	https://topostext.org...	India	30.0000	74.00000	167
57	https://topostext.org...	Arabia	28.0000	40.00000	123
320	https://topostext.org...	Syria	35.5000	39.00000	109
255	https://topostext.org...	Cyprus	35.0000	33.00000	85
109	https://topostext.org...	Nile	30.0918	31.23130	85
2282	https://topostext.org...	Alps	44.1420	7.34300	82
766	https://topostext.org...	Sicily	37.6000	14.50000	71
275	https://topostext.org...	Crete	35.2052	25.18360	64

	ToposText_ID	Place_Name	Lat	Long	Count
7	https://topostext.org...	Ethiopia	13.0100	35.01000	58
417	https://topostext.org...	Rhodes	36.4408	28.22440	56
966	https://topostext.org...	Athens	37.9718	23.72793	56
2043	https://topostext.org...	Capitol	41.8933	12.48300	52
298	https://topostext.org...	Euphrates	35.2791	40.27080	47
2241	https://topostext.org...	Pontus	43.5000	33.50000	47
1839	https://topostext.org...	Campania	41.1000	14.60000	46
1480	https://topostext.org...	Armenia	39.7020	44.29800	45
17	https://topostext.org...	Red Sea	19.5000	39.00000	42
545	https://topostext.org...	Carthage	36.8500	10.32000	42
602	https://topostext.org...	Cilicia	37.0100	34.01000	42

The place names referenced in *Natural History* are geographically mapped, with each location marked on the map using its corresponding coordinates. A dot is assigned to represent each place, with the size and color of the dot reflecting the frequency of its mention in the book. The larger and darker the dot, the more frequently the place is referenced within the context of Natural History.

An intriguing observation from the output, as depicted in Figure 2, is the prominence of India, a region outside the Mediterranean, despite its high frequency of mentions.

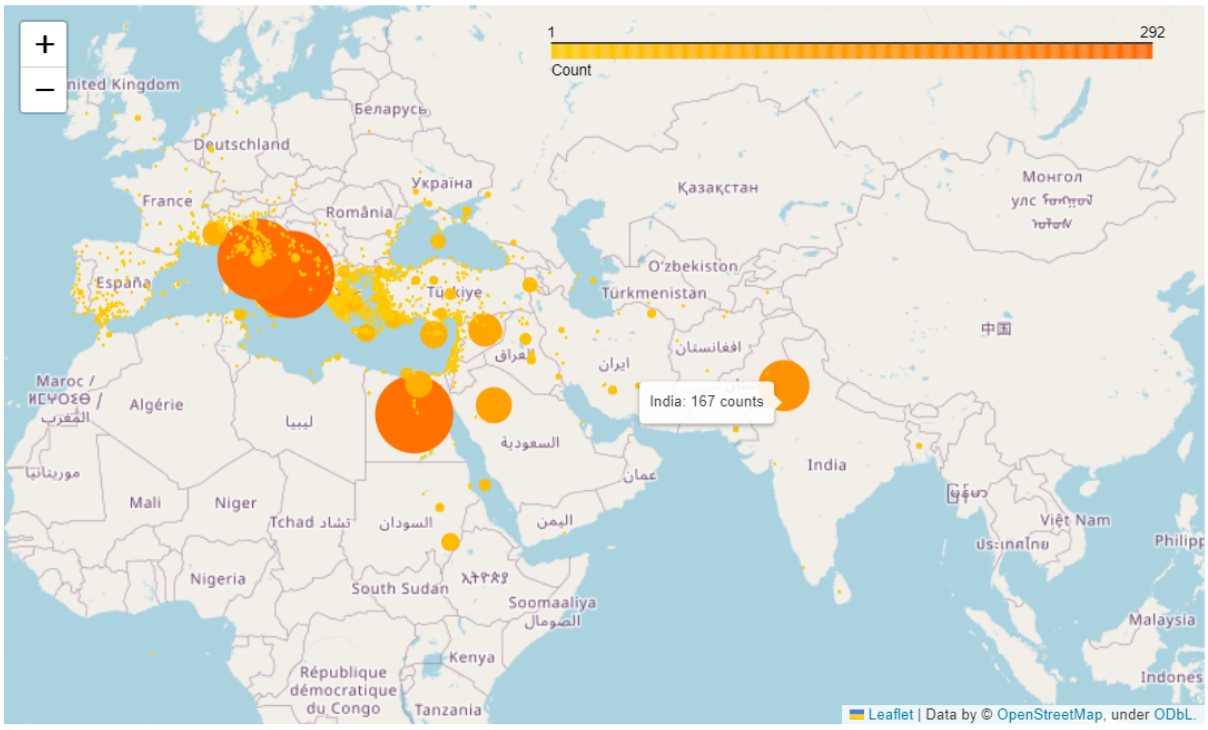

Figure 2: Place name distribution map

2.2 Why India?

Geographically, India presents itself as a distant and disconnected territory from the Roman Empire, lacking any direct aquatic or land routes with the Mediterranean region. Despite this apparent physical separation, the exotic curiosity Pliny attempted to integrate, as well as the Indo-roman goods exchange network reflected in the work, may contribute to an explaination of the prominent mentioning of India in *Natural History* as the broader context.

As suggested by (Murphy 2003), the *mirabilia*, encompassing accounts of extraordinary landscapes, peoples, plants, and animals, assumes a substantial proportion within the books of *Natural History*. Pliny's inclusion of such exotic elements not only catered to the prevailing curiosity of his Roman readers but also fostered a comparative perspective between distant locales, exemplified by his references to India, and their natural counterparts within Rome (Naas 2011). Within research framework of Roman Imperialism, the detailed portrayal of foreign lands, such as India, holds significant importance in shaping both Pliny's and his contemporary Roman readers' perception of their place within the global landscape (Pollard 2009).

In addition, *Natural History* serves as a valuable reference for tracking the Indo-Mediterranean network of exchange (Pollard 2009). Through the depiction of cities, ports, and rivers along the trade routes, the work provides substantive evidence of the flourishing trade relations between the Roman Empire and the Indian subcontinent (Neelis 2011). The extensive exemplify of diverse commodities, such as gemstones, glass, spices, textiles, plants, wine, along with the accounts of the currency *sestertii* involved in the merchandise exchange in the work shed lights to the compelling details and social and cultural implications of this long-distance trade (Székely 2006; Pollard 2009). Furthermore, the direct criticisms regarding the high cost for the luxury items imported from India implies both the magnitude of the trade volume and Pliny's stance towards this commercial interaction (Neelis 2011).

2.3 India-related text as a case study

In light of the observations and foundational research mentioned above, the present study centers its investigation on the spatial perspective within Pliny's *Natural History*, with a specific focus on the texts pertaining to India, seeking to delve into the discourse surrounding this region. To achieve this goal, distant reading methodologies, including statistical analysis, topic modeling, and social network analysis, will be employed.

The main aim of this study is to explore how is India described, and how is the information about India structured in *Natural History*, which may also contribute to a more profound comprehension of the inherent complexity and interconnectivity that permeates this monumental work.

3 Methodology

3.1 Workflow

The workflow for this study involved the following key stages:

Data Collection:

As mentioned in the Introduction chapter (Section 1), the text employed for this study is obtained from the digitized English translation (by Henry T. Riley (1816-1878) and John Bostock (1773-1846)) of Pliny's *Natural History* available on TOPOSText project.

The two parts of *Natural History* (Book1-11, Book12-37) are scraped for their the textual contents together with the annotated information of the geographical coordinates of the ancient places mentioned in the work, and the book, chapter and paragraph affiliations with the function provided in Beautiful Soup library of Python.

Data Preprocessing:

The information extracted from the html is structured into separate columns as Pandas dataframe, a dataframe for plain text of the entire work, and a dataframe for geographical-related text in *Natural History* with the geographical annotations are generated and stored in CSV format respectively.

After a preliminary exploration, the research focus is narrowed down to India-related text in *Natural History*. With a reference to the geographical territories in the consideration of ancient Greek and Roman world (Talbert 2000b), a dataframe for India-related text is filtered from the abovementioned dataframe for geographical-related text with the range of geographical coordinates of India subcontinent in the era of *Natural History*. The flitered India-related text dataframe is also stored in CSV format.

The location names mentioned in the India-related text were checked manually for its completeness. If any location names were identified and not being annotated in the TOPOSText, they were added to the India-related text dataset.

Additionally, the textual contents in the datasets were processed to make them suitable for textual analysis. This processing involved tokenization, lemmatization and the exclusion of stop-words.

Data Analysis:

Statistical analysis is conducted in the preliminary exploration of the extracted dataframes. A nomalized frequency of geographical name occurence in each book is calculated for an overview of the place name distribution in *Natural History*. And the top 1% prominently mentioned place names in the entire work are sorted out with the

time of their occurencies. The specific attention on India-related text as a case study is drown from this initial observation.

In the analysis of the India-related text (target corpus) in *Natural History*, three analysis methods are employed:

1. Word frequency: single word frequency and bi-gram collocation of the target corpus are measured with the functions in NLTK package for an overview of the keywords relating to India in *Natural History*.

2. Topic modeling: Genism library is used for semantic vectorization and implemetion of Latent Dirichlet Allocation (LDA) model for the topic modeling of the India-ralated text, and the library of pyLDAvis is utilized for an interactive visualization. The output of this method shows the potential topics in the India-related text in *Natural History*.

3. Network analysis for Named Entities: Person names mentioned in the target corpus are retrieved from the tagging of the text given by the pretrained multilingual Named Entity Recognition model Flair. The person name entities are cross checked with the annotation on TOPOSText. Stone names, river names, mountain names, person names and the book number are extracted as nodes, and the co-occurence between the nodes are calculated as edges for network analysis. The output of this method is a graph showing the clusters of the nodes in the target corpus, indicating the structure of the content related to India in *Natural History*.

Interpretation and Conclusion:

The workflow and parameter setting of each research method is explained in the beginning of each analysis section. The results aquired from each method is interpreted with a dialouge to the broader literature and close reading of the related text.

In the Conclusion chapter, the findings are illustrated comprehensively in the context of the research questions. And the limitations of each method is discussed and evaluated.

3.2 Data preparation

The present section provides an overview of the data preparation process, encompassing three sub-sections: HTML scraping from TOPOSText, creation of a filtered dataset of “India-related text,” completeness checks and preprocessing of textual data. The tools and procedures employed in data collection and dataset generation for the study are elucidated in the subsequent content.

3.2.1 HTML scraping from TOPOSText

As previously stated, the textual contents of Pliny's *Natural History* are available on the TOPOSText project, presented in two distinct parts: Book1-11, Book12-37. Both parts are provided in HTML format, offering separate sections of the complete work.

To extract the relevant data, the Beautiful Soup tool, a Python library renowned for parsing HTML and XML documents, was employed. This process involved navigating the HTML structure effectively to retrieve essential information.

The text in the HTML documents is organized into paragraphs, each uniquely identified by an "id" attribute that specifies its corresponding book, chapter, and paragraph number. For instance, a typical paragraph has an "id" tag as follows:

<p id='urn:cts:latinLit:phi0978.phi001:3.9.7'>

Utilizing these "id" attributes, the paragraphs were meticulously associated with their respective book, chapter, and paragraph information.

As a result of this data extraction process, a reference dataset was obtained, comprising the plain text of *Natural History* divided into paragraphs, with each paragraph assigned a unique identifier, and separate columns indicating its affliated book, chapter, and paragraph number. An illustrative example of the dataset's structure can be refered as Table 3.

Table 3: Example for the reference dataset containing the plain text in paragraphs of *Natural History*

	UUID4	Reference	Book	Chapter	Paragraph	Text
0	e9e67565-bb...	urn:cts:lat...	1	1	1.0	PREFACE IN ...
1	010b853d-b8...	urn:cts:lat...	1	2	1.0	But who cou...
2	2d10e332-9c...	urn:cts:lat...	1	3	1.0	But if Luci...
3	113e0b4c-5b...	urn:cts:lat...	1	4	1.0	My own pres...
4	19115032-9f...	urn:cts:lat...	1	5	1.0	For my own ...

There are a total of 3493 paragraphs in the English translated version of *Natural History* used in this study. The extracted text contains 343096 tokens and 28606 types after preprocessed. This reference dataset has been saved in CSV format for record.

