# The Text Analysis Prototype for Galileo's Library and Letters Online: GaLiLeO

### anonym

[cc-by](https://creativecommons.org/licenses/by/4.0/) ©. Published by De Gruyter in cooperation with the University of Luxembourg Centre for Contemporary and Digital History. This is an Open Access article distributed under the terms of the [Creative Commons Attribution License CC-BY](https://creativecommons.org/licenses/by/4.0/){:target="\_blank"}

library("IRdisplay")  
display\_png(file="./media/CoverImage.png")



This article is the author's report on a collaborative project, which currently exists as a prototype that was presented to a small group of subject experts in September 2018. The article, with permission, makes use of the author's contributions to a presentation made at that prototype demonstration workshop with collaborator Hannah Marcus, Historian of Science at Harvard University. Where Marcus' work has focused on metadata and a scholarly editorial apparatus for correspondence to and from Galileo Galilei, this article takes as its object of evaluation the code written in R to explore and contextualize the texts, including the letters, but also books written by Galileo and those that he owned. Portions of this article appeared for a time on the author's personal website, but have now been removed.

text analysis, history of science, digital archives, corpus creation, Galileo Galilei

This article reports on a prototype digital laboratory of tools for the discovery of interpretative pathways through historical materials related to Galileo Galilei (1563-1642). GaLiLeO: Galileo's Library and Letters Online allows a user to contextualize a document or term within a broader corpus of personal and semi-public letters, prefatory letters, and book-length texts from Galileo's library. Given three features of the tool set and corpus, the article argues for a third digital, analytical space between the digital archive and a corpus used for computational analysis. First, the code for GaLiLeO was developed in 2016-2018 based on specific modeling principles from Digital Humanities that are now several years old and merit re-contextualization within shifts and DH and adjacent fields (,). Second, the tools for exploration, question making, and question answering in GaLiLeO do not aim for corpus-scale descriptive models or interpretations, but rather to highlight documents of interest for historians and literary scholars working outside DH with humanistic methods of document selection and interpretation rather than those offered by text mining, computational lingustics, and data science. Finally, the project is unfinished and based on an incomplete corpus, but has nonetheless yielded research results because of the tools' focus on identification of pathways of a few potentially related documents, not corpus-wide patterns. By avoiding corpus-level declarations, the preliminary computational tools in GaLiLeO represent a prototype for emphasizing relationships across documents that may have resisted or otherwise been excluded from the organizational logic of (digital) archival and other post-hoc metadata.

## Project History and Contextualization

This article engages with questions related to the usefulness of an unfinished, yet entirely functional, project that was designed to be experimental and iterative in recognition of its incomplete corpus of documents and their condition. Galileo's Library and Letters Online (GaLiLeO) was developed to start a conversation about the characteristics of new digital tools in the interdisciplinary field of Galileo Studies. The GaLiLeO prototype tools and the process of their creation offered a way to brainstorm about a future larger-scale project, work with a cohort of experts in various areas of 17th-century Italian culture, and draft new digital and computational analytical tools alongside scholarly editorial apparati. The GaLiLeO team's enthusiasm and imagination quickly outpaced the resources of time and funding. Feedback on a 2017 proposal for a Digital Humanities Advancement Grant through the National Endowment for the Humanities in the United States () pointed to ways to scale back the initial scope to a proof-of-concept for the combined text analysis of the two corpora and their metadata, working with materials at hand to focus more on method. The audience for the prototype of this tool suite was a group of content specialists that do not consider themselves digital humanists, but are historians and literary scholars in search of tools for meaningful exploration of large digitized archives and primary materials related to Galileo Galilei (1543-1642).

Although digitized resources related to Galileo are plentiful, computational and quantitative means of navigating the materials are not. The largest digital resource of texts for Galileo specialists is the galileo//thek@ (Galileoteca) developed through 2016 and maintained by the Galileo Museum in Florence, Italy. () The Galileoteca provides access to scans of primary documents, scholarly editions, and selected secondary scholarship alongside explanatory materials related to the Museum's holdings. A subset of manuscripts are searchable by their shelfmarks within the Galileian manuscripts at the National Central Library of Florence, a person or place name, dates, and words in titles or incipits. A second digital layer offers access to OCR from scans of the documents included in the late-19th century National Edition of Galileo's works and correspondence compiled by Antonio Favaro (1847-1922) and his team of collaborators. () These 20 volumes, now supplemented with appendices recently edited by Galileo experts, represent a corpus too large to memorize and disorienting to read in a linear fashion. Long texts (including some letters) are included in thematic volumes, all other letters by date. Scans and transcriptions of documents are organized according to their location in the national edition and the digitized text reflects Favaro's editorial interventions, most often in materials written by Galileo, rarely in texts authored by others. The excellent curatorial and editorial team at the Museo Galileo has also added other interpretive layers to the digital interface to help visitors find appropriate documents easily through a searchable critical lexicon.

This kind of resource requires the user to have details about the documents' creation or contents prior to beginning a search. In the Galileoteca, specialists who know what person or document they want to find, those who are studying a chronological period in Galileo's lifetime, or have a thematic or topical research question in mind can use the resources to access related materials. Users can search word forms and lemmas within certain areas of this digital archive or across eleven kinds of materials that also include biographies, indices, and iconography. It is an excellent finding aid for those who already know the specific text(s) they seek and can navigate the organizational logic of the manuscripts or the national edition. This digital archive has successfully remediated the structures that have made these documents findable while adding the functionality of keyword searching in metadata and any machine-readable texts.

At the same time, the collection of documents is extensive and scholars continue to discover new features that broaden understandings of Galileo and his contemporaries even without a corpus-level computational study of the contents. These mutually compatible aspects of Galileo Studies, that is, the continued interest in small-scale patterns and the possibilities for larger-scale analysis, sparked two digital projects that evolved separately, but simultaneously until 2016: developing an editorial apparatus for Galileo's correspondence based on close reading of the available letters and creating a digital and computational library of the books that Galileo owned. From the perspective of History of Science, Hannah Marcus had worked with a team of scholars to identify thematic overlap in the surviving correspondence related to Galileo. Part of that project also involved careful attention to metadata about the letters, which has resulted in new findings. () When the project leaders met in April 2016 at Stanford University, the points of connection and similar goals of the two collaborative efforts coalesced around two concerns: not repeating existing resources and establishing research outcomes while the larger projects unfolded. Neither corpus was complete, nor may they ever be definitive, but the expanded access they provide is fertile territory for new investigation - with the right tools.

Galileo's Library and Letters Online derives its title from this tension in its identity. Naming credit goes to Hannah Marcus and Morgan MacLeod after much discussion about what this project would be. The creators struggled with the identity of the hybrid tools-text nature of the platform. The tools were not meant to create a digital archive, which would emphasize findability and display a range of related material, a role already filled by the Galileoteca. The aspirational goal was similar to the ARTFL Encyclopédie (), but the projects were being developed by the two PIs and a few rotating graduate and undergraduate students without dedicated funding or broad technical support. Even though Marcus' team had created a rigorous and extensive scholarly interpretive layer for the correspondence, a digital edition would have implied a single organization and interpretation of the texts. Using the vague "online" descriptor avoided these problems of identity. It gestures toward actions of aggregation, exploration, and experimentation that are enabled by the web. While bringing together plain, full text of the letters and limited contents from the library would create bridges across otherwise siloed resources, the immediate aim was not to make a claim about a collection, but to evaluate the connections found between and across the texts of which it was composed as it grew.

