Mapping the spatial references in Pliny the Elder's Natural History through distant reading

Dawn, Lizao Zhuang (r0914937)

this is an abstract

```
AttributeError: module 'pandas' has no attribute 'DataFrame'
```

```
AttributeError: module 'pandas' has no attribute '__version__'
```

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.

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

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, (**schultze2011?**) 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 AFC and tree analysis. These two studies also demonstrate how distant reading methodologies offer novel insights into the understanding of ancient treatises.

1.2 Spatial perstpective in *Natural History*

As pointed out by (**beagon2011?**), 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.

In light of this, the present study adopts a focus on the spatial perspective within Pliny's Natural History. A case study employing distant reading methodologies such as statistical analysis, topic modeling, and social network analysis is delved into Indian-related texts, aiming to explore the discourse surrounding India within the work. This endeavor seeks to contribute to a deeper understanding of the inherent complexity and interconnectivity present 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 geographically-related content.

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: 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

Book	Total_length	Place_count	Place_freq
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

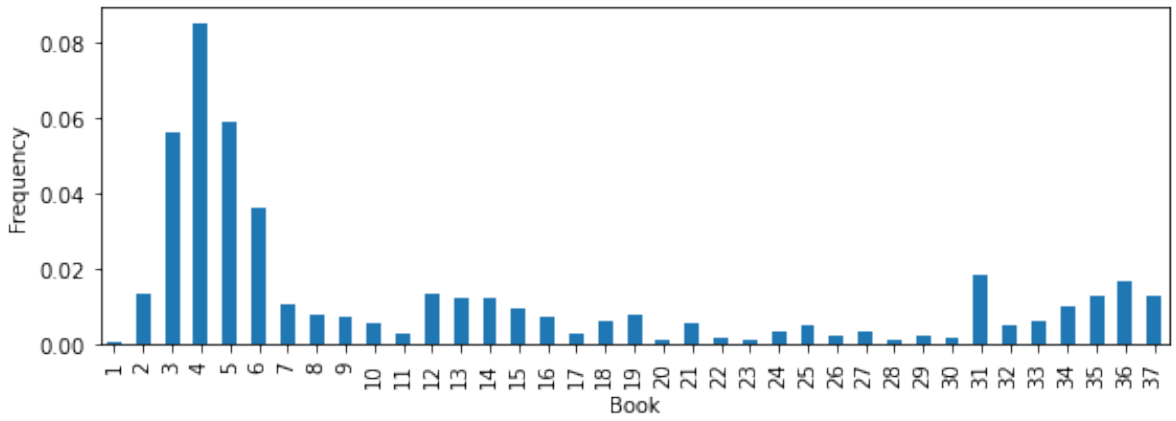

Figure 1: Place name distribution in *Natural History*

2 Research Question

2.1 Spatial perspective

2.2 Focus on India

It emerges in the top 10 places mentioned in the text; it has been already highlighted that the NH is a source of information of the interest of Rome for India (bibliography)

2.3 Discourse about India in Natural History

What information is provided about India? What Indian places are described? How is India described? How is this information structured?

In light of this, the present study adopts a focus on the spatial perspective within Pliny's *Natural History*, employing distant reading methodologies such as statistical analysis, topic modeling, and social network analysis. As a case study, the research delves into Indian-related texts, aiming to explore the discourse surrounding India within the work. This endeavor seeks to contribute to a deeper understanding of the inherent complexity and interconnectivity present in *Natural History*.

3 Methodology

Description of the workflow

4 Data preparation

4.1 ToposText

The original text was written in Latin, and in the

the English version translated by Henry T. Riley (1816-1878) and John Bostock (1773-1846), first published 1855, text from the Perseus Project, licensed under a Creative Commons Attribution-Share-Alike 3.0 U.S. License is used for exploration in the thesis.

containing 797810 tokens and 34548 types, with a scope of encyclopedia in the following structure:

4.2 Description of the dataset of place names

4.3 Description of the dataset of textual passages related to 'India' and 'Indian places' (seize of the corpus: how many tokens?) – check the indication in the pdf file for the coordinates of the India region.

Texts mentioning regions related to "India" in the context were extracted with a range of coordinates referring to Barrington Atlas of the Greek and Roman World by R. J.A. Talbert.

4.4 Manual check (by close reading) of the passages in the dataset: are all the Indian places mentioned in these passages correctly annotated in ToposText?

4.5 Tokenization, lemmatization, remove stop words

4.6 Scrape text with geographical annotation

The text of the whole book has been digitized and annotated with people's name, places' name and coordinates by TOPOSText project since 2012. 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 geographically-related content.

