Process Book

in viz veritas, com-480 data visualization
spring semester 2023

Hannah Casey
Zoé Jeandupeux
Erik Wengle

Introduction

The dataset for the visualizations is a collection of old travel guides for Italy from the rare book collection of the Bibliotheca Hertziana, Max-Planck institute for art history in Rome. The collection itself contains around 800 documents from the years 1550 up until 1930, all of which have been digitized, but we are working on only the part that has also been transcribed using Transkribus, an AI supported platform for text recognition and annotations.

The dataset was obtained by scraping the Transkribus collection of the Bibliotheca Hertziana, who kindly gave us access to the data.

The Path to the final result

One of the main challenges in our project was dealing with the unstructured data of the books. We started very early implementing NLP methods to extract interesting information on the books in order to create engaging visualizations. In our initial design plans for the website, we decided to split our website into two major overviews: One describing the entire set of the books, focusing on the given metadata such as the publication year and location. We then wanted to offer a closer glimpse on single books, and decided to restrict ourselves to the set of English books in order to efficiently employ NLP methods on the data.
We had very ambitious goals for the closer inspection of the books, such as implementing a timeline of the places visited throughout the book, displayed on a map. However, this proved to be rather challenging, which is why we opted to create a diverse set of visualizations focusing on different aspects, such as locations mentioned in the book, the sentiment development throughout a single book and the combination of sentiment and locations.

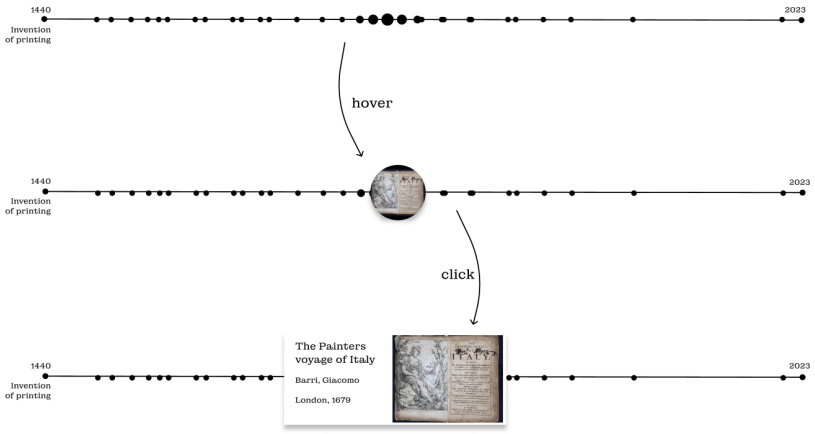
In the following subparts, we shall explain how the different visualizations came to be and how we justify our design and implementation choices.

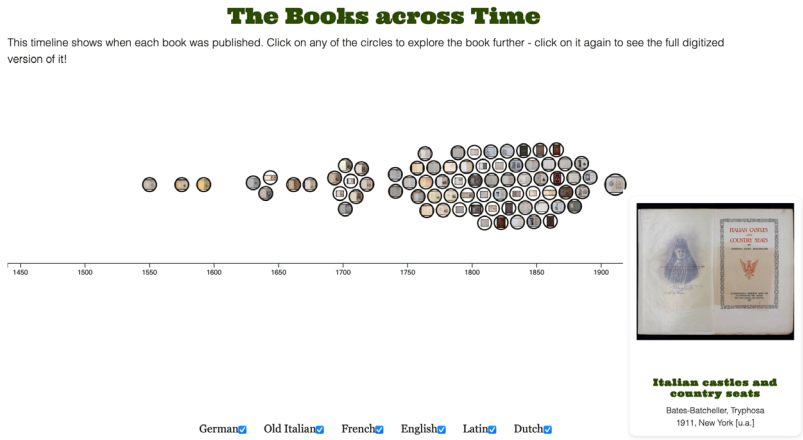
Visualizations

Timeline

The goal for the timeline was to showcase the age of the books as well as giving an overview of the entire collection at once.