This short term goal shaped design in a way that links traditional humanistic study of primary materials with the outcomes of larger-scale analyses using methods from Digital Humanities, data science, and computational linguistics. Because of the iterative nature of the corpus development (described in more detail in section 1.3), the tools ultimately needed to suggest pathways, not patterns. That is, quantitative and graphical results could not be the end in themselves, they were designed to surface small subsections of the corpus with potential relationships. The tools needed to be designed in such a way as to facilitate contextualization, comparison, and expansion of points of entry into the primary sources, keeping in mind students new to the historical material and peer scholars outside the Digital Humanities looking for a way to investigate their own texts. While the corpus may be of interest for large scale study, the audience for the tools was colleagues who conduct traditional historical and literary interpretations of primary sources. GaLiLeO aimed to provide an alternative entry point to the corpus of Galileian materials for posing contextualized questions about individual documents or authors through computation.

This article will evaluate the extent to which the digital and computational tools in GaLiLeO point to small-scale interpretive pathways through a corpus (the realm of non-digital humanists) rather than declare a quantitative definition of an aspect of a corpus for interpretation (which is often the aim of DH tools, methods, and research questions).

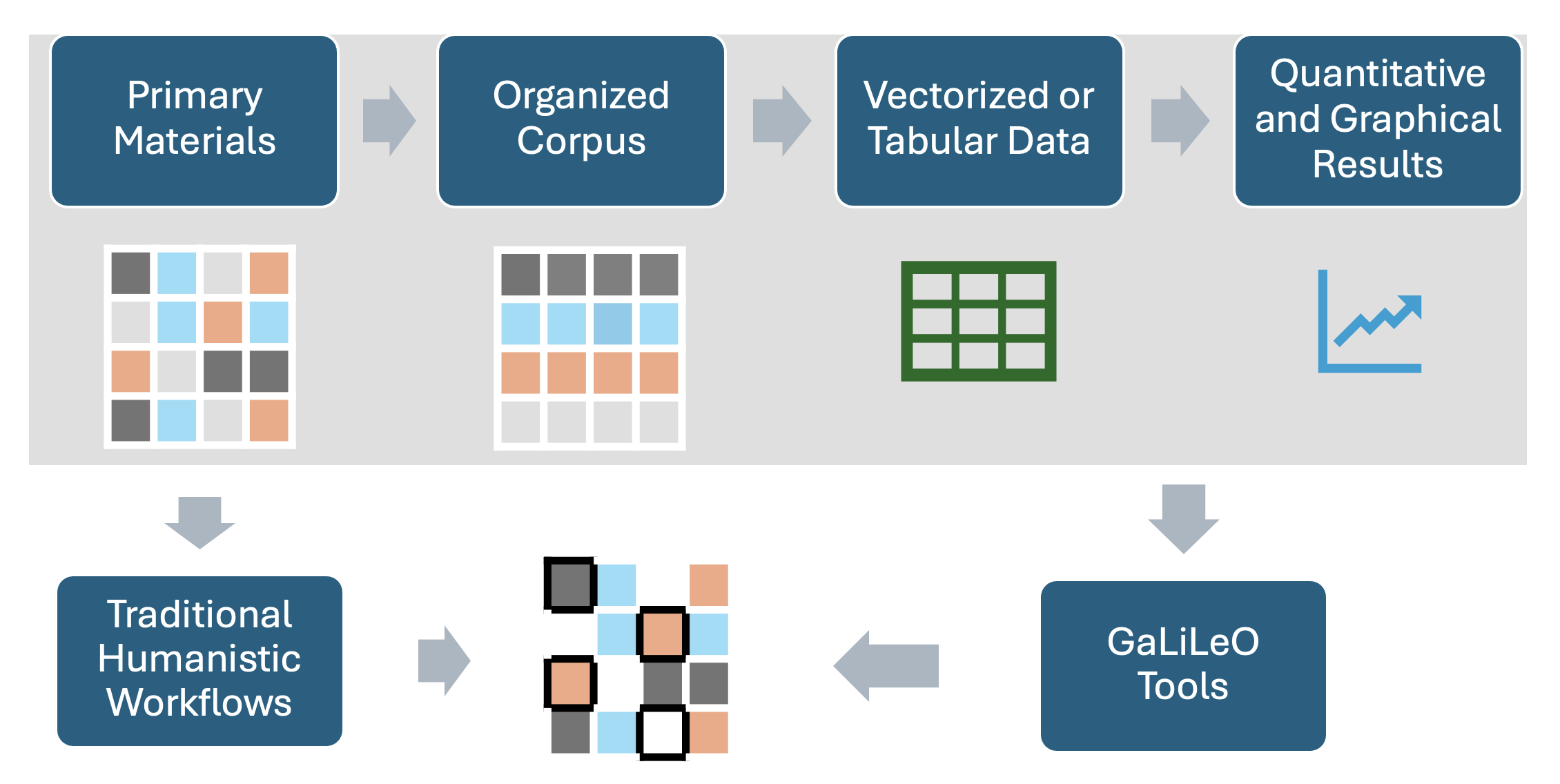
## Method for Tool Development

The tools were developed most intensively in 2016-2018, at a particular moment in the expansion and seemingly perpetual redefinition of the field of Digital Humanities, including digital history and computational literary studies. On the one hand, communities of stylometric experts were working to develop more accurate tools for classification (; ; ) and topic modeling was being applied to economic history () and the history of biology (). On the other hand, these were also the years that culminated in Nan Z. Da's polemical critique of methods and reproducibility in the field (), Katherine Bode's call for an articulation of the constructedness of corpora (), and greater visibility for the gaps, silences, and situatedness of archives and methods for their computational study. (; ) Taken together, this meant developing a suite of tools that incorporated features from both sets of trends in the field in pursuit of small-scale inquiry in a large corpus.

Non-digital historical and literary methods often involve making a selection from heterogeneous sets of primary materials that have been organized by their authors, archives, libraries, and editors. (See the left side of Figure 1, below.) That selection can be based on biographical (chronological), thematic, or lexical similarities or differences, but the features of what drive that selection were of interest for tool development in GaLiLeO. The choice of terms, themes, or historic moments can be driven by pre-established notable names, dates, and concepts for which existing tools like the Galileoteca are well suited. Yet, new discoveries also arise from a specialist noticing something unusual, witnessing variety, or recognizing an absence (among other possibilities). In a large corpus of thousands of documents, absence and anomaly are much harder to detect than in one main text or dozens of letters. This is where computational methods can assist, although many of the existing tools during GaLiLeO's development had other aims.