Moreover, the geographical annotations concerning the ancient places mentioned in the text are labeled with a class attribute denoted as "place", exemplified by the following HTML code snippet:

<a about="https://topostext.org/place/419125LPal" class="place" lat="41.8896"

long="12.4884">Palatine

To compile a comprehensive dataset encompassing all the annotated ancient places, along with their corresponding geographical coordinates and contextual information (such as book, chapter, and paragraph numbers), all annotations under the "place" class are extracted. This dataset enables an analysis of the distribution of place names within *Natural History*.

As certain places may possess multiple names, ToposText_ID, which is the unique identifier assigned to distinct places available on TOPOSText is also extracted as a reference information. An example of the dataset presenting the geographical-related text in *Natural History* is provided in Table 4 for reference.

Table 4: Example for the geographical-related text dataset

	UUID4	ToposText_ID	Place_Name	Reference	Lat	Long	Book	Chapter	Paragraph	Text
0	bf12...	http...	Academy	urn:...	37.9920	23.7070	1	8	1.0	For ...
1	f782...	http...	Pala...	urn:...	41.8896	12.4884	2	5	1.0	For ...
2	a0f9...	http...	Esqu...	urn:...	41.8950	12.4960	2	5	1.0	For ...
3	b8d8...	http...	Capitol	urn:...	41.8933	12.4830	2	5	1.0	For ...
4	f81b...	http...	Rome	urn:...	41.8910	12.4860	2	6	3.0	Belo...

According to the geographical annotations of the ancient places occured in *Natural History*, there are 5595 occurences of place names in book 1-11 and 3281 in book 12-37, adding up to a combined total of 8876 annotated places throughout the work. The geographical-related text in *Natural History* contains 199507 tokens and 23937 types after preprocessed. This dataset including place names and their textual context in *Natural History* is saved in CSV format for record.

3.2.2 Filtered dataset of "India-related text"

As outlined in the Research Question chapter (Section 2), this thesis examines texts concerning the Indian region in Pliny's *Natural History* as a case study. The objective is to explore how India is described, portrayed, and imagined within this extensive work, providing valuable insights into its complexity.

To ensure a comprehensive contextual analysis, the dataset creation considers not only instances where the word "India" is directly mentioned but also text related to the Indian region. This broader approach aims to encompass a wider scope of relevant information. Drawing from the research and mapping of the Indian region in the perception of the ancient Greek and Roman world, as explained and manifested in the *Barrington Atlas of the Greek and Roman World* (Talbert 2000a, 2000b), the approxi-

mate coordinates defining the target region are as follows[2]:

- Latitude: 5-35 degrees North
- Longitude: 65-95 degrees East

Utilizing the aforementioned dataset of geographical-related text in *Natural History*, the text having annotations with geographical coordinates falling within the specified range are extracted to construct a dataset relevant to the discourse about Indian region in the work. The filtering process ensures not only the text explicitly mentioning "India" but also those including other place names situated within the defined boundaries of the Indian region were retained.

The new dataset comprises the textual content as well as the geographical coordinates of the mentioned Indian place in *Natural History*. An example of the structure of the dataset of India-related text is showed as Table 5.

Table

	UUID4	ToposText_ID	Place_Nar
85	94ea9a38-cc2c-4765-bf4e-feec573b3775	https://topostext.org/place/300740RInd	India
92	01a2d650-13d5-4609-a716-359c673c57f6	https://topostext.org/place/300740RInd	India
93	edb23c91-5745-4982-abb1-d63878a5e8dd	https://topostext.org/place/300740RInd	India
218	afa83f27-a731-4277-9471-91d1c239d229	https://topostext.org/place/254683WInd	Indus
343	37c891bf-f779-4fe1-8af5-a31ff4edf190	https://topostext.org/place/300740RInd	India

There are 229 occurrences of paragraphs mentioning the places in Indian region with geographical coordinates annotation. And the distinct places mentioned are ['India' 'Indus' 'Ganges' 'Acesinus' 'Hydaspes' 'Taprobane' 'Arachosia' 'Muziris' 'Baragaza' 'Ceylon']. The textual content pertaining India compiles 18029 tokens and 5384 types after preprocessed. The dataset and corpus for India-related text in *Natural History* are saved respectively in CSV format for further reference.

3.2.3 Data completeness check

The paragraphs extracted from the India-related text dataset undergo manual verification for the completeness of Indian place name annotations. Each distinct paragraph in the dataset is individually extracted and stored in TXT format as separate files within a corpus folder. The file names contain information about the affiliating book, chapter, and paragraph numbers.

[2]As indicated in the map-by-map directory, the range spans territories of "modern states of India (minus the Punjab), Bangladesh, Bhutan, Burma, Nepal, and Sri Lanka".

There are in total 146 distinct paragraghs metioning India places in *Natural History* accroding to the annotations on TOPOSText.

An example of the exported file name can be referred as follows:

`Exported india_corpus\37.77.1_text.txt`

The text files are uploaded to Recogito platform, which offers a semantic annotation tool and automatic geographical annotation suggestions from its supported gazetteers. This process is used to find Indian place names mentioned in the text paragraphs related to India that were not annotated in TOPOSText. These unidentified place names are then marked on the Recogito workspace with the available geographical coordinates information as additional annotations. And the identified annotations are exported in CSV format for supplement to the dataset of India-related text in *Natural History*.

As shown in Table 6, the supplement annotations are organized in the following manner:

FILE: This column contains the name of the file indicating the book, chapter, and paragraph number where the mentioned place name appears.

QUOTE_TRANSCRIPTION: This column contains the textual name of the place as mentioned in the text.

URI: The URI column contains the geographical information obtained from the gazetteers available on the Recogito platform. The URI provides a unique identifier for the specific location.

VOCAB_LABEL: This column contains the confirmed automatically matched geographical name with the corresponding place name mentioned in the text.

LAT&LNG: The LAT and LNG columns represent the geographical coordinates (latitude and longitude) associated with the marked place name. Note that some marked names may not have matching coordinates.

PLACE_TYPE: This column contains the automatically matched geographical role provided by the gazetteers. It describes the type of place the name represents.

VERIFICATION_STATUS: The VERIFICATION_STATUS column indicates whether the place names have been “verified” with confirmed coordinates that match the gazetteers’ information.

COMMENTS: The COMMENTS column includes manual remarks for the place names that do not have matching coordinates but are believed to indicate Indian place names

based on the context.

Table 6: Examp

	FILE	QUOTE_TRANSCRIPTION	TYPE	URI	VOCAB
0	2.75.1_text.txt	hypasis	PLACE	http://pleiades.stoa.org/places/60110	Zadadro
3	6.21.4_text.txt	sydrus	PLACE	http://pleiades.stoa.org/places/60110	Zadadro
4	6.21.4_text.txt	rhodapha	PLACE	http://pleiades.stoa.org/places/60019	Rhodop
5	6.21.4_text.txt	palibothra	PLACE	http://pleiades.stoa.org/places/59978	Paliboth
6	6.21.5_text.txt	prinas	PLACE	http://pleiades.stoa.org/places/60008	Prinas (

```
PLACE_TYPE
river                   14
settlement              13
island                   6
unknown                  4
cape                     3
mountain                 2
people                   2
lake                     1
unlocated                1
unlocated,river          1
unlocated,settlement     1
Name: UUID, dtype: int64
```

After the mannual annotation process, 56 Indian place names were identified and can be added as supplementary annotations to the existing dataset, most of which are names of rivers, settlements, and islands. Among these, 45 place names have confirmed geographical coordinates based on the reference in Recogito. For the other 11 place names, though have no matching coordinates on Recogito, there are contextual clues indicating that they are probably Indian location names.

The supplemented place name annotations were added to the India-related text dataset. The updated dataset contains 285 occurrences of paragraphs mentioning Indian places. And the distinct places mentioned are [‘India’ ‘Indus’ ‘Ganges’ ‘Acesinus’ ‘Hydaspes’ ‘Taprobane’ ‘Arachosia’ ‘Muziris’ ‘Baragaza’ ‘Ceylon’ ‘Hypasis’ ‘Sydrus’ ‘Rhodapha’ ‘Palibothra’ ‘Prinas’ ‘Cainas’ ‘Condochates’ ‘Erannoboas’ ‘Cosoagus’ ‘Sonus’ ‘Protalis’ ‘Peucolaitis’ ‘Taxilla’ ‘Modogalinga’ ‘Andarae’ ‘Dardae’ ‘Methora’ ‘Chrysobora’ ‘Dandaguda’ ‘Tropina’ ‘Patala’ ‘Capitalia’ ‘Automula’ ‘Amenda’ ‘Cantaba’ ‘Prasiane’ ‘Argyre’ ‘Crocala’ ‘Bibraga’ ‘Toralliba’ ‘Hippuros’ ‘Palaesimundus’ ‘Megisba’ ‘Palesimundus’ ‘Cydara’ ‘Coliacum’ ‘Emodian mountains’ ‘Capisa’ ‘Parabeste’ ‘Cartana’ ‘Tonberos’ ‘Arosapes’ ‘Gedrusi’ ‘Arbis’ ‘Sigerus’ ‘Catarcludi’ ‘Meros’ ‘Perimula’ ‘Chenab’ ‘Oratae’].

3.2.4 Preprocessing of texts

The textual contents stored in the “TEXT” column of the mentioned datasets are utilized as corpora for different analyses with three distinct scales: the entire work’s text, text specifically related to geographical content, and text related to Indian content. To prepare the data for analysis, a preprocessing process is applied using a defined function, which employs tools from the NLTK package.

During the preprocessing, the texts are tokenized, preserving punctuation marks, and lemmatized to their base forms. Furthermore, common English stopwords are excluded from the corpus, considering the text is in an English translation version. To reduce noise of short strings, tokens with length lower than two will not be appended to the output token list. The output of this preprocessing is a refined corpus presented as a nested list structure, with paragraphs forming the smallest nesting unit.

The size computed for each corpus mentioned earlier corresponds to the outcome of this preprocessing procedure. By preprocessing the data, the corpora are optimally organized, ensuring that they are conducive to meaningful analyses and facilitating the extraction of valuable insights from the text at varying scales.

4 Data Analysis

4.1 Place name distribution in India-related text

The comparison between the total number of place names and the place names specifically related to the Indian subcontinent mentioned in each book, is depicted in Figure 3. The difference in numbers between the two categories is significant, as indicated by the large disparity.

To facilitate a more effective comparison of the referencing trends across different books, Figure 4 presents subplots with varying y-axis scales. This approach allows for a clearer visualization of the trends and patterns in place name references throughout the various books.

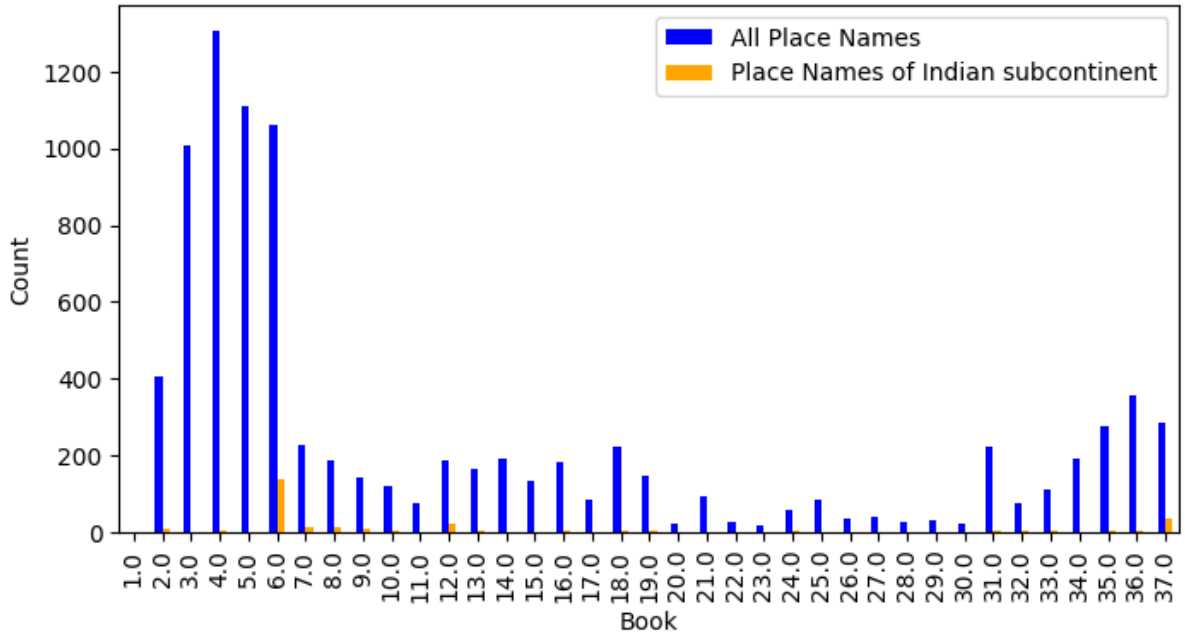

Figure 3: Occurence count for all place names and place names of Indian subcontinent in each book

