Exploration of Italian-published ancient book collection in Leuven

Dawn Zhuang (r0914937)

Introduction to Digital Humanities: Final Report

1 Introduction

Based on a pre-processing of the dataset via OpenRefine, which provides cleaned data with separated information about language, affiliated library name, carrier/media type, publication details (including publish time/period, place of publication, and author) and title of over 90% books in the dataset, it is feasible to conduct a further exploration about the collection with help of data analysis and visualisation tools like Tableau and Python. The exploration covered three aspects of the collection: an overview of the library collection, distribution trend revealed in the collection and an attempt on figuring out the trendy keyword of book titles.

2 Tools and Method

There are two analysing tools applied in the exploration, Tableau Public is used for statistic analysing and data visualisation, and Python is used for text analysis. Additional manipulation on the records of the dataset was conducted via OpenRefine.

Tableau Public

Various charts were created on Tableau to display different aspects of the dataset. General overview and data description were made through text tables. Bar charts and packed bubbles helped with visualising statistical comparisons. The periodical analysis were visualised by line graphs. And symbol map and tree-maps were applied in geographical mapping and manifesting the result of text analysis.

The built-in functions of percentage calculating and ranking were applied to enrich the dimensions of statistical comparisons.

And an interactive storyboard was created to showcase the combination of data visualisation, readers can click on different dimensions as sorting or filtering condition to compare the statistical changes and trends.

Python

Modules and applications in Python 3 are listed in table 1:

Python3 Module	**Application**
NLTK	Tokenize Italian texts; Generate collocation bigrams
CLTK (The Classical Language Toolkit)	Tokenize Latin texts
pandas	Create data frame from tabular dataset
string	Exclude punctuation in texts
csv	Create csv file for visualisation in Tableau

Table 1: Modules and their applications

3 Data and Results

With the pre-processed dataset, an analysis of the 1161 Italian-published ancient books collected in Leuven libraries in aspects of library collection, book production, and book topic trend has been conducted.

Library collection

By sorting the data in different ways, an overview of how the libraries hold this set of collection is revealed.

1. **Majority of publication period and language of the books**
   By sorting the collection by publication period (the year of publication corresponds to half a century) and language (books with two languages will be counted twice), it shows that the vast majority of the collection was published during 1541-1600, occupying over 83% of total, and the major languages of the books are Latin (over 76%) and Italian (over 20%).

2. **Overview of the carrier and media type of the collection**
   Sorted by different carrier and media types, there are 1143 books kept as unmediated volumes, 16 of them were stored in microfiche or other microforms, and 2 of them were in form of computer disc.

3. **Overview of the number of collections in different libraries in Leuven**
   It shows that over 68% of the Italian-published ancient books are collected in Maurits Sabbe Library, which is the research and heritage library of the Faculty of Theology and Religious Studies of KU Leuven.

Figure 1: Collection sorted by library (top 3)

Book distribution

In addition, with the information on book production, the geographical trend of publication in the publishing period can be further analysed.

1. **Geographical publication center revealed**
   Despite there being some unclear geographical information in the dataset, it is clear that over 60% of the books in this collection were published in Venice.

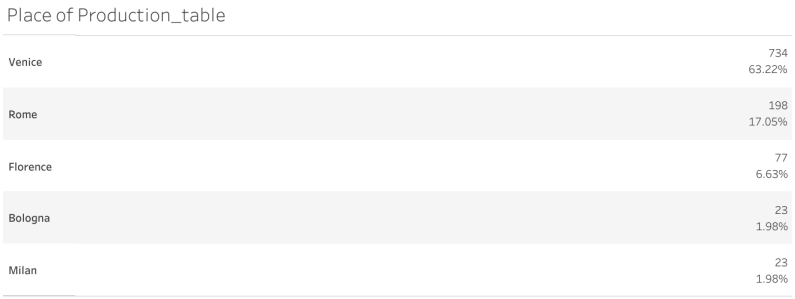

Figure 2: Collection sorted by place of publication (top 5)

And as the comparison of publication centers in different time periods shown in Figure 3, Venice kept ranking at the top during all time periods for this specific collection.