The gray box in Figure 1 offers an overview of a typical workflow for computational study of a textual corpus. While sharing a similar starting point with other humanistic inquiry, the data carpentry and digital humanistic processes begin with potential outputs in mind, which shape how the texts are organized and what data is created from them. Coding textbooks for humanists model this approach. (; ; ) A research question such as "How did early modern intellectual academies engage with the telescope in letters to Galileo?" could then prompt organization by the academy in which the author participated, creation of bag-of-words tables of word counts, and then stylometric modeling (among other ways to explore this idea). A parallel question that does not require digital methods for interpretation would be "What significance can be derived from an individual's departure from their academy's style of engagement with the telescope in letters to Galileo?" Again, specialist knowledge could determine the person and the academy due to previously-established notoriety, but that forecloses opportunities to engage with the lesser-studied or unstudied documents in the corpus. The scale is nonetheless overwhelming: With dozens of academies and hundreds of authors to choose from, where might a scholar begin?

display\_png(file="./media/GaLiLeOContextDiagram.png")



By creating a corpus-level view on the texts, that recognizes the imperfections of the corpus, GaLiLeO seeks a return to the texts to explore new interpretive pathways that have not been predetermined by centuries of scholarship. Granted, the underlying computational methods employed for the corpus-scale analysis in GaLiLeO have been modified from those used for larger scale study of similar materials to provide authorship attribution, thematic mapping, and rhetorical similarity. () Given the condition of the corpus, described in more detail in Section 1.3, such study was not an option. It was also not the immediate goal, since the team wanted to spark a conversation about tools that would be of use beyond the few digital humanists that also study Galileo. While the code could now be updated to reflect advances in text mining and computational linguistics, they do not address GaLiLeO's primary question of how to bridge the scales of question making and answering in the humanities.

This focus meant creating textual models and query possibilities that would highlight similarities or differences in a way that gives non-digital researchers enough context to both evaluate and follow leads suggested by those circumstances allowed by the corpus. In this sense, GaLiLeO aspires to going one step beyond the ePistolarium project () on 17th century correspondence in the Dutch Republic, which provides spatial and network visualizations of the searched person's locations and relationships. The openness of the planned, ideal query can also address DH priorities of de-colonizing archives. In order to know what is interesting enough to search in a corpus, a user needs to arrive with prior knowledge of the values associated with its contents. While not always problematic, systemic exclusion, majority perspectives, and other kinds of bias limit the interpretive pathways of a collection of text structured by technology. In that regard, GaLiLeO addressed a point raised by Ryan Cordell during a visit to Bowdoin College in 2014: What happens when we do not know the person, term, or year that might be interesting in a corpus of texts?

Consequently, rather than a model of the documents related to Galileo, the tools presented here provide a model for asking questions in non-digital humanistic fashion. The approach borrows this phrasing from Willard McCarty, who distinguishes between an archeological map *of* a historic site (i.e. a computational reading of the National Edition of Galileo's works) and the blueprint *for* a new edifice (inquiries related to the documents). () A model *of* Galilean texts would be forever incomplete and many parts already exist in the current digital ecosystem. A model *for* inquiry on the other hand would be extensible to the new data sets, for Galileo, but potentially for other authors and collections. The tools in GaLiLeO thus embrace the epistemology of *Hermeneutica*, but not the implementation. Since, as Rockwell and Sinclair point out, an "instrument implements a theory of interpreting the phenomenon it was designed to bring into view" (, 162), a new instrument was necssary. The historians involved in testing the prototype were seeking mechanisms to understand tone, reliability (through consistency), corroboration with other viewpoints, and verification. Literary scholars wanted to engage with the multiple contexts through which an expression conveys potentially different meanings. This required bringing together not just words and documents, but the overall corpus attributes. Again, as Rockwell and Sinclair so efficiently state: "developing tools involves bearing hermeneutical theories." (, 161) For the GaLiLeO designers, the result was a type of digital laboratory: a series of tools for exploration, question-making, and question answering.

One might ask why the team did not repurpose Rockwell and Sinclair's Voyant Tools for this purpose, since it was one of the primary inspirations for the functions in GaLiLeO. () The evaluations in subsequent sections of this article will provide a more detailed comparison, but one challenge was simply the size of the corpus, thousands of documents, which are difficult to see graphically.

The remaining sections of this article evaluate the tools as they were demonstrated via Jupyter notebook in September 2018, with identification of improvements that could be made in the future and slight modifications to function with the tools of this journal. While not necessarily an obstacle, the code ought to be optimized for runtime efficiency, since their designer was a Digital Humanist who codes, but not a software engineer. In addition, although the Jupyter notebook tried to offer novice coders instructions for use, the project would benefit from a more user-friendly interface.

## Preliminary Corpus for Data Development

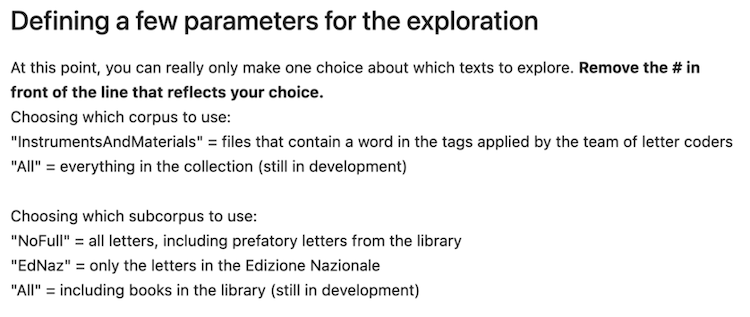
The two projects that merged involved metadata about Galileo's correspondence and cleaned OCR of books from his library, including both those that he wrote and others that he was known to have owned. The prepared corpus reflected Galileo's correspondence tagged by the Marcus team and the contents of his library prepared by the author's team as of July 2017. Metadata for these texts was hand keyed, with the clarification that sometimes authors (including letter writers) are simply unknown. Yet, the correspondence itself needed to be available for use as plain text. Even for an individual as famous as Galileo (or perhaps because of it), there is no freely-available set of plain-text files for analysis.

The correspondence in the National Edition of Galileo’s works included over 4,300 letters from the 20 volumes edited by Favaro in the 19th century (). Many letters were only excerpted in Favaro's edition. While a scan of the first printing of the National Edition is available, images of manuscripts are housed by different institutions with strict reproduction rules. Although materials existed on the Galileoteca site, and conversations with the director after 2018 indicated that there might be opportunities for bulk export and resource sharing, the process of working from the available print edition raised valuable questions about the state of Galilean correspondence. By working with the National Edition rather than manuscripts, the resources were not limited by the confines of physical archives, yet the editorial practices and remediation of the National Edition in the Galileoteca imposed critical changes in orthography, punctuation, and sometimes word choice. (, X, 11) These editorial layers foreclosed a corpus-scale analysis, which would have amounted to a study of a 19th-century version of the language of Galileo and his contemporaries. An ideal corpus would not remediate post-hoc editorial lenses with their own epistemological priorities. The GaLiLeO tools aspired to identify the categories, organizing principles, affinities, disparities, patterns, and outliers in the documents as they were expressed at the time. Even if the texts were readily available, they nonetheless overrepresent Galileo and the authors skew heavily male. This fundamental resource in the field was used to start the conversation about what might be built next, but should not be the foundation for that new digital corpus.