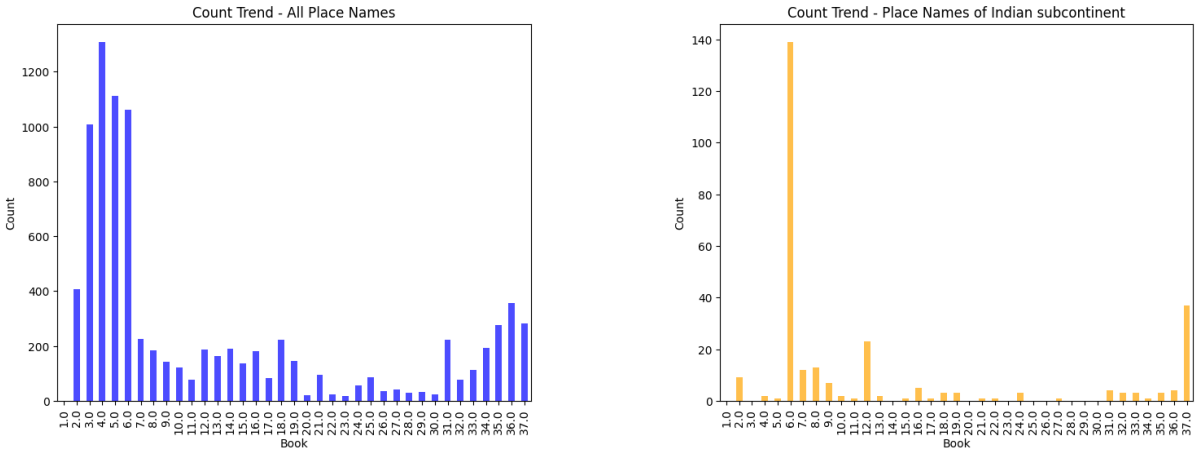

Figure 4: Occurence count for all place names and place names of Indian subcontinent in each book_different y-axis scales

The figures reveal a distinct difference between the occurrence trends of place names related to the Indian subcontinent and all place names collectively. Specifically, the referencing of the Indian subcontinent is highly concentrated in books 6, 12, and 37 of Pliny's narrative. This discrepancy indicates that the mentioning of place names from the Indian subcontinent is closely tied to specific themes and topics within Pliny's work.

In this regard, three methodologies have been employed to analyze the texts pertaining to the Indian subcontinent in *Natural History*, including word frequency and collocation analysis, topic modeling, and network analysis. The objective of these analyses is to

delve deeper into the textual content, unraveling the intricate relationships and uncovering the underlying themes and connections associated with the place names of the Indian subcontinent.

Through word frequency and collocation analysis, the aim is to identify keyword and significant word combinations co-occur in the textual content about India in *Natural History*. This analysis provides insights into the specific linguistic patterns and contextual associations surrounding the Indian places mentioned in the work, providing an overview of the keyword in the discourse.

Topic modeling allows for a broader exploration of the thematic landscape within which the Indian subcontinent place names are embedded. By clustering related words and identifying prevalent topics, this methodology helps to discern the major themes and subject matters that emerge from Pliny's narrative, providing a comprehensive understanding of the broader context in which these place names are mentioned.

Furthermore, network analysis offers a visual representation of the interconnections among the place names of the Indian subcontinent and other entities in Pliny's work. By examining the relationships between different locations and named entities, this analysis uncovers the geographical and conceptual networks that exist within the text, revealing how the Indian subcontinent place names contribute to the overall structure and narrative flow of *Natural History*.

Together, these methodologies aim to provide a nuanced and comprehensive exploration of the texts related to the Indian subcontinent in *Natural History*. By delving into the linguistic, thematic, and network aspects of these place names, a deeper understanding of their role in shaping Pliny's narrative can be achieved.

4.2 Word frequency and collocating bi-grams

By utilizing the measurements available in the NLTK package, a word frequency list and collocating bi-grams were generated from the text associated with Indian place names in *Natural History*. These outputs provide an overview of the prevalent words and word patterns, as potential keywords in the text.

In the initial observation, the words "India" and "one" ranked high in the frequency list. However, it is apparent that the passages would include the word "India" when discussing about India, making it less informative as a keyword. Likewise, the word "one" appeared as a generic descriptor for bringing up a type of tribe, plant, or attributes like distance, volume, or range, offering limited insight as a keyword. To enhance the relevance and descriptive nature of the frequency list, these two common but less informative words, "India" and "one", are further excluded from the token list.

Among 17729 tokens excluding "India" and "one", 201 (the top 1%) frequently occuring

words in the India-related text in *Natural History* are shown in Figure 5 and Figure 6.

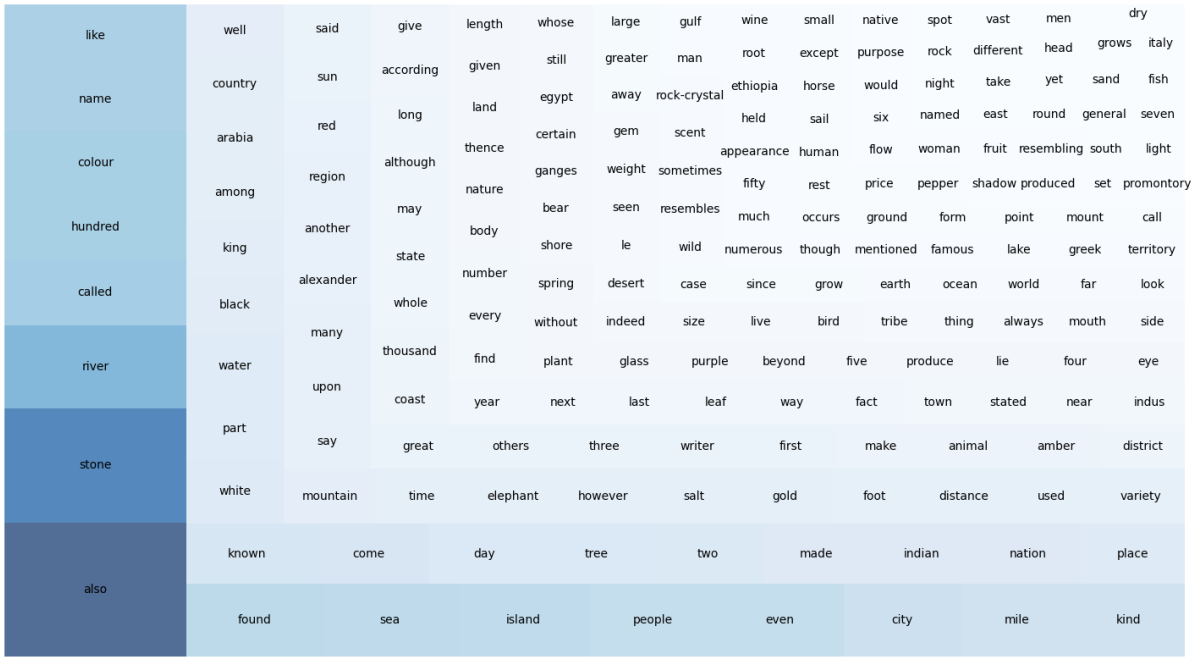

Figure 5: Top 1% frequent words in Indian subcontinent related text as tree map

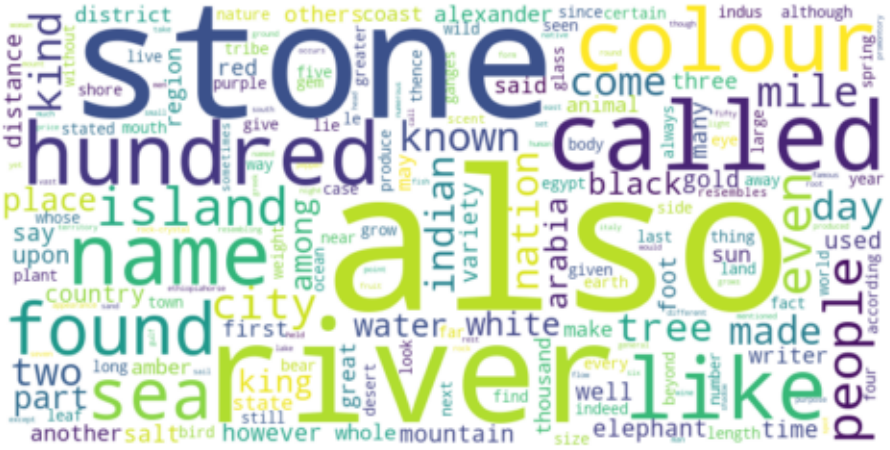

Figure 6: Top 1% frequent words in Indian subcontinent related text as word cloud

An intriguing observation from the word frequency sorting is the prominence of the word “also” in the given text. The word “also” appears frequently, which can be attributed to the encyclopedic nature of the work, where it is often used to draw comparisons in introductions about species and natural phenomena.

As shown in the following examples where “also” appears in the India-related text in *Natural History*, it is indeed used when comparing the counterparts in India after introducing a natural phenomenon, plant or human activity. In this regard, the common use of “also” may imply that India holds significance as a contrast in the broader narrative.

2.75.1 “**Also** in India at the well-known port of Patala the sun rises……”

12.10.1 “In India there is **also** a thorn the wood of which resembles ebony……”

12.15.1 “There is **also** in India a grain resembling that of pepper, but larger and more brittle……”

12.17.1 “Arabia **also** produces cane-sugar, but that grown in India is more esteemed.”

The hypothesis is further confirmed towards the end of the work in book 37, where Pliny concludes his comprehensive discourse on “Nature”. In 37.77.1, Pliny bestows the highest praise upon Italy, considering it to have earned “Nature’s crown”. And in this context, when expressing his preference and overall judgment, Pliny makes one final mention of “India”. He indicates that, “if we leave aside the fabulous marvels of India”[3], Spain can be appreciated as a significant and attractive destination, second only to Italy. This unintentional highlight of India suggests that it holds considerable importance as a distant contrast to the Mediterranean area, where the focal point locates in the *Natural History*’s world scope.

And following by “also”, the words “stone”, “river”, “called” and “colour” notably stand out in the word frequeny sorting. These frequent occurrences suggest some potential themes related to India in the content of *Natural History*, which aligns with the distribution of Indian place names as depicted in Figure 4.

Looking into the text in the three books pertaining the most mentions of Indian place names, book 6 includes specific topics on Nations of India, the Ganges and Indus (two main rivers in India) and routes of voyages to India, and book 12 contains introductions to trees and the economic values of their roots and leaves, as well as plants and their medical and flavouring effects. While book 37 focuses on descriptions of different types of gemstones, where the priciple types are introduced in a sequence/category of colours, alongside critics about the luxury trade they represents.

The frequent use of the word “river” in India-related text may be related to the mention of voyage and trading routes concerning India in *Natural History*. On the other hand, “stone” and “color” clearly connect to the content in book 37, which deals with gemstones.

These two potential themes observed from the frequent occuring words indicate that the geographical location and routes toward Indian subcontinent, and its role as an origin of many plants, animals and gemstones, pocesses a significance in the content about India in *Natural History*.

[3] *Natural History* 37.77.1 (https://topostext.org/work/153)

In addition to word frequency observation, collocation analysis is utilized to explore the common word patterns within the India-related text in *Natural History*.

The top 0.1% of the most likely collocating bi-grams are extracted using the likelihood ratio measurement. This selection process yields 18 out of 17728 bi-grams that are most significant and likely to co-occur together in the target corpus.

However, during the initial observation, it was noted that approximately one-third of the extracted bi-grams contained the word “hundred”, such as in (‘hundred’, ‘fifty’) and (‘six’, ‘hundred’). These bi-grams typically denoted measurements for distance, object length, or quantity, offering limited descriptive information about the content of the text. Consequently, the word “hundred” was excluded from further bi-gram extraction to focus on more informative and relevant co-occuring words.