The first sketch for milestone 2 was a simple timeline with circles representing each book, also showcasing the interaction a user could have with the visualization. The circles should increase in size when hovered and when clicked, a more detailed view of the book would pop up.





Due to the nature of the data, specifically that multiple books have the same publication year, the representation of the books was changed from a line to a more organic structure, where the books with the same year of publication are close together but not on top of each other.

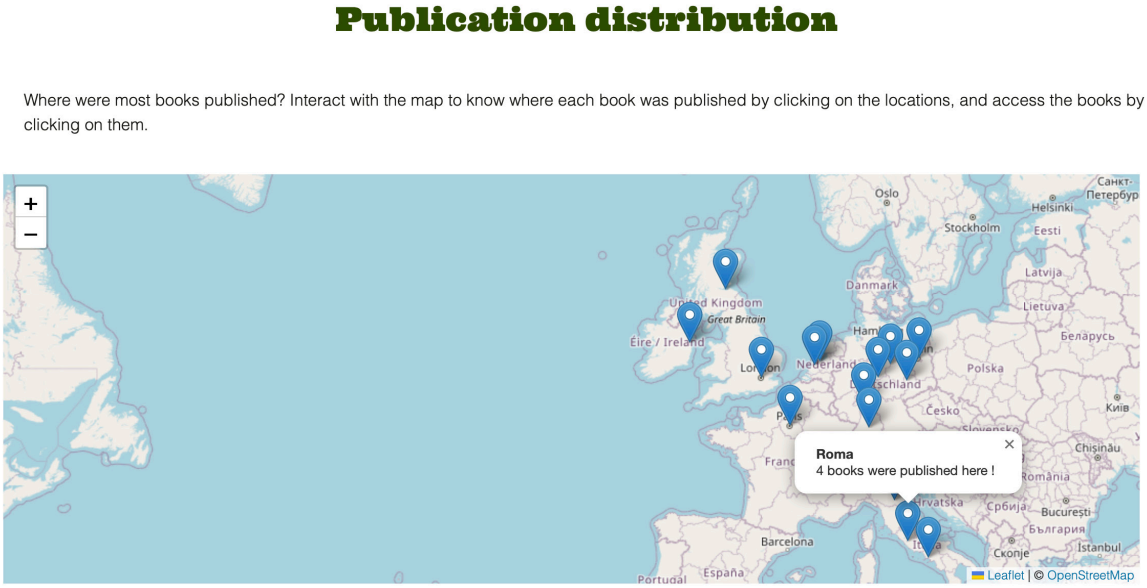
For each book a thumbnail image was selected, to best showcase the books as a collection of both rare and precious objects.

In order to connect the visualization with the database, on clicking on the book cards a new window opens up, giving the visitor the opportunity to explore even further. The links used for the books use IIIF manifests and hence should be stable.

Another feature added to the final visualization are checkboxes underneath it, to brush the data according to its language.

Publication map

This graph was not part of the milestone 2, but it was made to show some metadata on the books while fitting the design plan of the website. The data came from the exploratory data analysis of the first milestone, and then was expanded to get the titles of the books and the urls to the transcripted books. Similarly to the timeline, it is a plot to visualize geographically all the books and where they came from.