For the purposes of the prototype development, and with the limitations of time and resources, the corpus was nonetheless the only one available to use during the demonstration alongside the library contents and editorial metadata. Abbyy Fine Reader was used to convert the scans of the national edition to plain text with minimum data cleaning. The texts represented a spectrum of quality from a first pass of OCR on scans of 19th-century volumes (National Edition) to OCR results cleaned by comparison to scans of pre-modern printed volumes (library contents) to documents typed by hand based on the original print or manuscript (library contents). Using the editorial layer created by the Marcus team, which had tagged letters for use of various concepts after reading the letters, the corpus of correspondence was reduced to the documents with "Instruments and Materials" tags. These represented a range of early modern tools for knowledge creation including the telescope and its component pieces, but also candles, compasses, magnets, scales, and vessels for water. The team of specialist readers had also tagged words that were used as descriptors and actions related to experiments and experiences of natural phenomena in service of learning or confirming knowledge. Overall this created a subcorpus of 2,346 letters for analysis during the workshop. The proof of concept identified several areas for future development in terms of customization: how to interact with the specialist editorial layer (or not), choosing a subcorpus of interest, limiting or expanding the number of languages considered, and how to handle Galileo's work.

As presented at the workshop, the overall corpus could be subset into specialized document groupings based on type of document and metadata tag categories. (Figure 3)

display\_png(file="./media/GaLiLeOParameters.png")



Any text from the library that used one of 905 different terms tagged in the "Instruments and Materials" category identified by the editors was also available as a subcorpus. Users could opt to include the matching prefatory letters from 119 titles in Galileo's library. Due to an OCR error, the shortest document only had 1 token. The median length of the texts in the subcorpus was 257 words (i.e. 2,178 documents were shorter than this). The corpus represents 9 different languages. A planned feature would allow further subsetting by English, Dutch, French, German, Greek, Italian, Neolatin, Paduan dialect, and Spanish. This would require a team of scholars trained in those languages to help align the tags and clean the OCR.

The first code cells in the prototype made visible the goals and limited implementation of the prototype while obscuring significant interpretive work. Users were asked to run some setup code behind the scenes before getting started. This loaded the necessary libraries and read the data into memory without needing to show lines of code that might have been overwhelming to many of the workshop participants who had never worked with Jupyter or seen code prior to joining us that day. (The code cells in this article retain some of the commenting that was provided for colleagues who had never programmed prior to the workshop.)

#Requires ~20 seconds to load everything into memory  
source("script/5\_GaLiLeOSetup1.R")  
  
corpus\_name <- "InstrumentsAndMaterials" #documents that were tagged with this label by Marcus' editorial team  
  
subcorpus\_name <- "NoFull" #includes prefatory letters, but not full text of books  
#subcorpus\_name <- "EdNaz" #only OCR from the National Edition of Galileo's works  
  
# Do not edit this line or the following code:  
correct\_corpus\_files <- prepare(corpus\_name, "script/data/")  
cm <- grep("CorpusMetadata", data\_files[correct\_corpus\_files])  
correct\_cm <- grep(subcorpus\_name, data\_files[correct\_corpus\_files[cm]])  
corpus <- readRDS(file = paste(in\_dir, data\_files[correct\_corpus\_files[cm[correct\_cm]]], sep = ""))  
ca <- grep("CorpusAttributes", data\_files[correct\_corpus\_files])  
correct\_ca <- grep(subcorpus\_name, data\_files[correct\_corpus\_files[ca]])  
corpus\_attributes <- readRDS(file = paste(in\_dir, data\_files[correct\_corpus\_files[ca[correct\_ca]]], sep = ""))  
text\_a <- grep("TextAttributes", data\_files[correct\_corpus\_files])  
correct\_text\_a <- grep(subcorpus\_name, data\_files[correct\_corpus\_files[text\_a]])  
text\_attributes <- readRDS(file = paste(in\_dir, data\_files[correct\_corpus\_files[text\_a[correct\_text\_a]]], sep = ""))  
type\_a <- grep("TypeAttributes", data\_files[correct\_corpus\_files])  
correct\_type\_a <- grep(subcorpus\_name, data\_files[correct\_corpus\_files[type\_a]])  
type\_attributes <- readRDS(file = paste(in\_dir, data\_files[correct\_corpus\_files[type\_a[correct\_type\_a]]], sep = ""))  
source("script/6\_GaLiLeOSetup2.R")

During the demonstration, users were able to interact with documents written by over 625 authors. As with the national edition overall, the letter writers skew male and the collection features Galileo more than other individual. 307 authors only contributed one item. The top contributors were:

* Galileo Galilei (441)
* Benedetto Castelli (251)
* Fulgenzio Micanzio (153)
* Federico Cesi (150)
* Buonaventura Cavalieri (134)
* Suor Maria Celeste Galilei (124)
* Giovanfrancesco Sagredo (102)

Importantly, the subcorpus shows the prominence of Galileo's students and collaborators (Castelli, Cavalieri, and Sagredo), patron and partner (Cesi), and daughter (Maria Celeste). A planned feature will allow for continued subsetting by author of the material to facilitate the comparison of their linguistic choices to other writers. NB: anyone who explores the current notebook in more detail will immediately notice numeric discrepancies related to trimming the data to conform to file size limits in GitHub and mybinder. One can also see encoding issues related to accented vowels and special characters in the names of letter writers and recipients.