And the updated output bi-grams are list as follows:

```
[('alexander', 'great'),
 ('father', 'liber'),
 ('caspian', 'gate'),
 ('fifty', 'mile'),
 ('denarii', 'pound'),
 ('gold', 'silver'),
 ('precious', 'stone'),
 ('river', 'indus'),
 ('fourteen', 'equinoctial'),
 ('olive', 'oil'),
 ('asia', 'minor'),
 ('equinoctial', 'hour'),
 ('red', 'sea'),
 ('lapis', 'lazuli'),
 ('mile', 'breadth'),
 ('emperor', 'nero'),
 ('already', 'mentioned'),
 ('ft.', 'long')]
```

The extracted bi-grams can be broadly categorized into four types:

Historical figures: (‘alexander’, ‘great’), (‘father’, ‘liber’), (‘emperor’, ‘nero’)

Geographical locations and features: (‘caspian’, ‘gate’), (‘river’, ‘indus’), (‘asia’, ‘minor’), (‘red’, ‘sea’)

Meseuarments (distance, currency, length, time): (‘fifty’, ‘mile’), (‘denarii’, ‘pound’), (‘fourteen’, ‘equinoctial’), (‘equinoctial’, ‘hour’), (‘mile’, ‘breadth’), (‘ft.’, ‘long’)

Trading goods: (‘gold’, ‘silver’), (‘precious’, ‘stone’), (‘olive’, ‘oil’), (‘lapis’, ‘lazuli’)

On the one hand, the presence of bi-grams associated with geographical locations, distance, and time measurements in the India-related text reaffirms India’s position as a geographic reference, consistent with the earlier findings from the word frequency list and literature review. On the other hand, within the context of the Indo-Mediterranean network of exchange, the occurrence of bi-grams related to geographical locations, currency measurements, and trading goods underscores the importance of India’s role in merchandise trade within the narratives of *Natural History*.

Furthermore, the occurrence of historical figures such as “Alexander III, the Great (king of Macedon)”, “Nero (Roman emperor)”, and “Father Liber (referring to Dionysus, the Greek god of winemaking and wine)” suggests their connections with India in the history of expeditions or mythical tales (Dionysus is believed to have conquered India in Greek epic). This observation opens up a perspective for clustering the human names mentioned in the text to reveal the content structure about India in *Natural History*, which will be further explored in the Network Analysis section.

In conclusion, the analysis of word frequency and collocation in the India-related text within *Natural History* reveals noteworthy word patterns. These patterns suggest that India holds a significant role as a geographical contrast, being compared in terms of distance, natural phenomena, and origin of products with other regions introduced in the narrative. Furthermore, it is highlighted for its prominent role in merchandise trade in the portrait of India in *Natural History*.

4.3 Topic modeling

To delve further into the underlying topics about India in the work, topic modeling approach is applied as a next step. Topic modeling is a widely used method for text analysis that infers the latent topics in a collection of documents(**bail?**; **ted2012?**). Latent Dirichlet Allocation (LDA), as its most commonly employed algorithm, operates under an assumption that each document contains a mixture of different topics, and each topic is defined as a collection of words with varying probabilities of appearance in the passages(**ted2012?**; **kapadia2022?**).

In this study, the collection of India-related text in *Natural History*, segmented into paragraphs, are considered as different “documents”. And the Genism library in Python is utilized for semantic vectorization and the implementation of the LDA model on the groups of words within these “documents”.

Since the corpus size for text pertaining to Indian place names is relatively small, and after several attempts, the number of topics is determined to be 3, with 40 passes to obtain the most optimal and non-overlapping topic clusters.

The list of 30 keywords, grouped by the 3 assigned topics, is presented below.

```
[(0,
  '0.014*"stone" + 0.011*"also" + 0.007*"colour" + 0.007*"india" + '
  '0.007*"found" + 0.006*"like" + 0.004*"one" + 0.004*"black" + 0.004*"tree" + '
  '0.004*"amber" + 0.004*"name" + 0.003*"known" + 0.003*"white" + 0.003*"kind" '
  '+ 0.003*"gold" + 0.003*"made" + 0.003*"even" + 0.003*"called" + '
  '0.003*"indian" + 0.003*"glass" + 0.002*"part" + 0.002*"people" + '
  '0.002*"used" + 0.002*"variety" + 0.002*"many" + 0.002*"river" + '
  '0.002*"make" + 0.002*"rock-crystal" + 0.002*"island" + 0.002*"red"'),
 (1,
  '0.009*"stone" + 0.008*"also" + 0.007*"india" + 0.006*"like" + 0.006*"kind" '
  '+ 0.006*"colour" + 0.005*"name" + 0.005*"one" + 0.004*"called" + '
  '0.004*"even" + 0.004*"white" + 0.003*"pepper" + 0.003*"variety" + '
  '0.003*"another" + 0.003*"known" + 0.003*"tree" + 0.003*"black" + '
  '0.003*"weight" + 0.003*"leaf" + 0.003*"purple" + 0.002*"grain" + '
  '0.002*"people" + 0.002*"denarii" + 0.002*"used" + 0.002*"nard" + '
  '0.002*"resembles" + 0.002*"pound" + 0.002*"taste" + 0.002*"found" + '
  '0.002*"small"'),
 (2,
  '0.011*"river" + 0.010*"hundred" + 0.008*"one" + 0.007*"india" + '
  '0.007*"also" + 0.007*"city" + 0.007*"mile" + 0.007*"sea" + 0.006*"island" + '
  '0.005*"nation" + 0.005*"called" + 0.005*"people" + 0.004*"day" + '
  '0.004*"come" + 0.004*"two" + 0.004*"distance" + 0.004*"name" + '
  '0.004*"place" + 0.004*"salt" + 0.004*"king" + 0.003*"elephant" + '
  '0.003*"water" + 0.003*"country" + 0.003*"foot" + 0.003*"alexander" + '
  '0.003*"thousand" + 0.003*"upon" + 0.003*"mountain" + 0.003*"even" + '
  '0.003*"part"')]
```

Based on the prominent category of the keywords and the possible interconnection inbetween in each group, a summary for interpretation of the topic emerged from each group of keywords can be drawn as follows.

Group 0: This group includes various materials relating to precious stones, such as “stone”, “amber”, “rock-crystal”and “glass”, with mentions of colours like “black”, “white”, and “red”. And the underlying topic could be summarized as description of precious stones.

Group 1: This group includes different kinds of natural products, such as “tree”, “leaf”, “pepper”,“grain” and “nard”. It also contains words of description to the prducts, such as “weight”, “pound” and “taste”. In addition, a currency of ancient Rome, “denarii”, appears in the group. Therefore the possible topic for this group is the merchandise trade with Indian subcontinent.

Group 3: This group contains different geographical features such as “river”, “island”, “sea” and “mountain”. It also includes terms related to cities, nations, and distances, with mentions of historical figures like Alexander, who has once conquered India. “elephant” also represents the power and size of the kings in India during that era. In this regard, the underlying topic for this group could be geography and political situation in India.

The top 30 keywords for each topic, along with their respective weights, which rank their contributions to the topic is shown and visualized as follows.

```
<IPython.core.display.HTML object>
```

The interactive visualisation of the 3 clusters of the topic modeling about India-related text can be accessed on the html version of this thesis.

The static demonstration of the visuslisation can be referred as Figure 7, Figure 8 and Figure 9.

In the left panel of the above interactive chart, each bubble represents a topic, and the size of the bulbble indicates the percentage of the texts in the corpus contributing to the topic. The distance between the bubbles implies the extent of difference between them. And a good topic model is expected to have big and non-overlapping bubbles scattered throughout the chart (Tran 2022).

And in the right panel, the blue bars represent the overall frequency of each word in the corpus. If no topic is selected, the blue bars of the most frequently used words will be displayed. When hovering on the bubbles in the left panel, there will be red bars in the right panel giving the estimated number of times a given term was generated by a given topic. The word with the longest red bar is estimated to be used the most in the texts belonging to that topic.