Bookshelf selector

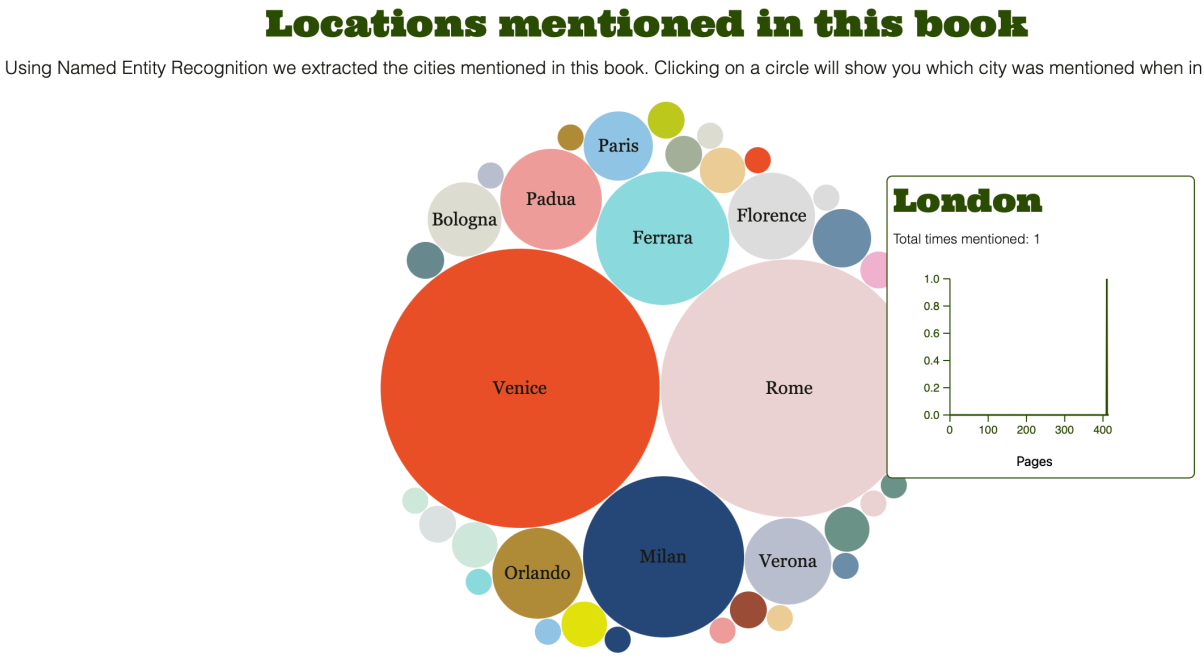
The book selector changed dramatically from the first sketches to the final design. Instead of selecting first the author and then the work, the books are represented as bookcards in a row, to imitate the way the books are found in the Bibliotheca Hertziana. This way the selector can be used to display the visualizations of the data regarding this single book. We opted for emulating the feeling of picking out an interesting book from a bookshelf to add playfulness to our design.

Bubblegraph

The main goal of this visualization is to showcase the place names and their frequency extracted from each of the English books in the collection.
The final design showcases more information than the sketches, as it also shows on which pages any selected city was mentioned when clicking on the circles.
The frequencies of the cities were extracted using Named Entity Recognition form the Spacy library, and cleaned afterwards to contain only true city names. There are still some minor mistakes in the data, but the visualization works very well on most books.

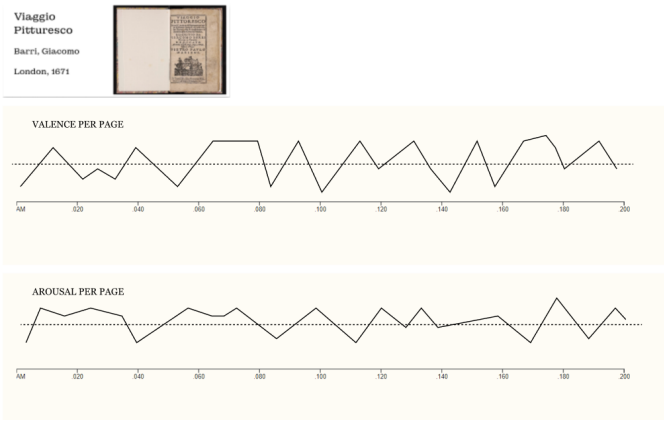
It allows users to explore the cities that were mentioned in different books, and shows that the cities mentioned really do change depending on the itinerary of the traveler, which can be often found in the title of the book.