authors(corpus) # Long output

AuthorName NumberOfDocuments  
1 Adami, Tobia 1  
2 Adimari, Alessandro 2  
3 Aggiunti, Niccolò 20  
4 Agucchi, Gio. Battista 10  
5 Aguilon, François 1  
6 Alamanni, Luigi 1  
7 Albergotti, Ulisse 2  
8 Alberto Cesare Galilei; Giacinto Cornacchioli 1  
9 Aleotti, Giovanni Battista\* 1  
10 Allegri, Alessandro 1  
11 Altemps, Gio. Angelo 1  
12 Altobelli, Ilario 4  
13 Ammannati Galilei, Giuilia 1  
14 Ammannati Galilei, Giulia 3  
15 Antonini, Alfonso 2  
16 Antonini, Alfonzo 1  
17 Antonini, Daniello 9  
18 Aproino, Paolo 6  
19 Arrighetti, Andrea 10  
20 Arrighetti, Niccolò 3  
21 Aversa, Raffaele 1  
22 Baitelli, Lodovico 2  
23 Baldelli, Francesco 1  
24 Baliani, Gio. Battista 13  
25 Bandini, Ottavio 1  
26 Barberini, Francesco 7  
27 Barberini, Maffeo 3  
28 Barbolani da Montauto, Asdrubale 1  
29 Bardi, Ferdinando 2  
30 Bardi, Giovanni 3  
31 Bardi, Girolamo 6  
32 Bartoli, Giovanni 9  
33 Bartolini, Giovanni 1  
34 Bartoluzzi, Giovanni 1  
35 Bassone, Sebastiano 1  
36 Bedini, Alessandro 1  
37 Bellarmino, Roberto 4  
38 Belloni, Giovanni 1  
39 Benedetti, Giovanni Battista 2  
40 Benivieni, Girolamo 1  
41 Bentivogli, Antioco 2  
42 Bernegger, Matthias 23  
43 Bettoli, Guido 2  
44 Betussi, Giuseppe 1  
45 Biancani, Giuseppe 2  
46 Bini, Lorenzo 1  
47 Bocchineri, Alessandro 2  
48 Bocchineri, Carlo 2  
49 Bocchineri, Geri 25  
50 Bocchineri, Geri and Alessandro 1  
51 Boiardi, Paolo Emilio 1  
52 Bolognetti, Giorgio 1  
53 Bombini, Paolo 1  
54 Bonhomi, Sebastiano 1  
55 Bonsi, Domenico 1  
56 Borghese, Scipione 1  
57 Borghi, Pier Battista 13  
58 Borrhaus, Martin 1  
59 Borro, Girolamo 1  
60 Borsacchi, Camillo 1  
61 Botti, Matteo 4  
62 Bottigella, Scipione 1  
63 Bouchard, Gio. Giacomo 3  
64 Brahe, Tycho 1  
65 Brenzoni, Ottavio 5  
66 Bruano, Giovanni 1  
67 Buonamici, Gio. Francesco 4  
68 Buonarroti, Michelangelo 3  
69 Caccini, Matteo 3  
70 Calamai, Giovanni Battista 1  
71 Calestani, Girolamo 1  
72 Campanella, Tommaso 10  
73 Capra, Baldassare 1  
74 Cardi da Cigoli, Lodovico 14  
75 Casati, Gio. Paolo 1  
76 Castelli, Benedetto 155  
77 Castelli, Onofrio 1  
78 Castelli, Ottaviano 1  
79 Cavalcanti, Orazio 1  
80 Cavalcantini, Guglielmo 1  
81 Cavalieri, Bonaventura 90  
82 Ceccarelli, Lorenzo 5  
83 Cenati, Bernardino 1  
84 Cesarini, Virginio 10  
85 Cesi, Federico 85  
86 Chiaramonti, Scipione 4  
87 Ciampoli, Giovanni 38  
88 Cicognini, Iacopo 1  
89 Cini, Niccolò 3  
90 Cioli, Andrea 17  
91 Ciotti, Gio. Battista 1  
92 Cittadini, Domenico 1  
93 Cittadini, Paolo Maria 1  
94 Clavius, Christopher 2  
95 Clemente VII 1  
96 Coccapani, Ghismondo 2  
97 Colonna, Fabio 14  
98 Contarini, Angelo 1  
99 Contarini, Francesco 2  
100 Contarini, Giacomo 1  
101 Conti, Bernardo 1  
102 Conti, Carlo 2  
103 Conti, Conte 2  
104 Conti, Francesco 1  
105 Cornarius, Ianus 1  
106 Cornaro, Giacomo Alvise 2  
107 Costa, Margherita 2  
108 Costanzi, Giuseppe 1  
109 Cozzolani, Gio. Giacomo 1  
110 Cremonini, Cesare 1  
111 Crivelli, Francesco 1  
112 d'Acquaviva, Giuseppe 1  
113 d'Austria, Leopold 2  
114 d'Elci, Orso 6  
115 d'Este, Alessandro 1  
116 d'Oriolo, Lelio 1  
117 da Sommaia, Girolamo 1  
118 dal Pozzo, Cassiano 2  
119 de Carcavy, Pietro 4  
120 de Ville, Antonio 2  
121 de' Medici, Antonio 2  
122 de' Medici, Cosimo 2  
123 de' Medici, Cosimo II 5  
124 de' Medici, Ferdinando II, Granduca di Toscana 1  
125 de' Medici, Francesco 2  
126 de' Medici, Giuliano 7  
127 de' Medici, Leopoldo 2  
128 de' Ricci, Pierfrancesco 1  
129 de' Ronchitti, Cecco 1  
130 de' Rossi, Bastiano 2  
131 Dee, John 1  
132 Degli Albizzi, Luca 1  
133 del Borgo, Esaù 6  
134 del Monte, Francesco Maria 9  
135 Del Monte, Guidobaldo 4  
136 del Monte, Orazio 1  
137 del Ricco, Giovanni 1  
138 della Porta, Gio. Battista 3  
139 delle Colombe, Lodovico 6  
140 Demisiani, Giovanni 1  
141 di Castro, Pietro 1  
142 di Cologna, Sigismondo 1  
143 di Grazia, Vincenzo 2  
144 di Guevara, Giovanni 1  
145 di Joyeuse, Francesco 2  
146 di Noailles, Francesco 1  
147 Diedo, Girolamo 1  
148 Dini, Piero 4  
149 Diodati, Elia 22  
150 Dionigi, Francesco 1  
151 Dolce, Lodovico 2  
152 Donati, Paolo 1  
153 Doni, Gio. Battista 3  
154 Duodo, Francesco 9  
155 Duodo, Pietro 2  
156 Elzevier, Lodovico 4  
157 Engelcke, Beniamino 1  
158 Ernesto, Elettore di Colonia 1  
159 Faber, Johann 12  
160 Fabri di Peiresc, Niccolò 7  
161 Failla, Pietro Iacopo 1  
162 Fancelli, Chiarissimo 1  
163 Farnese, Odoardo 4  
164 Ferrari, Cristoph 1  
165 Figliucci, Flaminio 1  
166 Filippo IV, Re di Spagna 1  
167 Fioravanti, Leonardo 2  
168 Fiorentini, Francesco Maria 1  
169 Fontana, Domenico 1  
170 Fontanella, Alessandro 1  
171 Foscarini, Paolo Antonio 2  
172 Frambotto, Pauolo 1  
173 Francesco Stelluti; Federico Cesi 1  
174 Franciosini, Lorenzo 1  
175 Gaio, Benardino 2  
176 Galilei, Alberto Cesare 4  
177 Galilei, Benedetto 1  
178 Galilei, Galileo 310  
179 Galilei, Livia 1  
180 Galilei, Maria Celeste 64  
181 Galilei, Michelangelo 12  
182 Galilei, Roberto 22  
183 Galilei, Vincenzio 3  
184 Galilei, Vincenzo 4  
185 Gallanzoni, Gallanzone 1  
186 Gassendi, Pietro 8  
187 Gerini, Giulio 1  
188 Germini, Camillo 1  
189 Gessi, Belinghiero 1  
190 Ghetaldi, Marino 1  
191 Giannini, Tommaso 1  
192 Giggi, Antonio 1  
193 Giraldi Cinzio, Giovanbattista 2  
194 Giraldi, Roberto 1  
195 Giugni, Vincenzo 5  
196 Giunti, Cosimo 1  
197 Glissenti, Fabio 1  
198 Gloriosi, Gio. Camillo 3  
199 Gondi, Gio. Battista 6  
200 Gonzaga, Ferdinando 1  
201 Grassi, Orazio 1  