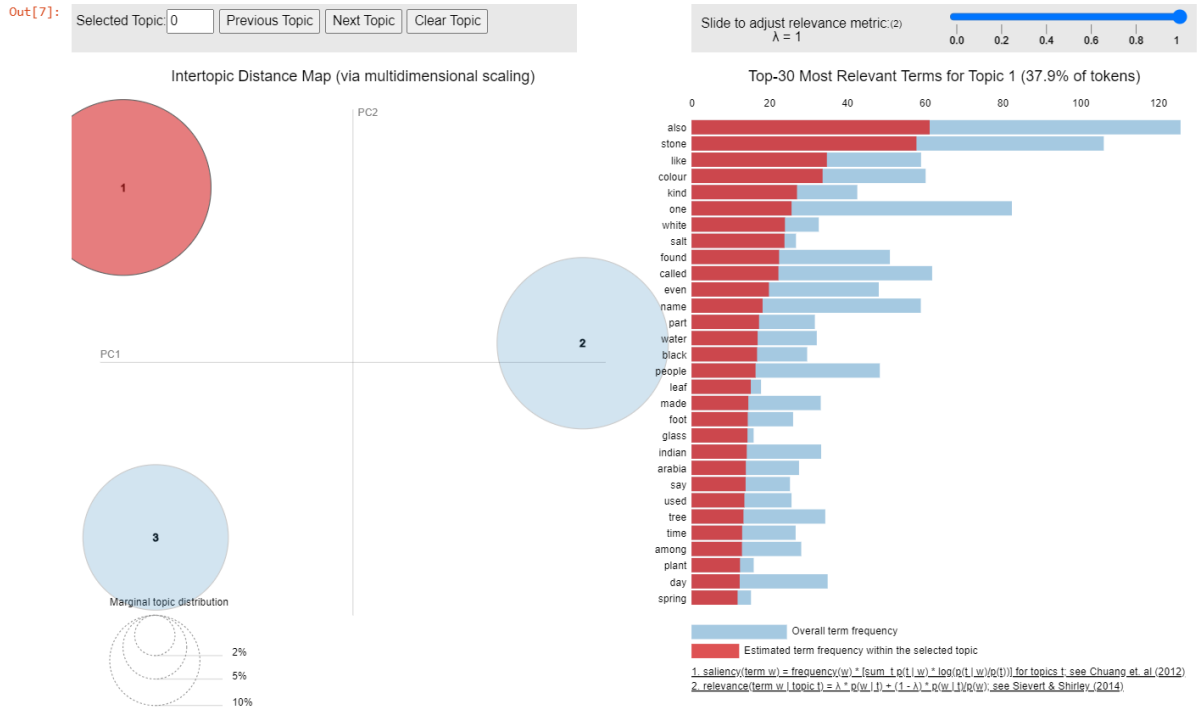

Figure 7: Topic cluster 1

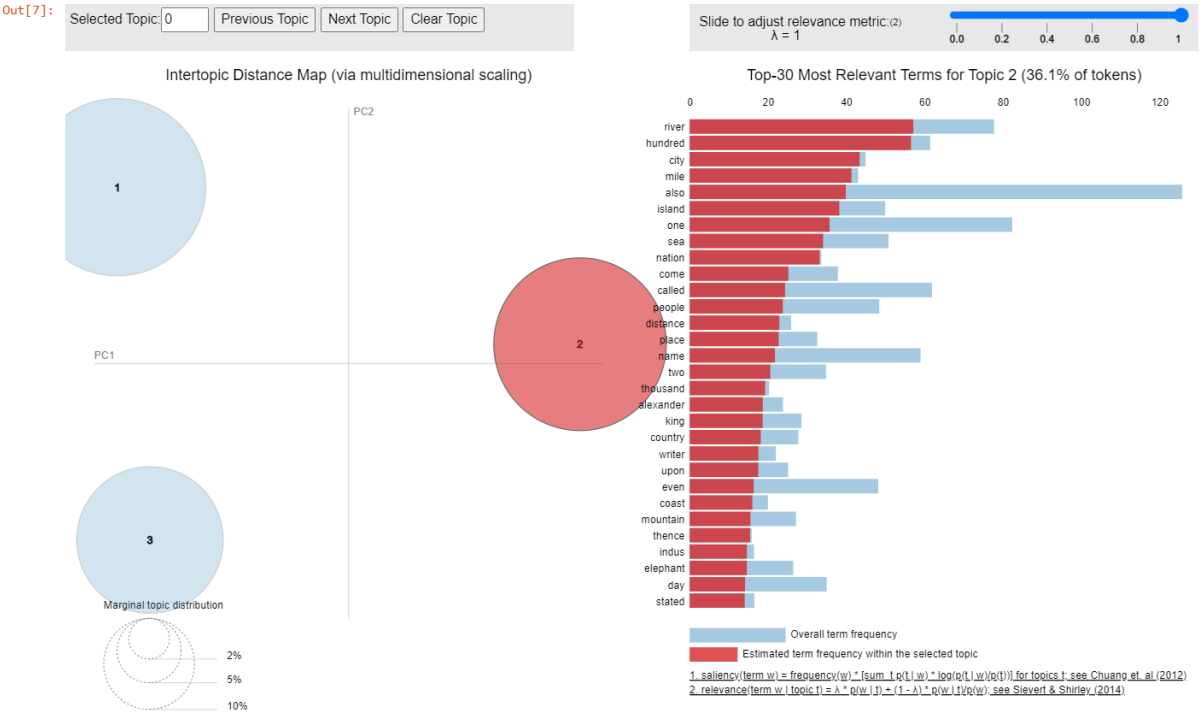

Figure 8: Topic cluster 2

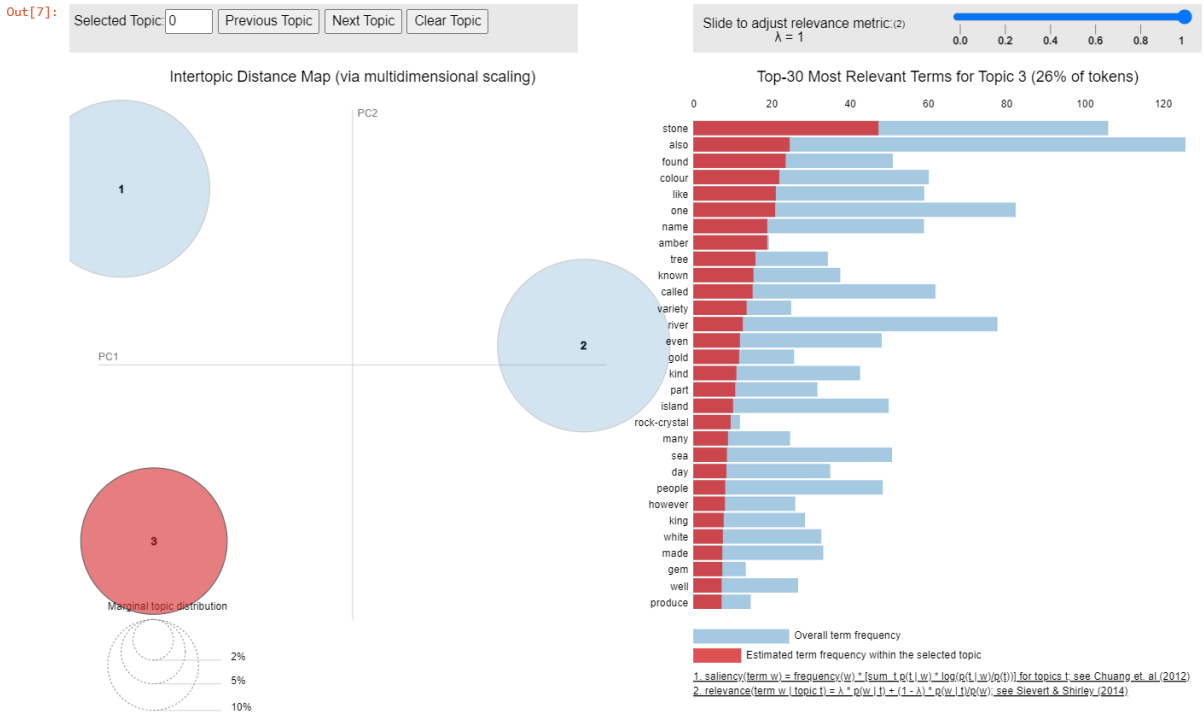

Figure 9: Topic cluster 3

4.4 Network analysis for Named Entity

5 Conclusions

6 Old structure

6.1 Overview of geographical related texts

What topics popped up from the context of place names?

6.1.1 Distribution of place names in the entire book

The normalized frequency of place name references in *Natural History* was calculated as the ratio of counts of the occurrences of place names in each book to the word lengths of the book (Table 7). As depicted in Figure 10, the findings indicate that books 3-6 prominently feature a higher frequency of place name references. This observation is consistent with content structure of *Natural History*, that books 3-6 centered around the themes of **"Geography and ethnography"**, is expected to contain a great number of location references.

Table 7: Distribution of place names in Natural History

Book	Total_length	Place_count	Place_freq
1	2778	1	0.000360
2	30570	406	0.013281
3	18037	1007	0.055830
4	15434	1309	0.084813
5	18872	1112	0.058923
6	27890	1012	0.036285
7	21204	225	0.010611
8	24176	185	0.007652
9	19197	140	0.007293
10	20816	121	0.005813
11	27345	77	0.002816
12	13906	188	0.013519
13	13243	164	0.012384
14	15277	189	0.012372
15	14552	135	0.009277
16	25442	180	0.007075
17	29387	82	0.002790
18	35850	222	0.006192
19	18822	146	0.007757
20	22743	21	0.000923
21	17896	95	0.005308
22	16491	24	0.001455
23	15764	17	0.001078
24	17491	56	0.003202
25	16734	85	0.005079
26	15448	35	0.002266
27	12444	40	0.003214
28	26476	28	0.001058
29	13976	31	0.002218
30	14395	23	0.001598
31	12204	222	0.018191
32	14635	76	0.005193
33	17946	113	0.006297
34	18972	193	0.010173
35	21282	277	0.013016
36	21295	357	0.016764
37	22255	282	0.012671

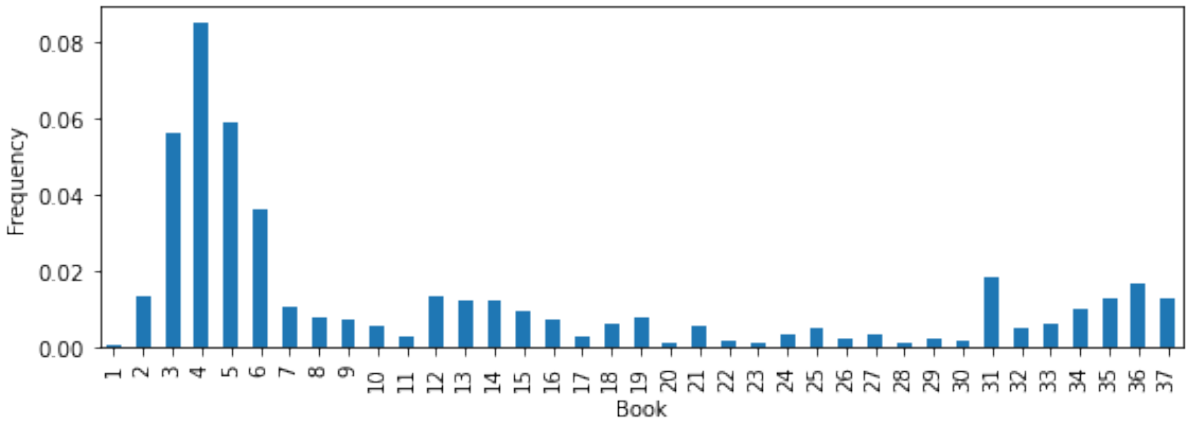

Figure 10: Place name distribution in Natural History

6.1.2 Topic modelling on geographical location related text

Genism library is used for semantic vectorization and implemetion of Latent Dirichlet Allocation (LDA) model for the topic modelling in the captioned text.

And the library of pyLDAvis is applied for an interactive visualization.

```
[(0,
  '0.010*"also" + 0.004*"picture" + 0.003*"painted" + 0.003*"milk" + '
  '0.003*"sponge" + 0.003*"first" + 0.003*"dung" + 0.003*"bird" + 0.002*"egg" '
  '+ 0.002*"made" + 0.002*"time" + 0.002*"horse" + 0.002*"year" + '
  '0.002*"caesar" + 0.002*"painting" + 0.002*"one" + 0.002*"goat" + '
  '0.002*"said" + 0.002*"two" + 0.002*"called" + 0.002*"give" + 0.002*"boy" + '
  '0.002*"onion" + 0.002*"among" + 0.002*"day" + 0.002*"great" + 0.002*"sheep" '
  '+ 0.002*"make" + 0.002*"famous" + 0.001*"wine"'),
 (1,
  '0.018*"also" + 0.011*"kind" + 0.007*"called" + 0.007*"stone" + 0.007*"like" '
  '+ 0.006*"wine" + 0.006*"colour" + 0.006*"leaf" + 0.006*"one" + '
  '0.005*"plant" + 0.005*"tree" + 0.005*"used" + 0.005*"water" + 0.005*"found" '
  '+ 0.005*"white" + 0.005*"root" + 0.004*"taken" + 0.004*"made" + '
  '0.004*"variety" + 0.004*"oil" + 0.004*"name" + 0.004*"seed" + 0.004*"black" '
  '+ 0.003*"make" + 0.003*"grows" + 0.003*"even" + 0.003*"honey" + '
  '0.003*"juice" + 0.003*"said" + 0.003*"two"'),
 (2,
  '0.007*"called" + 0.007*"also" + 0.007*"island" + 0.006*"day" + 0.006*"sea" '
  '+ 0.006*"one" + 0.005*"name" + 0.005*"place" + 0.005*"river" + 0.004*"wind" '
  '+ 0.004*"mile" + 0.004*"city" + 0.004*"soil" + 0.004*"time" + 0.003*"year" '
  '+ 0.003*"district" + 0.003*"earth" + 0.003*"land" + 0.003*"people" + '
  '0.003*"sun" + 0.003*"country" + 0.003*"part" + 0.003*"great" + 0.003*"two" '
  '+ 0.003*"italy" + 0.003*"nation" + 0.003*"region" + 0.002*"spring" + '
```

```
 '0.002*"first" + 0.002*"distance"'),
(3,
 '0.016*"river" + 0.016*"mile" + 0.013*"town" + 0.011*"called" + 0.009*"sea" '
 '+ 0.009*"name" + 0.007*"distance" + 0.006*"water" + 0.006*"also" + '
 '0.006*"city" + 0.005*"island" + 0.005*"come" + 0.005*"two" + '
 '0.005*"hundred" + 0.004*"gulf" + 0.004*"upon" + 0.004*"place" + '
 '0.004*"formerly" + 0.004*"promontory" + 0.004*"people" + 0.004*"coast" + '
 '0.004*"one" + 0.003*"part" + 0.003*"nation" + 0.003*"mountain" + '
 '0.003*"side" + 0.003*"lie" + 0.003*"length" + 0.003*"distant" + '
 '0.003*"mouth"'),
(4,
 '0.010*"also" + 0.008*"even" + 0.007*"one" + 0.005*"made" + 0.004*"first" + '
 '0.004*"year" + 0.004*"time" + 0.003*"rome" + 0.003*"man" + 0.003*"king" + '
 '0.003*"people" + 0.003*"work" + 0.003*"statue" + 0.003*"day" + 0.003*"tree" '
 '+ 0.003*"place" + 0.003*"used" + 0.003*"name" + 0.003*"great" + '
 '0.003*"gold" + 0.003*"temple" + 0.002*"among" + 0.002*"men" + 0.002*"case" '
 '+ 0.002*"animal" + 0.002*"many" + 0.002*"two" + 0.002*"called" + '
 '0.002*"although" + 0.002*"life"')]
```

The text data undergoes tokenization and lemmatization using functions from the NLTK package. This preprocessing step aims to obtain meaningful words that facilitate the inference of potential topics based on grouped keywords. To ensure the modeling results consist of words with descriptive meaning, stop words in English are excluded, along with tokens having a length less than 2, when preparing the corpus for input into the LDA module.

After several tryouts, the number of topics is set to 5, and the passes is set to 20, in order to generate distinct and non-overlapping topic clusters.

The following visualization presents the top 30 keywords for each topic, along with their respective weights, which rank their contributions to the topic.

```
<IPython.core.display.HTML object>
```

In the left panel of the above interactive chart, each bubble represents a topic, and the size of the bulbble indicates the percentage of the texts in the corpus contributing to the topic. The distance between the bubbles implies the extent of difference between them. And a good topic model is expected to have big and non-overlapping bubbles scattered throughout the chart (Tran 2022).

And in the right panel, the blue bars represent the overall frequency of each word in the corpus. If no topic is selected, the blue bars of the most frequently used words will be displayed. When hovering on the bubbles in the left panel, there will be red bars in

the right panel giving the estimated number of times a given term was generated by a given topic. The word with the longest red bar is estimated to be used the most in the texts belonging to that topic.

An intriguing observation about the overall result of the topic modelling is that the word “also” comprises a large protion in the given text, and appears in all assigned topics. Taking the encyclopedia scope of Natural History into consideration, it may imply that the place names are prone to be mentioned in a context of enumerason and comparison. In the literary studies by Pollard (2009) and Murphy (2003), Pliny gave a critical description of the geographical surroundings and their exotic counterparts (e.g., Po River and Nile River), which may confirms it worthwhile getting a deeper exploration in the usage and reference of the place names in Natual History in order to map the scope and vision he attempted to display in the encyclopedia by Pliny the Elder.