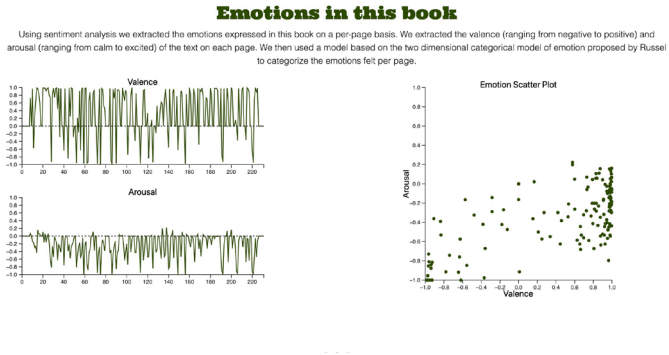
5



Sentiment Analysis

To employ the sentiment analysis, we focused on the two dimensional categorical model of emotion proposed by Russel. For this we extract the valence and arousal for each page and categorize it using Russel's model. While implementing the visualization, it quickly became evident that the obtained values were relatively noisy, making it hard to achieve the storytelling we envisioned initially. Our goal was to display the valence and arousal on a per-page basis, in order to see how the emotions of the author develops throughout the book and whether it could be affected by the locations visited or other factors mentioned in the book itself. In the end, we tried other means of extracting the emotions or normalizing them, but could not come to a conclusion. We still kept the visualization in, as in essence it represents our core idea - the visualization would be much more effective if coupled with a dataset that accurately describes the sentiment per page of a given book.

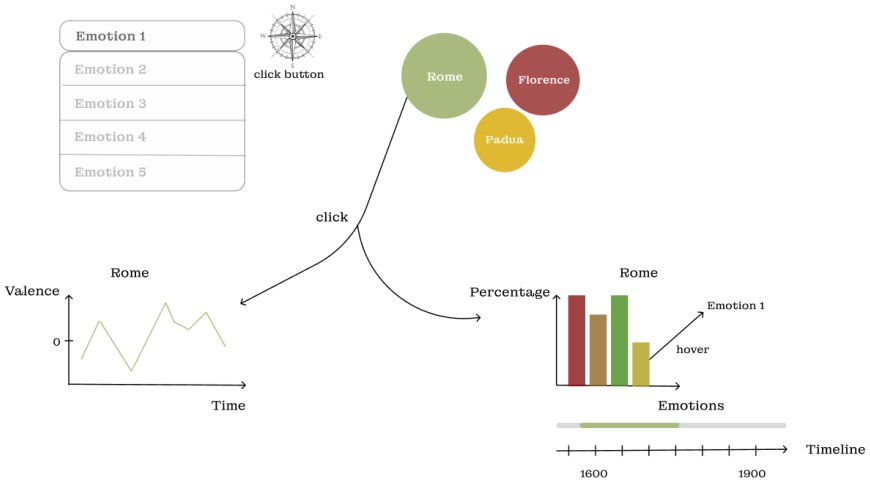




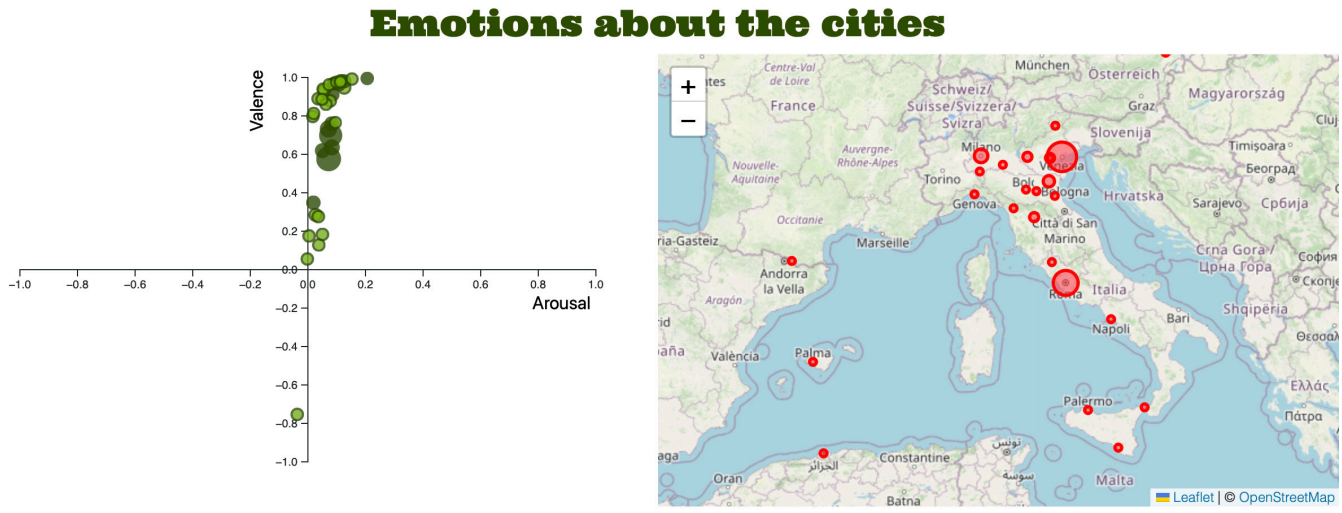
The overall design of the viz changed in order to fit the theme of our bookshelf selector. We also decided to omit displaying the page itself, since the reliability of the sentiment analysis is shaky, and the page contents per transcribed book varied widely, making it impossible to display it in a uniform manner along with the corresponding visualizations.