The geographical annotations can be parsed with functions available in Beautiful Soup library, and the first five returned annotations are shown as follows:

```
<a about="https://topostext.org/place/380237SAca" class="place" lat="37.992" long="23.707">A
<a about="https://topostext.org/place/419125LPal" class="place" lat="41.8896" long="12.4884">
<a about="https://topostext.org/place/419125LEsq" class="place" lat="41.895" long="12.496">Es
<a about="https://topostext.org/place/419125SCap" class="place" lat="41.8933" long="12.483">C
<a about="https://topostext.org/place/419125PRom" class="place" lat="41.891" long="12.486">Ro
```

With defining a function, all the texts with the geographical annotations can be parsed and stored as a dataframe, containing information in 8 columns as:

1. Unique ID assigned
2. ToposText_ID (which identifies the distinct location)
3. Place name
4. Reference (indicate where the place name occurs in the book)
5. Latitude
6. Longitude
7. Book number the place mentioned in
8. Chapter number the place mentioned in

9. Paragraph number the place mentioned in
10. Plain text of the paragraph where the place is mentioned

	UUID4	ToposText_ID	Place_Name	Refer
0	f8bbe55f-283c-49a2-ac89-35e0b24b46fe	https://topostext.org/place/380237SAca	Academy	urn:c
1	a804496a-f5b4-4359-abfe-a19f43ca58ba	https://topostext.org/place/419125LPal	Palatine	urn:c
2	2d69bd88-f0f7-411c-9261-989e4575cdcc	https://topostext.org/place/419125LEsq	Esquiline	urn:c
3	3e44e671-8506-46b5-9d1f-6bb9179427bd	https://topostext.org/place/419125SCap	Capitol	urn:c
4	18d4d160-548b-4f2e-a2ba-481cbe7b29c3	https://topostext.org/place/419125PRom	Rome	urn:c

There are 5595 locations mentioned in book 1-11 and 3281 locations mentioned in book 12-37. The combined dataframe for the whole book, has the shape of (8876, 10). And the output has been stored as .csv for record.

4.7 Scrape text of the entire book

The text of the entire book is also scraped as a reference.

	UUID4	Reference	Book	Chapter	Paragrap
0	9dc3ebec-2788-41d5-aae7-c5822472d873	urn:cts:latinLit:phi0978.phi001:1.1.1	1	1	1.0
1	5a3b400b-e2fe-42e0-b8e8-425b69257ae0	urn:cts:latinLit:phi0978.phi001:1.2.1	1	2	1.0
2	3e6bdc07-568b-43ae-8ebd-537355f33d4f	urn:cts:latinLit:phi0978.phi001:1.3.1	1	3	1.0
3	88f9960b-bf90-40da-b140-48344acaf23e	urn:cts:latinLit:phi0978.phi001:1.4.1	1	4	1.0
4	95522a01-f6dd-43ea-94b0-82eacc2b3a8f	urn:cts:latinLit:phi0978.phi001:1.5.1	1	5	1.0

The combined dataframe for texts in the whole book has the shape of (3493, 6). And the output has been stored as .csv for record.

```
(797810, 34548)
```

```
(2812850, 28405)
```

```
(85665, 6608)
```

5 Data Analysis

5.1 Word frequency

5.2 Topic modelling

5.3 Network analysis for Named Entity

6 Conclusions

7 Old structure

7.1 Overview of geographical related texts

What topics popped up from the context of place names?

7.1.1 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 4). As depicted in Figure 2, 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 4: 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

Book	Total_length	Place_count	Place_freq
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

7.1.2 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" + '
```

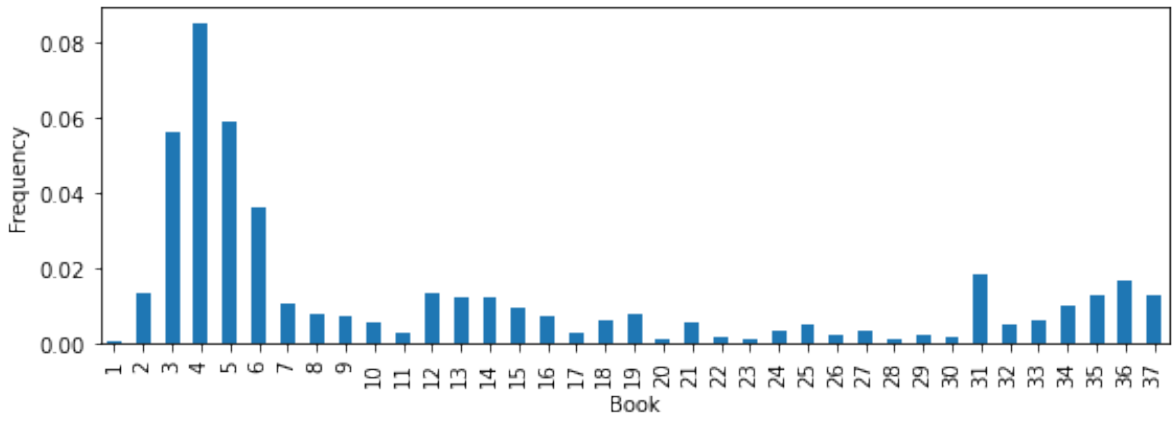

Figure 2: Place name distribution in Natural History