More specifically, a rough generalization can be drown for each topic with the dominant words in it as follows, which may help to conclude the themes and keywords for geography related context in *Natural History*.

Topic 1: **Artistic Elements and Objects** - The presence of paintings, milk, sponges, and other objects adds to the artistic and visual aspects of the context.

Topic 2: **Botanical and Natural Elements** - Various plants, trees, colors, and natural materials contribute to the botanical richness depicted in the book.

Topic 3: **Geographic Features and Places** - Islands, rivers, cities, and other geographical features play a significant role in the narrative, highlighting the diverse landscapes explored in the text.

Topic 4: **Distance and Proximity** - Distances, towns, rivers, and seas provide insights into the spatial relationships and navigational aspects within the book.

Topic 5: **Historical and Cultural References** - Roman history, statues, temples, and notable figures showcase the historical and cultural context prevalent in the book.

In addition, as shown in the visualization chart, the Topic 5: **Historical and Cultural References** and Topic 2: **Botanical and Natural Elements** seem to be the most prominent topics about geographical location related text in *Natural History*.

In conclusion, the general exploratory analysis about geographical location related text in *Natural History* shows that in the books about geography and ethnography, and mining and mineralogy, place names are most frequently referred. And the potential topics about geograpical location related contents are “Artistic Elements and Objects”, “Geographic Features and Places”, “Distance and Proximity”, “Historical and Cultural References” and “Botanical and Natural Elements”, with the latter two as the most

prominent topics in the context.

Considering the comprehensive scope of *Natural History*, the presence of concrete place names provides a valuable opportunity to delve deeper into Pliny the Elder's perception and imagination of landscapes. Therefore, it is worthwhile to embark on a more detailed examination of the distribution, significance, and contextualization of place names in *Natural History* to gain insights into how Pliny the Elder crafted the narrative and conveyed his understanding of the world.

6.2 Prominent location mentioned in Natural History

What place stands out in the narrative? And how does it align with the scope and underlying concept of *Natural History*?

6.2.1 Place name distribution

By grouping the "ToposText_ID" (as indicator for distinct geographical loactions in the text) in the earlier constructed dataframe, there are 2052 unique places mentioned in *Natural History*.

The top 20 most frequent place names mentioned (as 1% of total) in *Natural History* is shown in Table 8.

Table 8: Top 20 mentioned place names in Natural History

	ToposText_ID	Place_Name	Lat	Long	Count
1687	https://topostext.org/place/406163RIta	Italy	40.6	16.3	292
2034	https://topostext.org/place/419125PRom	Rome	41.891	12.486	269
52	https://topostext.org/place/271307REgy	Egypt	27.1	30.7	261
82	https://topostext.org/place/300740RInd	India	30	74	167
57	https://topostext.org/place/280400RAra	Arabia	28	40	123
320	https://topostext.org/place/355390RSyr	Syria	35.5	39	109
255	https://topostext.org/place/350330RCyp	Cyprus	35	33	85
109	https://topostext.org/place/312301WNil	Nile	30.0918	31.2313	85
2282	https://topostext.org/place/441073LAlp	Alps	44.142	7.343	82
766	https://topostext.org/place/376145RSic	Sicily	37.6	14.5	71
275	https://topostext.org/place/352252IKre	Crete	35.2052	25.1836	64
7	https://topostext.org/place/130350REth	Ethiopia	13.01	35.01	58
417	https://topostext.org/place/364282IRho	Rhodes	36.4408	28.2244	56
966	https://topostext.org/place/380237PAth	Athens	37.9718	23.72793	56
2043	https://topostext.org/place/419125SCap	Capitol	41.8933	12.483	52
298	https://topostext.org/place/353403WEup	Euphrates	35.2791	40.2708	47
2241	https://topostext.org/place/435335WPon	Pontus	43.5	33.5	47

	ToposText_ID	Place_Name	Lat	Long	Count
1839	https://topostext.org/place/411146RCam	Campania	41.1	14.6	46
1480	https://topostext.org/place/397443RArm	Armenia	39.702	44.298	45
17	https://topostext.org/place/195390WEry	Red Sea	19.5	39	42
545	https://topostext.org/place/369103PCar	Carthage	36.85	10.32	42
602	https://topostext.org/place/370340RCil	Cilicia	37.01	34.01	42

The place names referenced in Natural History are geographically mapped, with each location marked on the map using its corresponding coordinates. A dot is assigned to represent each place, with the size and color of the dot reflecting the frequency of its mention in the book. The larger and darker the dot, the more frequently the place is referenced within the context of *Natural History*.

An intriguing observation from the output, as depicted in Figure 11, is the prominence of India—a region outside the Mediterranean—despite its high frequency of mentions.

```
<folium.folium.Map at 0x17c69eb4490>
```

Figure 11: Place name distribution map

6.2.2 Zooming into "India"

As highlighted in the research conducted by Nappo (2017), the era of Pliny the Elder's writing of *Natural History* witnessed a thriving Indo-Roman trade relationship. The prominence of the term "India" within the text suggests that this trade connection holds considerable significance in the narrative of *Natural History*.

To provide more comprehensive contextual analysis, the focus is extended beyond solely "India" to the regions that encompass the empires of the Indian subcontinent. The approximate range of coordinates defining the target region is as follows:[4]

Latitude: Northernmost point: Approximately 37.6 degrees North (located in the region of Jammu and Kashmir in India) Southernmost point: Approximately 5.5 degrees North (located in the region of Dondra Head in Sri Lanka)

Longitude: Westernmost point: Approximately 60.9 degrees East (located in the region of Gwadar in Pakistan) Easternmost point: Approximately 97.4 degrees East (located in the region of Kibithu in India)

And a dataframe for Indian subcontinent related texts can be filetered with the captioned coordinates range.

[4]Given the challenges in determining the precise coordinates of the Empires in the Indian region during the 1st century AD, an approximate range of coordinates for the current Indian subcontinent is used as a rough estimation.

	UUID4	ToposText_ID	Place_Nar
85	1c5714c4-00f4-4ed3-81db-691ae52ac72d	https://topostext.org/place/300740RInd	India
92	a3f8e7ea-fd33-40b5-8095-a1feb0c19919	https://topostext.org/place/300740RInd	India
93	70343d33-a104-4566-ab63-b3e70527a21c	https://topostext.org/place/300740RInd	India
218	da1d2386-a176-4995-9a2a-37cdf7e07ec8	https://topostext.org/place/254683WInd	Indus
326	305c95c4-ae38-4751-a686-3b72a6bed815	https://topostext.org/place/340670RBac	Bactria

The shape of the filtered dateframe for texts and place coordinates related to Indian subcontinent is (241, 10). And the dataframe is also saved as .csv for further reference.

And the places refered in the captioned region in the dataframe are: [‘India’ ‘Indus’ ‘Ganges’ ‘Acesinus’ ‘Hydaspes’ ‘Taprobane’ ‘Arachosia’ ‘Muziris’ ‘Baragaza’ ‘Ceylon’].

The comparison between the total number of place names and the place names specifically related to the Indian subcontinent mentioned in each book, is depicted in Figure 12. The difference in numbers between the two categories is significant, as indicated by the large disparity.

To facilitate a more effective comparison of the referencing trends across different books, Figure 13 presents subplots with varying y-axis scales. This approach allows for a clearer visualization of the trends and patterns in place name references throughout the various books.

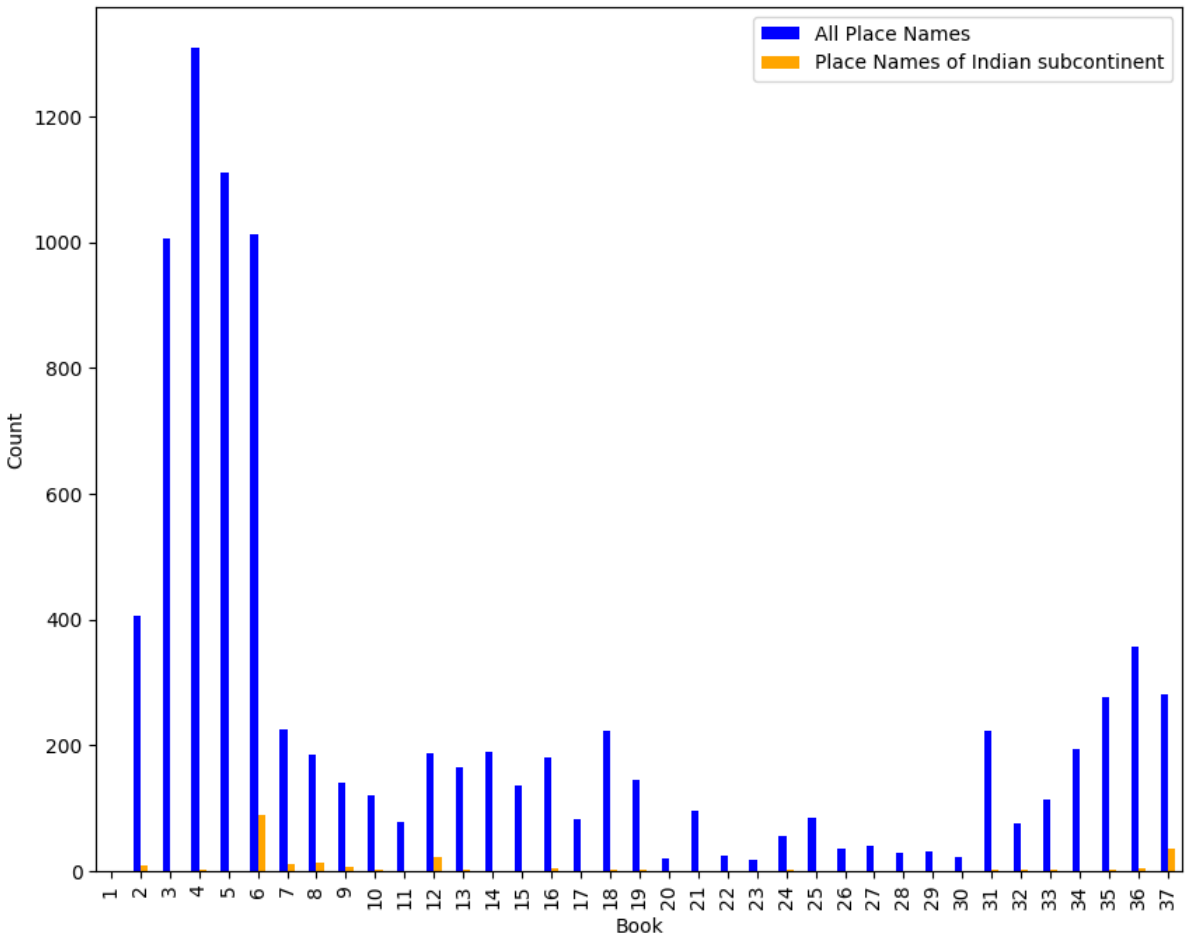

Figure 12: Occurence count for all place names and place names of Indian subcontinent in each book

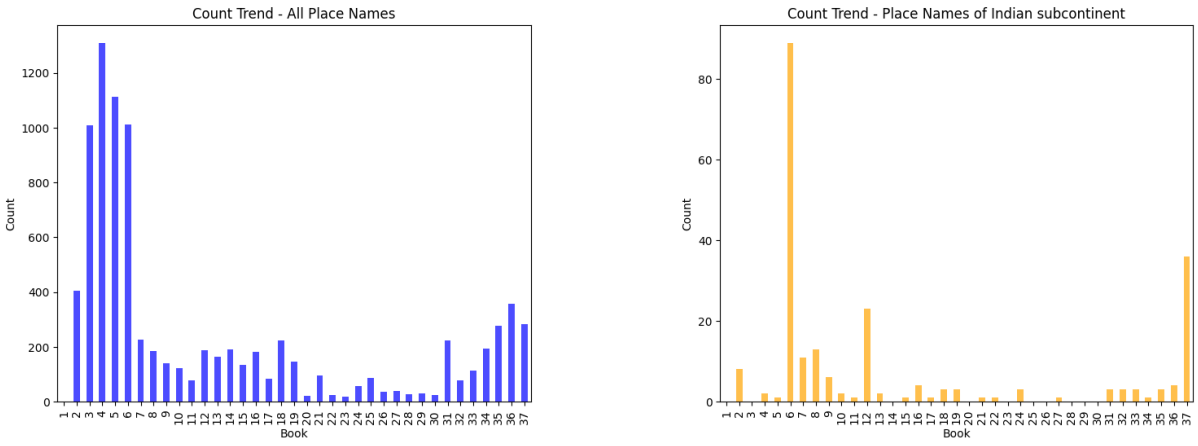

Figure 13: Occurence count for all place names and place names of Indian subcontinent in each book_different y-axis scales

The figures reveal a distinct difference between the occurrence trends of place names related to the Indian subcontinent and all place names collectively. Specifically, the

referencing of the Indian subcontinent is highly concentrated in books 6, 12, and 37 of Pliny's narrative. This discrepancy indicates that the mentioning of place names from the Indian subcontinent is closely tied to specific themes and topics within Pliny's work.