202 Gregory XIII, Pope 1  
203 Grienberger, Christoph 1  
204 Grini, Domenico 1  
205 Groto, Luigi 4  
206 Grozio, Ugo 1  
207 Guadagni Salviati, Ortensia 1  
208 Gualdo, Paolo 15  
209 Gualterotti, Raffaello 8  
210 Guerrini, Benedetto 2  
211 Guevara, Giovanni 1  
212 Guicciardini, Piero 3  
213 Guiducci, Mario 35  
214 Guzzaroni, Gio. Battista 1  
215 Hasdale, Martin 9  
216 Holste, Luca 1  
217 I Matematici del Collegio Romano 1  
218 Imperiali, Bartolomeo 9  
219 Incontri, Lodovico 2  
220 Jauffred, Giacomo 2  
221 Kepler, Johannes 6  
222 Kepler, Lodovico 1  
223 Labia, Andrea 1  
224 Lagalla, Giulio Cesare 4  
225 Landi, Federico 1  
226 Landi, Lorenzo 2  
227 Landini, Gio. Battista 2  
228 Landini, Silvestro 1  
229 Landucci, Virginia 1  
230 Liceti, Fortunio 21  
231 Lingelsheim, Gio. Michele 2  
232 Lodi, Giacinto 1  
233 Lodovici, Lodovico 1  
234 Lorenzini, Antonio 1  
235 Lorii, Antonio 1  
236 Lorini, Niccolò 1  
237 Maculano, Vincenzo 1  
238 Magagnati, Girolamo 6  
239 Magalotti, Filippo 2  
240 Magalotti, Lorenzo 1  
241 Magazzini, Vitale 1  
242 Magini, Gio. Antonio 16  
243 Magiotti, Lattanzio 1  
244 Magiotti, Raffaello 19  
245 Malacreta, Gio. Pietro 1  
246 Malaspina, Pietro Francesco 2  
247 Mannucci, Filippo 2  
248 Manso, Gio. Battista 2  
249 Maraffi, Luigi 1  
250 Marci, Gio. Marco 1  
251 Marsili, Alessandro 3  
252 Marsili, Cesare 17  
253 Mattei, Girolamo 1  
254 Maximillian I 1  
255 Mercuriale, Girolamo 3  
256 Mermanni, Tommaso 1  
257 Mersenne, Marino 1  
258 Micanzio, Fulgenzio 117  
259 Michelini, Famiano 9  
260 Millini, Benedetto 2  
261 Minucci, Andrea 1  
262 Mirabella, Vincenzo 2  
263 Morandi, Orazio 1  
264 Morin, Gio. Battista 2  
265 Morosini, Andrea 2  
266 Muti, Carlo 3  
267 Muzzarelli, Giovanni 2  
268 Naldi, Matteo 1  
269 Nardi, Antonio 8  
270 Nardi, Baldassarre 2  
271 Neri, Giuseppe 1  
272 Nerli, Maria Felice 1  
273 Niccolini, Francesco 18  
274 Niccolò Aggiunti; Dino Peri 1  
275 Nicolas-Claude Fabri de Peiresc; Pietro Gassendi 1  
276 Ninci, Alessandro 9  
277 Noghera, Vincenzo 1  
278 Not indicated by Favaro 1  
279 Oddi, Muzio 1  
280 Orsini, Alessandro 1  
281 Orsini, Franciotto 1  
282 Orsini, Paolo Giordano 3  
283 Ortensio, Martino 1  
284 Pannocchieschi d'Elci, Arturo 4  
285 Paolo Gualdo; Lorenzo Pignoria 1  
286 Paradiso, Romolo 1  
287 Passignani, Domenico 2  
288 Passionei, Don Francesco 1  
289 Pecci, Francesco 2  
290 Peregrino, Filenio 1  
291 Peri, Dino 20  
292 Perozze, Stefano 1  
293 Perugino, Innocenzo 2  
294 Petrangeli, Lorenzo 2  
295 Picchena, Curzio 22  
296 Piccolomini Aragona, Enea 4  
297 Piccolomini, Ascanio 16  
298 Piccolomini, Francesco 1  
299 Pieralli, Marcantonio 8  
300 Pieroni, Giovanni 13  
301 Pierucci, Gio. Michele 8  
302 Pignoria, Lorenzo 15  
303 Pinelli, Cosimo 1  
304 Pinelli, Francesco 1  
305 Pinelli, Giovan Vincenzo 2  
306 Piochi, Cristofano 1  
307 Pisani, Ottavio 5  
308 Porro, Gio. Giacomo 4  
309 Porta, Malatesta 2  
310 Pozzobonelli, Paolo 3  
311 Pusterla, Gio. Battista 1  
312 Querengo, Antonio 2  
313 Ramponi, Gio. Lodovico 4  
314 Rasi, Francesco 1  
315 Realio, Lorenzo 2  
316 Reijusk, Giovanni 1  
317 Remo, Giovanni 2  
318 Renieri, Vincenzo 35  
319 Reynier, Giulio 1  
320 Ricasoli Baroni, Giovanbattista 1  
321 Riccardi Niccolini, Caterina 4  
322 Riccoboni, Antonio 1  
323 Rinuccini, Carlo 2  
324 Rinuccini, Francesco 24  
325 Rinuccini, Gio. Battista 2  
326 Rinuccini, Pier Francesco 1  
327 Rinuccini, Tommaso 3  
328 Riques, David 1  
329 Risenti, Bernardo 1  
330 Risneri, Federico 2  
331 Robinson, Arrigo 1  
332 Roffeni, Gio. Antonio 4  
333 Rota, Angelo 1  
334 Ruscelli, Girolamo 3  
335 Ruschi, Gio. Battista 2  
336 Sacchetti, Niccolò 1  
337 Sagredo, Giovanfrancesco 90  
338 Sagredo, Zaccaria 7  
339 Salviati, Filippo 4  
340 Sandelli, Martino 4  
341 Santini, Antonio 24  
342 Santorio, Santorre 1  
343 Saracinelli, Cipriano 2  
344 Saracinelli, Ferdinando 1  
345 Saracini, Gherardo 2  
346 Sarpi, Paolo 2  
347 Sarrocchi, Margherita 4  
348 Sassetti, Cosimo 1  
349 Scalandroni, Benedetto 2  
350 Scalandroni, Sebastiano 1  
351 Scheiner, Christoph 1  
352 Schickhardt, Guglielmo 2  
353 Schroeter, Bartholomaeus 1  
354 Seggetti, Tommaso 1  
355 Selvatico, Girolamo 1  
356 Serafini, Orazio 1  
357 Sertini, Alessandro 7  
358 Settala, Lodovico 1  
359 Settimi, Clemente 1  
360 Sfondrati, Paolo 1  
361 Shorer, Emanuele 1  
362 Silvi, Giovanni 2  
363 Sizzi, Francesco 2  
364 Soldani, Iacopo 2  
365 Spinelli, Girolamo 1  
366 Spinola, Daniele 7  
367 Spinola, Tiberio 3  
368 Staccoli, Raffaello 1  
369 Stecchini, Paolo 1  
370 Stelliola, Nicolò 3  
371 Stelluti, Francesco 27  
372 Stigliani, Tommaso 1  
373 Strozzi, Gio. Battista 1  
374 Strozzi, Roberto 1  
375 Tadino, Alessandro 1  
376 Taletone, Giovanni 1  
377 Tamburalli, Cornelio 1  
378 Tedaldi, Maria 4  
379 Tedaldi, Muzio 4  
380 Tedeschi, Leonardo 1  
381 Titius, Benedictus 1  
382 Tolomei, Gio. Francesco 5  
383 Torricelli, Evangelista 5  
384 Tremazzi, Filippo 1  
385 Turtorini, Ascanio 1  
386 Ubaldini, Roberto 1  
387 Uguccioni, Giovanni 1  
388 unsigned 4  
389 Usimbardi, Lorenzo 3  
390 Vaiani, Anna Maria 1  
391 Valerio, Luca 13  
392 Van Maelcote, Oddo 1  
393 van Weert, Francesco 1  
394 Vannuccini, Giovanni 2  
395 Vasa, Władysław IV 1  
396 Vialardi, Francesco Maria 1  
397 Vincenti, Giacomo 1  
398 Vinta, Belisario 14  
399 Vitale, Constantino 1  
400 Wells, Giovanni 1  
401 Welser, Mark 24  
402 Welser, Matthias 1  
403 Wiffeldich, Giusto 1  
404 Zbaraski, Krystof 2  
405 Ziletti, Giordan & fratelli 1