```
 '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.

7.2 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*?

7.2.1 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 5.

Table 5: 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
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 3, is the prominence of India—a region outside the Mediterranean—despite its high frequency of mentions.

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

Figure 3: Place name distribution map

7.2.2 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:[2]

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.

	UUID4	ToposText_ID	Place_Name	Refe
85	30629c68-c5a2-45ec-9c6e-ae2e0ceaecab	https://topostext.org/place/300740RInd	India	urn:
92	958f3c3d-bbfc-4488-8009-0185494ace05	https://topostext.org/place/300740RInd	India	urn:
93	11517672-3e13-4fcc-ac4a-f7ce58f814a7	https://topostext.org/place/300740RInd	India	urn:
218	617dbd2d-e1cf-4858-91b4-ed6a959c9cee	https://topostext.org/place/254683WInd	Indus	urn:
326	c2dad95b-6687-4cca-9482-4c449fafdd88	https://topostext.org/place/340670RBac	Bactria	urn:

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 data frame are: ['India' 'Indus' 'Bactria' 'Ganges' 'Acesinus' 'Oxus' 'Hydaspes' 'Taprobane' 'Arachosia' 'Muziris' 'Baragaza' 'Aria' 'Ceylon'].

[2]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.

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 4. 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 5 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.

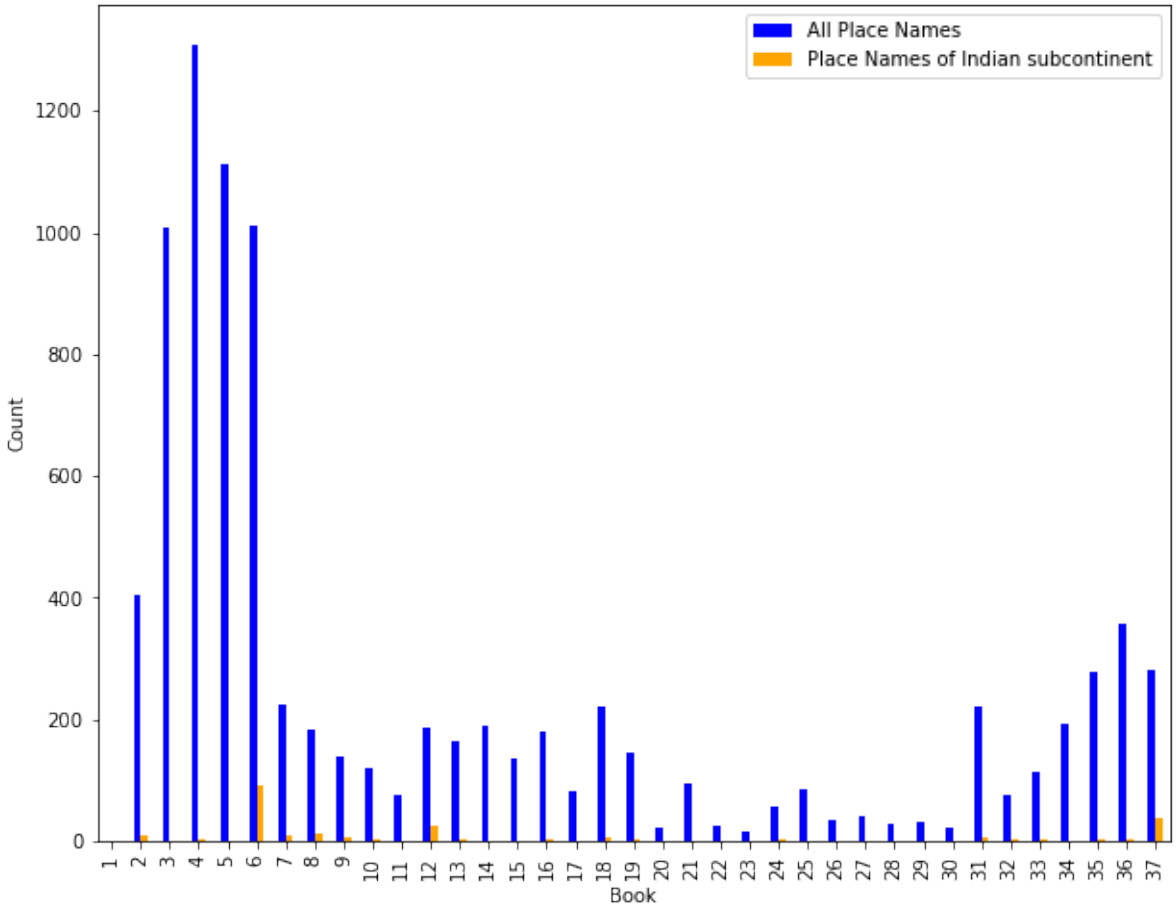

Figure 4: Occurence count for all place names and place names of Indian subcontinent in each book

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.

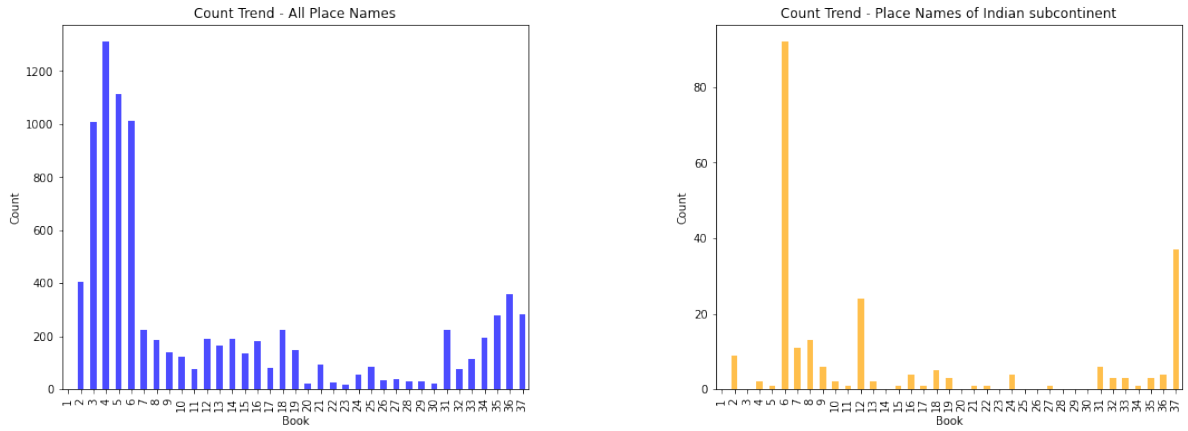

Figure 5: Occurence count for all place names and place names of Indian subcontinent in each book_different y-axis scales

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.

7.2.2.1 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 6 and Figure 7.