In this regard, three methodologies have been employed to analyze the texts pertaining to the Indian subcontinent in *Natural History*, including collocation analysis, topic modeling, and network analysis. The objective of these analyses is to delve deeper into the textual content, unraveling the intricate relationships and uncovering the underlying themes and connections associated with the place names of the Indian subcontinent.

Through collocation analysis, the aim is to identify significant word combinations and phrases that co-occur with the place names of the Indian subcontinent. This analysis provides insights into the specific linguistic patterns and contextual associations surrounding these locations, shedding light on their cultural, historical, and geographical significance.

Topic modeling allows for a broader exploration of the thematic landscape within which the Indian subcontinent place names are embedded. By clustering related words and identifying prevalent topics, this methodology helps to discern the major themes and subject matters that emerge from Pliny's narrative, providing a comprehensive understanding of the broader context in which these place names are referenced.

Furthermore, network analysis offers a visual representation of the interconnections among the place names of the Indian subcontinent and other entities in Pliny's work. By examining the relationships between different locations and named entities, this analysis uncovers the geographical and conceptual networks that exist within the text, revealing how the Indian subcontinent place names contribute to the overall structure and narrative flow of *Natural History*.

Together, these methodologies aim to provide a nuanced and comprehensive exploration of the texts related to the Indian subcontinent in *Natural History*. By delving into the linguistic, thematic, and network aspects of these place names, a deeper understanding of their significance and their role in shaping Pliny's narrative can be achieved.

6.2.2.1 Frequency list and collocations in Indian subcontinent related texts

Through the utilization of measures available in the NLTK package, a word frequency list and a list of collocating bi-grams of the texts pertaining to the Indian subcontinent are generated to investigate potential keywords and themes of interest.

To enhance the relevance and descriptive nature of the frequency list, particular attention has been given to exclude two commonly encountered but less informative words, namely "india" and "also", from the token list.

Among 18775 tokens of the whole corpus for Indian subcontinent related text, 197 (the top 1%) frequent words is filterd out and shown in Figure 5 and Figure 6.

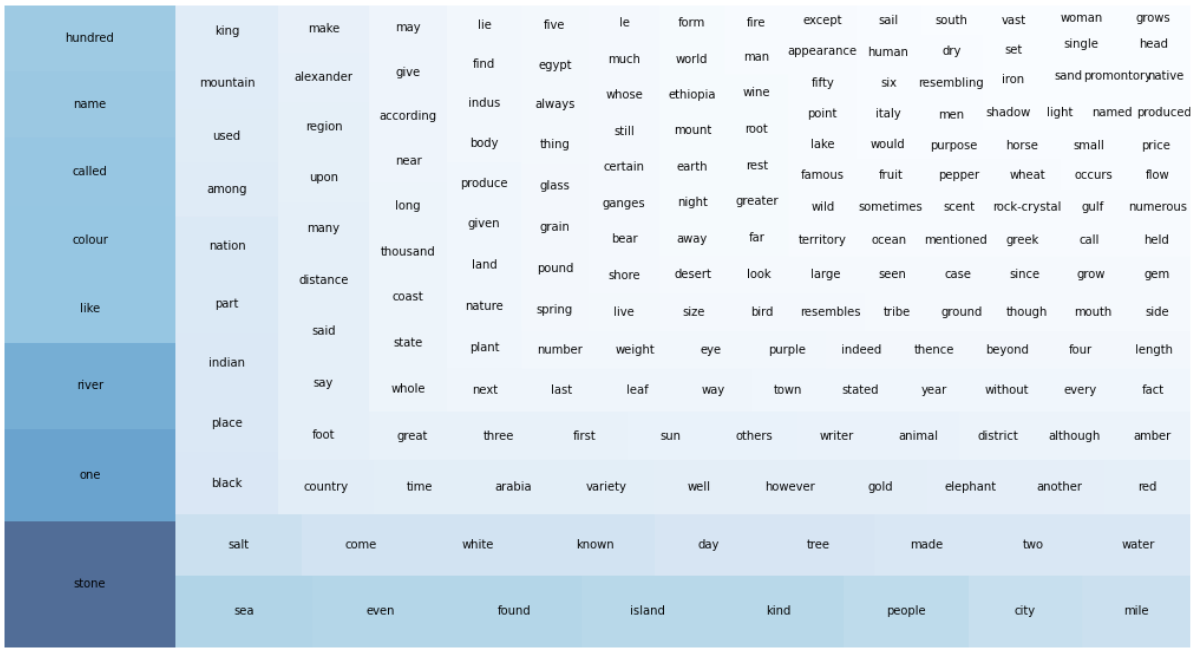

Figure 14: Top 1% frequent words in Indian subcontinent related text as tree map

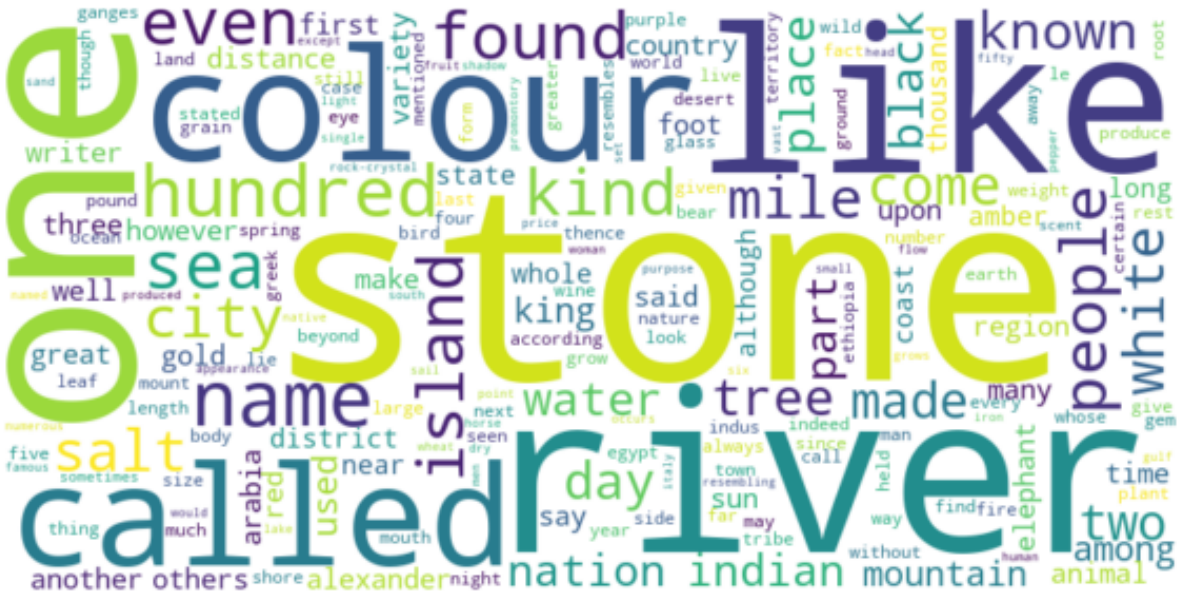

Figure 15: Top 1% frequent words in Indian subcontinent related text as word cloud

As depicted in the visualizations, the words “stone,” “river,” and “color” notably stand out, suggesting their prominence in the narrative pertaining to the regions of the Indian subcontinent. This observation is indicative of the significant references to precious stones and the origins and transportation routes associated with the trade of such valuable commodities.

The collocating bi-grams associated with place names of the Indian subcontinent region are extracted based on the top 20 highest scores in the likelihood ration measurement. A higher likelihood ratio score indicates a stronger association or collocation between the words, suggesting that they are more likely to appear together in the given text.

The extracted collocations undergo a filtering process that specifically includes those involving keywords of place names within the regions of the Indian subcontinent, whihc enables a focused analysis of collocations directly relevant to the geographic context.

```
[('already', 'mentioned'),
 ('present', 'day'),
 ('alexander', 'great'),
 ('father', 'liber'),
 ('taken', 'drink'),
 ('formerly', 'called'),
 ('majesty', 'augustus'),
 ('fifty', 'mile'),
 ('late', 'majesty'),
 ('next', 'come'),
 ('roman', 'citizen'),
 ('mile', 'circumference'),
 ('human', 'being'),
 ('greek', 'name'),
 ('late', 'lamented'),
 ('marcus', 'varro'),
 ('one', 'hundred'),
 ('hundred', 'fifty'),
 ('rising', 'dog-star'),
 ('emperor', 'nero')]
```

Interestingly, in the flitered bi-grams, 20% of them are referring to human names or names of gods in myths (e.g. Alexander III, the Great (king of Macedon); Octavius Caesar Augustus (Roman Emperor); Nero (Roman emperor); Marcus Varro (ancient Latin scholar), Father Liber (referring to Dionysus, Greek god of winemaking and wine)).

As shown in the quotation of Book 16, Chapter 62, Paragraph 1, the word “India” was mentioned in the context of an introduction of a plant, as a counterpart in the plant origin, and as a conquered land intertwining with the historical story about how the plant was brought to Rome by Alexander the Great.

> 16.62.1 It is said that ivy now grows in Asia Minor. Theophrastus about 314 BC. had stated that it did not grow there, nor yet in **India** except on Mount Meros, and indeed that Harpalus had used every effort to grow it

> in Media without success, while **Alexander** had come back victorious from **India** with his army wearing wreaths of ivy, because of its rarity, in imitation of **Father Liber**; and it is even now used at solemn festivals among the peoples of Thrace to decorate the wands of that god, and also the worshippers' helmets and shields, although it is injurious to all trees and plants and destructive to tombs and walls, and very agreeable to chilly snakes, so that it is surprising that any honour has been paid to it.

##(More detailed analysis and illustration will be further conducted for the pattern of interactions between Indian subcontinent place names and human names in the book.)

6.2.3 Topic modelling about Indian subcontinent region related texts

Since the corpus size for text pertaining Indian subcontient region is rather small, with certain tryouts, the the number of topics is set as 3 and the passes is set as 40 to get the most noon-overlapping topic clusters.

The word "India" is excluded from the corpus in order to get more descriptive keywords which may contribe to a more concrete topic summary.

The top 30 keywords for each topic, along with their respective weights, which rank their contributions to the topic is shown and visualized as follows.

```
[(0,
  '0.025*"stone" + 0.007*"also" + 0.007*"river" + 0.007*"found" + '
  '0.007*"colour" + 0.006*"like" + 0.005*"one" + 0.005*"name" + 0.005*"island" '
  '+ 0.005*"white" + 0.004*"hundred" + 0.004*"mile" + 0.004*"gold" + '
  '0.004*"variety" + 0.003*"come" + 0.003*"glass" + 0.003*"city" + '
  '0.003*"known" + 0.003*"many" + 0.003*"sea" + 0.003*"gem" + 0.003*"black" + '
  '0.003*"even" + 0.003*"nation" + 0.003*"thence" + 0.003*"place" + '
  '0.002*"according" + 0.002*"distance" + 0.002*"kind" + 0.002*"alexander"'),
 (1,
  '0.010*"also" + 0.006*"called" + 0.006*"one" + 0.005*"hundred" + '
  '0.005*"people" + 0.004*"name" + 0.004*"tree" + 0.004*"kind" + 0.004*"river" '
  '+ 0.004*"colour" + 0.004*"like" + 0.004*"city" + 0.003*"even" + '
  '0.003*"mile" + 0.003*"two" + 0.003*"black" + 0.003*"known" + 0.003*"island" '
  '+ 0.003*"used" + 0.003*"part" + 0.003*"amber" + 0.003*"indian" + '
  '0.003*"made" + 0.003*"foot" + 0.003*"come" + 0.003*"mountain" + 0.003*"sea" '
  '+ 0.002*"make" + 0.002*"arabia" + 0.002*"king"'),
 (2,
  '0.010*"salt" + 0.007*"also" + 0.006*"sea" + 0.006*"one" + 0.005*"even" + '
  '0.004*"day" + 0.004*"river" + 0.004*"water" + 0.004*"time" + 0.003*"among" '
```

```
  '+ 0.003*"great" + 0.003*"spring" + 0.003*"kind" + 0.003*"elephant" + '
  '0.003*"found" + 0.003*"island" + 0.002*"animal" + 0.002*"alexander" + '
  '0.002*"called" + 0.002*"made" + 0.002*"near" + 0.002*"night" + '
  '0.002*"country" + 0.002*"king" + 0.002*"people" + 0.002*"well" + '
  '0.002*"land" + 0.002*"two" + 0.002*"name" + 0.002*"place"')]
```

```
<IPython.core.display.HTML object>
```

The three generated topics for the Indian subcontinent related texts can be summarized based on the dominant words as follows:

Topic 1: **Stones, Rivers, and Islands** - various elements related to stones, rivers, and islands. It also touches upon the notion of distance and the mention of gold and gems.