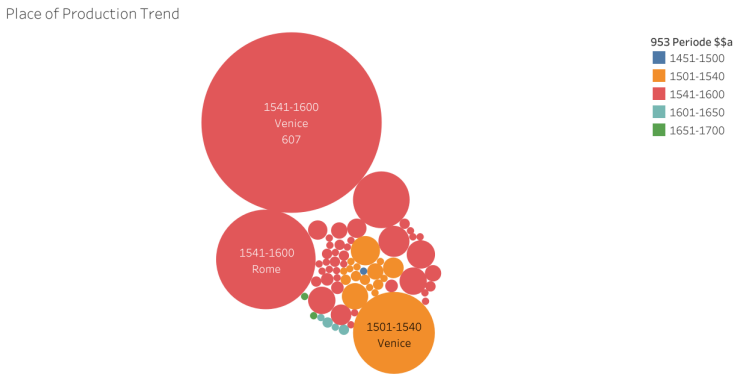

Figure 3: Trend in place of production

According to van der Sman (2000), Venice indeed acted as one of the centers in Euroupe in book publication and distribution during the early modern period. The trend shown in the captioned collection complies with this finding.

2. **A hypothesis of publication peak period**
   Figure 4 compares publication quantity counted by year for different languages. It manifests that most of the books in this collection, no matter in what language, were published during 1567-1607. It might indicate there was a peak in book publishing in Italy around this period. However, the conclusion could not be drawn without supporting from more data about Italian-published books in the same period.

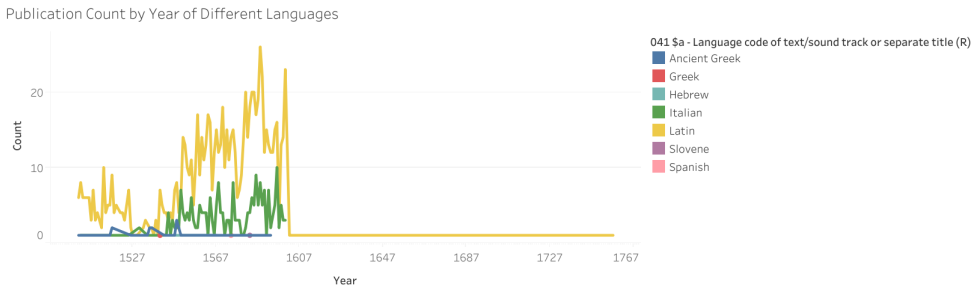

Figure 4: Book publication counted by year (grouped by languages)

3. **Popular authors of the collection**
   Sorting the books by author, the top five authors with the most publications in the collction are Thomas Aquinas, Nibolas Cleynaerts, Duns Scotus, Francisco de Toledo, Pius and Aristoteles.

   This finding may indicate the popular authors of publications in Italy during 16-18 centuries. Still, further references are needed to support the conclusion.

Figure 5: The five most frequent authors of the collection

Book title trend

And in order to have a grasp of the trendy topic in the collection, a frequency list and a collocation bigram for the words in book titles were generated for the two major book

languages, Latin and Italian. Different language processing modules have been used and the stop words were excluded for analysis.

1. **Frequency list and collocation bigram of Latin titles:**
   The 10 most frequent Latin titles in the collection are: ('libri', 107), ('libros', 60), ('eiusdem', 54), ('liber', 53), ('aristotelis', 52), ('tres', 43), ('ioannis', 42), ('quibus', 39), ('francisci', 39), ('omnia', 36), ('petri', 31).

   The top 10 collocation bigram of the Latin titles are: ('libri', 'tres'), ('tomus', 'primus'), ('pont', 'max'), ('societatis', 'iesv'), ('onvphrii', 'panvinii'), ('dvns', 'scoti'), ('pars', 'prima'), ('librum', 'sententiarum'), ('totius', 'anni'), ('commentariorvm', 'theologicorvm').