Emotions in locations

The initial design of this plot, chosen for milestone 2, was abandoned. Firstly, the data did not allow us to render insightful visualizations for the emotions associated with the cities, because the emotions were too similar from one city to another. The lack of diversity in the emotions for each city forced us to find another way to link both locations and sentiment.



The second reason was the website design idea to separate into two parts the website, with on one side plots showing all the books and their metadata and on the other the books individually. The plot was then adapted by computing the emotions per city per book. We then displayed each city on a valence-arousal graph. The size of the dots depends on the frequency a given city was mentioned throughout the book, it therefore also combines the bubblegraph and the emotion plot in this way. The map was added to show geographically where the places named are, almost allowing to create an itinerary and distinguishing the locations actually visited by the author from the ones they only mentioned. It can be noticed that it is a difficult task to determine which emotion could be attached in a book to which city and there are many reasons for it. Firstly, to compute the emotion per city per book, we simply average the valence and arousal scores on the pages where the given city was named. In milestone 2 we wanted to average by putting more weight on the pages where the city was named more times, but it actually heavily depends on the writing style of the author.
The emotion in itself also depends on the style of the author, and may not represent accurately what they feel, although inside a same book with the same writing style we can notice some great differences in the valence score, giving us some insights in which locations were written with most positive or negative vocabulary.



Have a look at the cities mentioned by the author . Can you guess his itinerary according to their position on the map ? How did he talk about them ? Discover it by hovering over the emotion plot. Admire images of the cities manually retrieved from the travel guides by clicking on the darker points.

Hovering on the map shows where the city is represented on the valence-arousal plot, by turning it red. To showcast some of the images found in the books without having to look for them manually inside the transcripted books, we manually retrieved images corresponding to the cities mentioned by the books. They can be displayed by clicking on the darker points of the valence-arousal graph. We could not retrieve images for all 440 different cities named throughout all books, but the most popular italian ones could be found. The images displayed do not necessarily come from the book that is currently selected, but come from the collection.

Peer assessment

Breakdown of parts

Hannah: Website design, Bubble Graph, Timeline, Book cards, Scraping + Preprocessing, Named Entity Recognition, Image selection for book thumbnails, communication with Bibliotheca Hertziana, Process Book

Erik: Website design, Implementation of the Bookshelf Selector, Scraping + Preprocessing, Sentiment Analysis, Emotions visualizations, Process Book, Screencast

Zoé: Publication distribution visualization, Emotion for cities visualization, Map for books visualization, Exploratory data analysis, Image selection for cities, Preprocessing, Process Book

Acknowledgments

Special thanks go to the Bibliotheca Hertziana, specifically Golo Maurer and Elisa Bastianello for their support in finding and obtaining the dataset.
Shoutout to Battsooj Enkhbaatar, who helped us with the layout of the process book and enormous thanks to Daniella Rogers, who supported us on selecting the fonts for the website.