Because the goal of GaLiLeO was to assist in the selection of texts, the contextual data about their contents had to be developed with the possible subcorpus variations that were planned. The data creation process occurred prior to the user's corpus selection. (This also solves a runtime problem, similarly to the functionality of the Text Analysis and Concording Tool (TACT), but of course restricts usability by this predefinition. (, 57)) Currently the analysis focuses only on the body of documents. A feature to develop would include words in titles and the metadata as part of the contextual information available. Given the known editorial interventions, two versions of each subcorpus were made: one with and one without punctuation. Every character was converted to lowercase.

One of the primary challenges was accounting for orthographical variance between the critically edited texts and the diplomatically edited ones. These are unrelated to the significant orthographical changes made in the 19th century edition, and are instead an artifact of printing practices in the 17th century. Early modern printing practice interchanged u and v, such that, for example, vno and uno were read as equivalents by contemporary readers. Since the digital characters v and u are encoded uniquely, substitutions were made in all texts to standardize how words appeared, meaning a departure from their print original, but the ability to search for terms without concern for orthographic variation. Using regular expressions would have changed too many vs that really were vs, so this catalog of substitutions was developed iteratively. It identifies and corrects most v/u substitutions, but likely still needs refinement. On the other hand, a universal substitution was made for j (called the long i) to i. This would need to be addressed with more nuance in a subsequent version of the tool kit, but allowed for significant experimentation even with the alpha prototype for texts in Italian (as opposed to Latin where it is more frequent).

v\_incorrect <- c("vb", "vc", "vd", "vf", "vg", "vh", "vj", "vk", "vl", "vm", "vn", "vp", "vq", "vs", "vt",   
 "vw", "vx", "vy", "vz", "bv", "cv", "fv", "gv", "hv", "jv", "kv", "mv",  
 "pv", "qv", "tv", "wv", "xv", "yv", "zv", " sve ", " sva ", " svo ", "trvi", "nvo ",  
 "nva ", "dvr", "dvi", " piv ", " lvi ", "svoi", " dv", "vr ", "nvare", " Vn", "Qv", "æ",  
 "vre ", " lv", " svo.", " svo,", " svo?", " svo!", " svo;", "vrea", "svet", " lvi.",  
 " lvi,", " lui?", "svra") # v does not normally appear before these letters  
u\_correct <- c("ub", "uc", "ud", "uf", "ug", "uh", "uj", "uk", "ul", "um", "un", "up", "uq", "us", "ut",   
 "uw", "ux", "uy", "uz","bu", "cu", "fu", "gu", "hu", "ju", "ku", "mu",  
 "pu", "qu", "tu", "wu", "xu", "yu", "zu", " sue ", " sua ", " suo ", "trui", "nuo ",  
 "nua ", "dur", "dui", " piu ", " lui ", "suoi", " du", "ur ", "nuare", "Un", "Qu", "ae",   
 "ure ", " lu", " suo."," suo,", " suo?", " suo!", " suo;", "urea", "suet", " lui.",  
 " lui,", " lui?", "sura")

The code is not elegant, but captures quantitative information about each document, the words in the document, and the author. As shown in the figure below, the underlying metadata is also contextualized within the corpus. The results are stored as RDS objects to be loaded into memory when the user chooses a subcorpus for exploration.

display\_png(file="./media/GaLiLeOTextAttributes.png")



Figure 4 outlines the specific and relative information created and saved for each document. Again, there are redundancies and inefficiencies since this is an alpha-prototype coded by a humanist. Many of the attributes are recorded with variations: tokenization includes and excludes punctuation, stylistic similarity based on three different feature classes and two distance measures, as well as distinguishing between documents written by Galileo and those written by other authors. The goals were twofold: achieve as many perspectives as possible on the texts and word types as well as explore the strengths or limitations of each perspective in comparison to the others. Creating many of the features also allows for querying by attribute range rather than the specific attribute. For example, by saving the contextual information, a user could identify authors who wrote many documents in the corpus, but were the recipient of none.

As Rockwell and Sinclair point out, these hermeneutica draw our attention to "the materiality of interpretation." (, 152) For GaLiLeO the materialities rest on two kinds of instantiations of interpretation: providing context around specific documents or words and experimenting with pattern-based inquiry. The contextual output draws attention to the marginal, mainstream, and missing information related to a query. Data about commonalities and differences appear together by default. Pattern-based functions, currently at the end of the prototype, push the materiality of interpretation further by asking the scholar to frame a question without knowing a specific document, word, or author that might meet the criteria.

Typical document-retrieval design would require the user to have one of two things:

* prior knowledge of key actors in the period, important dates, sites of intellectual exchange, and significant keywords (all likely brought to light by previous scholarship)
* enough curiosity to provoke browsing ordered lists of people, dates, places, or concordances in the hopes of finding something interesting

While our underlying data does capture information about people and places (with Named Entity Recognition to be implemented), it also tries to bring to light the bigger picture in which they appeared. GaLiLeO gives relevance to inconsistencies, prizes the anomalous, renews the priority of the individual text or author, draws attention to gaps, and capitalizes on juxtaposition to provoke new inquiry.