2. **Frequency list and collocation bigram of Italian titles:**
   The top 10 frequent Italian titles in the collection are: ('libri', 33), ('sopra', 29), ('libro', 17), ('trattato', 16), ('tre', 16), ('discorso', 15), ('dvello', 14), ('principi', 14), ('battista', 13), ('lettere', 12), ('vita', 11).

   The top 10 collocation bigram of the Italian titles are:: ('accessere', 'porrò'), ('capponi', 'porrecta'), ('cvm', 'elvcidationibvs'), ('editis', 'accessere'), ('elvcidationibvs', 'formalibvs'), ('formalibvs', 'seraphinvm'), ('luculentissima', 'subtilissimaq́'), ('mvtio', 'ivstinopolitano'), ('porrecta', 'editis'), ('porrò', 'luculentissima').

The result shows that the most frequent words of both languages are related to "libri" ("books" in English) , which may indicate the collection includes a large proportion of comments or notes of classics, or the book titles were added by the later collectors (maybe the book did not have an original title and was annotated as "book of..."). Precise translations and cultural and historical contexts are needed for further discussion.

4 Assessment on cleaning status

Although the dataset has been cleaned and enriched during the pre-processing session, additional manipulations were conducted to fit the necessity of further exploration.

1. **Append record ID and fill down values for split cells**
   In the pre-processed file, cells with multiple values were split into new rows under the same record for clearer display. However, when sorting the data by specific dimensions, split rows without a record ID could not be counted. To solve this problem, a new column has been added filling with the same record ID of separated rows via grel:row.record.index. And values of the same record in separated rows were filled for the columns being analysed with the built-in function of "fill down" in OpenRefine.

   After the manipulation, each row was assigned an ID, and rows of the same record have the same ID and filled values of other columns. It enables records to be counted (set "count distinct" as a counting restriction in Tableau) by different sorting dimensions even if they have separated rows in one column.

2. **Further cluster and reconfirmation of city names**
   In order to apply a built-in mapping function of Tableau Public, city names of production locations should be as precise as possible to correspond to the geographical

coordinate automatically. Thus the values of the separated column of the "production place name" in English were manually checked and re-clustered with a reference to the "Latin Place Names" database. But there are still several ambiguous values that can not be determined for the city name. Table 2 lists the undetermined city names for a record.

City name value	**Note**
Camerini	might be an ancient city
Dominicum de Farris	publisher information, did not match its metadata
Fani	failed to locate
fin officina hęredum Luceantonij Juntę	failed to locate
Mvssiponti	failed to locate
Pervsiae	failed to locate
Placentini	failed to locate
Romæ in Vaticano	not sure if it refers to Rome or Vatican City
Rua	original value [Rua], failed to locate
S.l.	abbrivation, failed to locate
V. 2	abbrivation, failed to locate
Ενετιεσι	Ancient Greek, failed to locate

Table 2: Undetermined city names

With the concrete city names, most of the records in the dataset could be mapped in the interactive dashboard, which shows clearly the city and number of book published during different time periods.

5 Future Work

Upon the findings in the exploration, the research on this dataset could be enriched with a comparison with other datasets of ancient books published in the same period or same region. And the text analysis is just a preliminary attempt, though processed with reliable modules, the correctness of tokenization in Latin and Italian should be evaluated, and precise translation of the words is needed for further analysis.

6 Visualisation output

- The general interactive data visualisation output could be referred in the link: Italian-published Ancient Book Collection in Leuven Libraries.
- The codes of text analysis and embedded Tableau storyboard can be referred in the attached Jupyter Notebook file.

References

van der Sman, G. J. (2000). Print Publishing in Venice in the Second Half of the Sixteenth Century. *Print Quarterly*, 17(3):235–247. Publisher: Print Quarterly Publications.