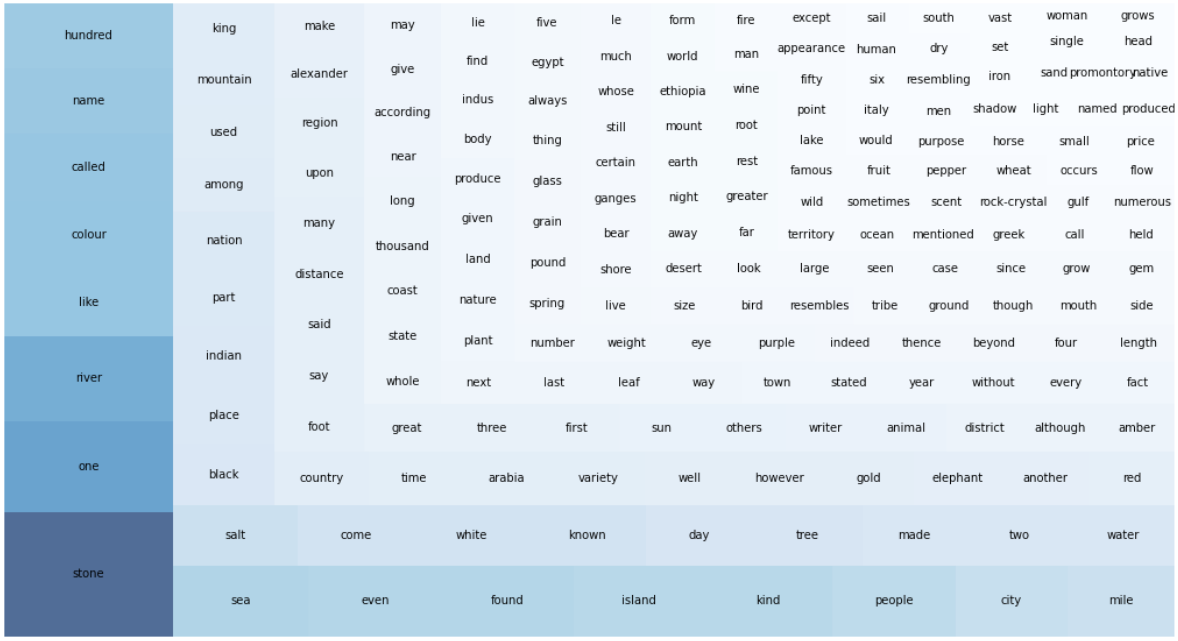

Figure 6: Top 1% frequent words in Indian subcontinent related text as tree map

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.

Figure 7: Top 1% frequent words in Indian subcontinent related text as word cloud

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.)

7.2.3 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.

7.2.3.1 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 8, the size of the node represents the betweeness centrality a place name mentioned in the book, and the weight of edge between two nodes represents 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.

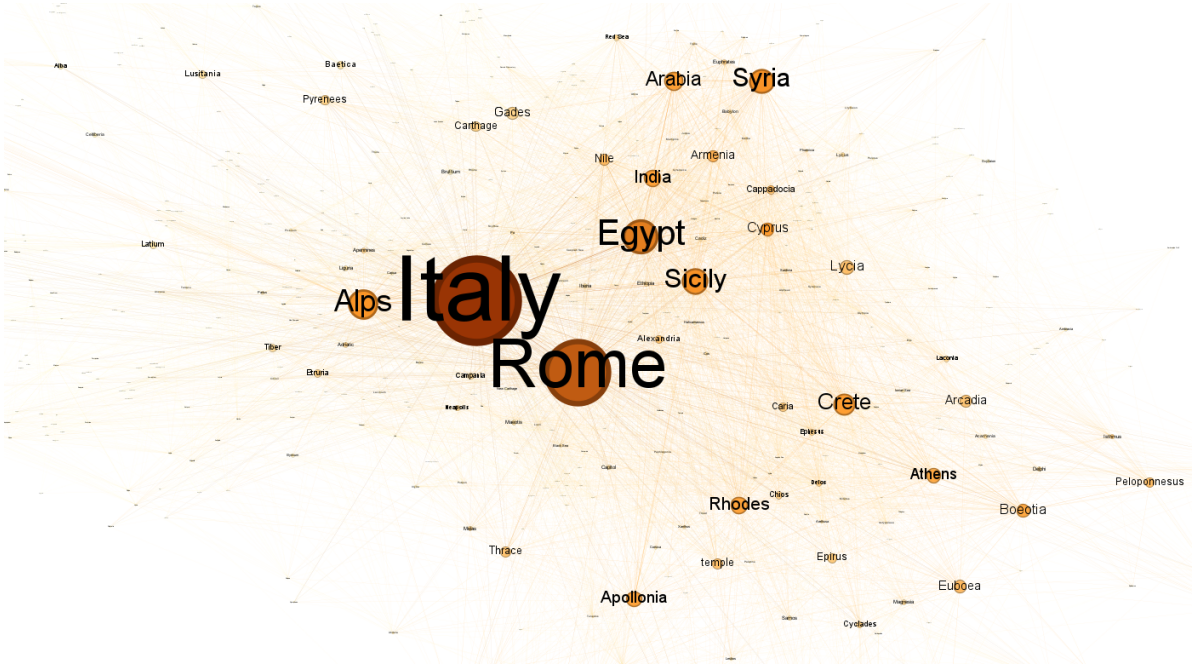

Figure 8: Network graph for place namesmentioned in Natural History

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
218	Indus	2	98	1.0	[{'entity_group': 'LOC', 'score': 0.999539, 'w...	[]
326	Bactria	2	110	1.0	[{'entity_group': 'LOC', 'score': 0.87196684, ...	[Ctesias,

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 164 nodes and 1458 edges in the prepared data.

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.)

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.

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.

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 <Small Class="caps" Xmlns:m="http://Www.w3.org/1998/Math/MathML" Xmlns:mml="http://Www.w3.org/1998/Math/Mat Xmlns:xlink="http://Www.w3.org/1999/Xlink">c.e</Small>. Rome." *Journal of World History* 20 (3): 309–38. https://doi.org/10.1353/jwh.0.0074.

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.

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.