Topic 2: **Cities, Trees, and Natural Features** - cities, trees, and natural features. It also mentions amber, mountains, and the connection to Arabia.

Topic 3: **Salt, Sea, and Water** - salt, the sea, and water-related concepts. It also touches upon topics such as animals, Alexander the Great, and the notion of a country.

And Topic 1: **Stones, Rivers, and Islands** takes the forefront among the other topics.

Consistent with the findings in the frequency list of the corpus, it is evident that “stones” and “rivers” hold a significant presence in the narrative concerning the Indian subcontinent.

6.2.3.1 Network analysis about Indian subcontinent region related texts

Two separate network analyses were conducted. The first analysis focused on exploring the relationships between place names mentioned throughout the entire book. The second analysis specifically examined the name entities of people and place names associated with the Indian subcontinent regions. Nodes and edges were generated for both analyses and imported into Gephi for visualization. By studying the clustering patterns of place names and people in the resulting network graphs, valuable insights can be gained into both the overall context of the book and the specific context of the Indian subcontinent within *Natural History*.

In the network analysis for place names throughout the entire book, unique place names are seen as nodes, and once two place names co-occur in the same paragraph, it will be counted as one edge. There are total 2255 nodes and 52602 edges in the prepared data.

As shown in Figure 16, the size of the node represents the betweeness centrality a place name mentioned in the book, and the weight of edge between two nodes rep-

resents the time the two place names appeared in the same paragraph (as seen in the same context). Gone through a Force Atlas 2 layout algorithm, the graph also demonstrates the rough cluster of place names which tend to be mentioned together.

In the case of “India”, it is observed mostly incooperates with “Egypt”, “Arabia” and “Nille”, which tend to be appearing in the description of trading route.

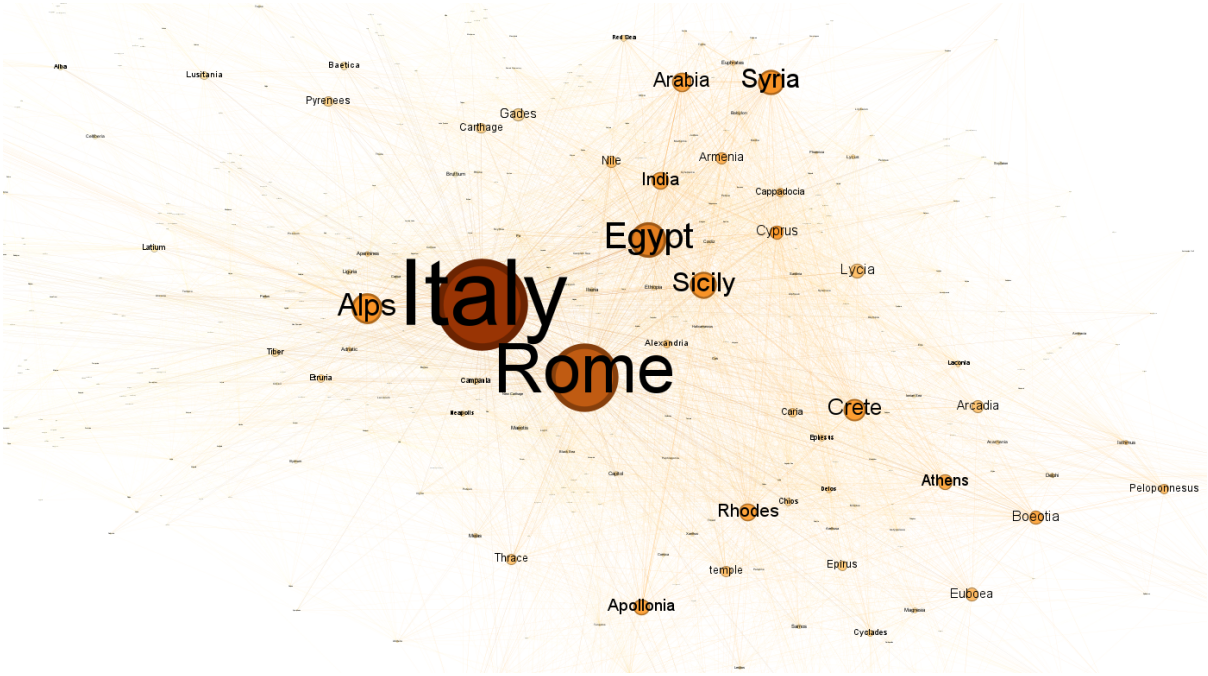

Figure 16: Network graph for place namesmentioned in Natural History

To gain a more detailed cluster of narrative contents about Indian subcontinent in *Natural History*, the idea is to generate a network for book number, place names and person names in the target corpus. The person name nodes are retrieved from the tagging of text given by the pretrained multilingual Name Entity Recognition model WikiNEuRal (Tedeschi et al. 2021).

##(will further compare with scraping person name annotations from ToposText, to see which way gets more accurate information.)

The tags for name entity groups retrieved from WikiNEuRal model is appended as a new column in the corpus dataframe. And the tags as “PER”, which means “person name” are further extracted as another column.

	Place_Name	Book	Chapter	Paragraph	Text_ner	PER_na
85	India	2	75	1.0	[{'entity_group': 'LOC', 'score': 0.99636984, ...	[Onesicr
92	India	2	75	1.0	[{'entity_group': 'LOC', 'score': 0.99636984, ...	[Onesicr
93	India	2	75	1.0	[{'entity_group': 'LOC', 'score': 0.99636984, ...	[Onesicr

	Place_Name	Book	Chapter	Paragraph	Text_ner	PER_na
218	Indus	2	98	1.0	[{'entity_group': 'LOC', 'score': 0.999539, 'w...	[]
343	India	2	112	1.0	[{'entity_group': 'LOC', 'score': 0.7412269, '...	[Artemid

The rows containing no person name were dropped and those with multiple person name records were exploded to separate rows.

	Place_Name	Book	PER_names
85	India	2	Onesicritus
85	India	2	Alexander
85	India	2	Alexander
85	India	2	Onesicritus
92	India	2	Onesicritus
...	...	...	...
8842	India	37	Jupiter
8847	India	37	Xenocrates
8866	Indus	37	Democritus
8873	India	37	Nature
8873	India	37	Nature

Within the Indian subcontinent context, the nodes consist of three types, namely **place name**, **person name** and **book number**.

And there are four types of edges being recorded and combined, including the co-occurence of:

1. **place name** and **person name** in the same paragraph
2. **person name** and **book number**
3. **place name** and **book number**
4. **place name** and **place name** in the same paragraph

In the network analysis for place names and person names within the Indian subcontinent context, there are total 158 nodes and 1353 edges in the prepared data.

	Link	Name
0	/people/54	Muses
1	/people/1881	Catullus
2	/people/2300	Cicero
4	/people/382	Manius
5	/people/1515	Persius
...	...	...

```
6972  /people/9015  Bostrychitis
6990  /people/7679       Eusebes
6999  /people/6553         Idaei
7001  /people/9782      Memnonia
7017  /people/6166          Adad

[2764 rows x 2 columns]
```

As manifested in **?@fig-indiantext_clustering**, there is obvious clustering of person names occuring in Indian subcontinent related texts. In other words, groups of person names are tend to be referenced in some specific topics.

##(more detailed illustration will be further conducted.)

Beagon, Mary. 2011. “Chapter Five. The Curious Eye Of The Elder Pliny.” In *Pliny the Elder: Themes and Contexts*, 71–88. Brill. https://brill.com/display/book/edcoll/9789004210073/Bej.9789004202344.i-248_006.xml.

Fantoli, Margherita. 2022. “Statistics and Linguistics: Can We Tell Something More about Pliny the Elder?” https://classics-at.chs.harvard.edu/statistics-and-linguistics-can-we-tell-something-more-about-pliny-the-elder/.

Healy, John F. 1999. *Pliny the Elder on Science and Technology*. Oxford: university press.

Lao, Eugenia. 2016. “Taxonomic Organization in Pliny’s Natural History.” In *Greek and Roman Poetry, the Elder Pliny*, edited by Francis Cairns and Roy Gibson, 209–46. Papers of the Langford Latin Seminar 16. Prenton: Francis Cairns Publications.

Murphy, Trevor. 2003. “11. Pliny’s Naturalis Historia: The Prodigal Text.” In, 301–22. BRILL. https://doi.org/10.1163/9789004217157_012.

Naas, Valérie. 2002. *Le Projet Encyclopédique de Pline l’Ancien*. Collection de l’école Française de Rome 303. Rome: Ecole française de Rome.

———. 2011. “Chapter Four. Imperialism, Mirabilia, And Knowledge: Some Paradoxes In The Naturalis Historia.” In *Pliny the Elder: Themes and Contexts*, 57–70. Brill. https://brill.com/display/book/edcoll/9789004210073/Bej.9789004202344.i-248_005.xml.

Nappo, Dario. 2017. *Money and Flows of Coinage in the Red Sea Trade*. Vol. 1. Oxford University Press. https://doi.org/10.1093/oso/9780198790662.003.0017.

Neelis, J. 2011. “Chapter Three. Trade Networks In Ancient South Asia.” In *Early Buddhist Transmission and Trade Networks*, 183–228. Brill. https://brill.com/display/book/9789004194588/Bej.9789004181595.i-372_004.xml.

Pinkster, Harm. 2005. “The Language of Pliny the Elder.” *Journal of Asthma - J ASTHMA* 129 (November): 239–56. https://doi.org/10.5871/bacad/9780197263327.003.0011.

Pollard, Elizabeth Ann. 2009. “Pliny’s Natural History and the Flavian Templum Pacis: Botanical Imperialism in First-Century C. E. Rome.” *Journal of World History* 20 (3):

309–38. https://www.jstor.org/stable/40542802.

Roller, D. W. 2022. “Introduction.” In *A Guide to the Geography of Pliny the Elder*, 1–14. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108693660.003.

Rydberg-Cox, Jeff. 2021. “Modeling the Sources and Topics of Pliny’s Natural History.” *Umanistica Digitale*, no. 11: 217–29. https://doi.org/10.6092/issn.2532-8816/12521.

Schultze, Clemence. 2011. “Chapter Ten. Encyclopaedic Exemplarity In Pliny The Elder.” In *Pliny the Elder: Themes and Contexts*, 167–86. Brill. https://brill.com/display/book/edcoll/9789004210073/Bej.9789004202344.i-248_011.xml.

Székely, Melinda. 2006. “Eastern Trade of the Roman Empire Based on Pliny the Elder’s Natural History.” *Chronica* 6 (January): 199–206. https://www.proquest.com/docview/2379648941/citation/93A42D142D614235PQ/1.

Talbert, Richard J. A. 2000a. *Barrington Atlas of the Greek and Roman World: Map-by-Map Directory*. Princeton (N.J.): Princeton university press.

———. 2000b. *Barrington Atlas of the Greek and Roman World.* Princeton (N.J.): Princeton university press.

Tedeschi, Simone, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. 2021. “WikiNEuRal: Combined Neural and Knowledge-Based Silver Data Creation for Multilingual NER.” In, 25212533. Punta Cana, Dominican Republic: Association for Computational Linguistics. https://aclanthology.org/2021.findings-emnlp.215.

Tran, Khuyen. 2022. “pyLDAvis: Topic Modelling Exploration Tool That Every NLP Data Scientist Should Know.” https://neptune.ai/blog/pyldavis-topic-modelling-exploration-tool-that-every-nlp-data-scientist-should-know.