For example, in the "Instruments and Materials" documents without full text, the dates of coverage were 1506-1642 (the year of Galileo's death). With mean and median years in the mid-1620s, there is potentially an emphasis on discussions of experimentation supported by tools and instruments in the correspondence written in later decades of Galileo's life. Yet, in terms of chronological representation of the overall corpus, it is also heavily skewed toward the later years of Galileo's life, when his fame was such that correspondence increased and the need for preserving a record of communication with famous individuals became apparent.

date\_range(all\_years)

The earliest year is 1565 .  
The latest year is 1641 .  
Here is a numerical summary of the spread of dates covered:  
 Min. 1st Qu. Median Mean 3rd Qu. Max.   
 1565 1613 1629 1624 1635 1641

The following code cell shows a simplistic example of how this was instantiated in terms of creating both the model of the data, but also a model for contextualizing any document. The context, ideally, was provocation for further inquiry based on the results (the circumstances influencing the events). Users are immediately greeted with information about the years of coverage in their subcorpus, shown in the code snippet and then rendered specifically for one document of interest that will serve as the example for the rest of this article.

#Get year information  
years <- as.numeric()  
for (d in 1:length(text\_attributes)){  
 year <- as.numeric(text\_attributes[[d]]$Year)  
 years <- c(years, year)  
}  
all\_years <- sort(years, decreasing = F)  
years\_df <- as.data.frame(table(all\_years))  
years\_df$Color <- "No"  
colnames(years\_df) <- c("Year", "Count", "Color")  
  
#Partial representation of the function to highlight contextual data about years  
#Commented out to avoid overwriting the full function loaded earlier  
  
#choose\_doc <- function(text\_list, corpus\_metadata, full\_ID){  
# tl\_index <- which(Full\_IDs == full\_ID)  
# for (j in 1:length(tl\_index)){  
# cat("This is 1 of ", text\_list[[tl\_index[j]]]$AuthorDocs, " documents written by ", text\_list[[tl\_index[j]]]$Author, "in the corpus.\n")  
# cat("The title is: ", corpus\_metadata[[tl\_index[j]]]$Metadata$Title, "\n")  
# cat("\nIt was written or published ", text\_list[[tl\_index[j]]]$Year, ".\n")  
# cat(text\_list[[tl\_index[j]]]$YearInfo$SameYear, "document(s) was (were) dated or (published) in the same year.\n")  
# cat(text\_list[[tl\_index[j]]]$YearInfo$Before, "document(s) was (were) dated or (published) in the years before it.\n",   
# text\_list[[tl\_index[j]]]$YearInfo$After, "document(s) was (were) dated or (published) in years after it.\n")  
# target\_y <- text\_list[[tl\_index[j]]]$Year  
# years\_df$Color[which(years\_df$Year == target\_y)] <- "Yes"  
#   
# y<-ggplot(years\_df, aes(x=as.numeric(Year), y=Count, fill=factor(Color))) +  
# geom\_bar(stat="identity")+theme\_minimal() +  
# theme(axis.text.x= element\_text(angle = 45)) +  
# theme(legend.position = "none") +  
# scale\_x\_continuous(name = "Year", breaks = seq(earliest,latest, by = 5)) +  
# labs(title = "Comparison of Selected Text (in Blue) by Year")  
# out\_dir <- "GaLiLeOResults/"  
# year\_plot\_name <- paste(out\_dir, full\_ID, "YearComparison.png", sep = "")  
# ggsave(filename = year\_plot\_name, plot = y, device = "png")  
# dev.off()  
# }  
#}

display\_png(file="./media/EdNaz574YearComparison.png")

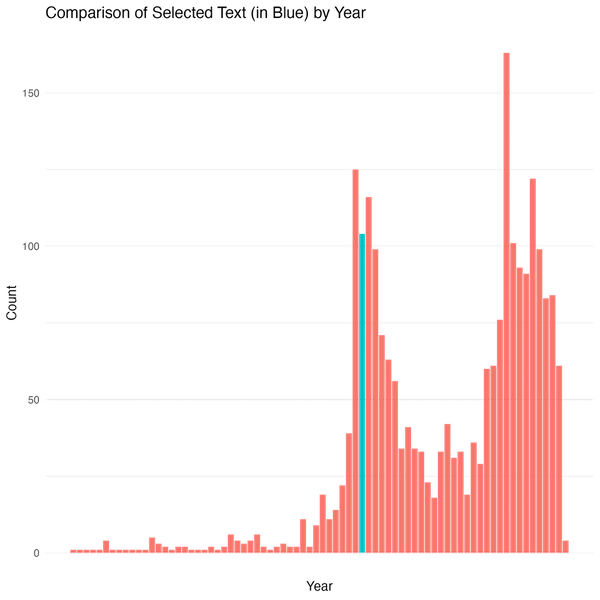


Figure 5 visually foregrounds the chronological context of a letter written in 1610, while the code does so numerically. The user can immediately see trends of representation in the corpus at large. A student new to the field can recognize that they are in a period of importance (determined by quantity of documents), while a scholar might be curious about the years with less documentation.

By creating this metadata for documents, GaLiLeO's tools facilitate structuring a query that identifies relationship intensities (central, peripheral or eccentric), frequencies (maximum, high, low, minimum, none), and document categories (private, public, print tradition in the library) of interest. Using this metadata, it would be helpful to both an expert scholar and a student to be able to search by the following attributes that can be calculated for each named person or place:

* how central the author is in Galileo’s library or letters
* that a letter recipient is only present in paratext in the library
* if a person mentioned has a higher centrality in the library than in correspondence
* that the destination of the letter is the most frequent city to which letters were sent and/or the second-most frequent city in which books in the library were published
* the date of the document is the year in which book collecting started to increase and/or a year with many missing letters, but still a high frequency for the overall correspondence
* a word has a high frequency in Galileo's letters, but very few or no occurences in printed materials, or is absent from other writers' documents entirely

These features use corpus-wide trends to point to small-scale patterns. In response to DH scholars Witmore and Hope, GaLiLeO aimed for tools that “allow you to have your books on multiple shelves simultaneously.” (, 18) Such a structure is an invitation, one that does not require knowing someone or something “interesting” before exploring. In this sense, the ideal tools would respond to a challenge that Ryan Cordell identified in the development of the large-scale digital text archives: even a motivated user often does not know what to put in a search bar. ()

The original plan for GaLiLeO anticipated accounting for the fact that any word in the letters or library is:

* a string of alphanumeric characters (ngram)
* a potential editorial tag provided by specialists
* a part of speech (POS)
* a keyword in multiple textual contexts
* a piece of a textual structure (sentence, verse, speaking part, marginalia)
* a piece of multiple semantic networks
* an abstraction (indicative of a dictionary of synonyms and related terms)
* a concept (a collection of instances across time and contexts)
* an expression in time
* a position in vectors of words
* part of metadata
* part of text or paratext
* a material object with a print and/or manuscript history and visual qualities
* an instance of a frequency of usage:
  + In the corpus
  + In works written by Galileo
  + In works written by others All of these features would thus build the query structure for the curious researcher that can formulate a question without already knowing a preliminary answer.

Scholars and students know what kinds of questions are interesting without necessarily knowing who or what fits those criteria, and the GaLiLeO team hopes to have built something to model (or provide a foundation for) that kind of query. In that sense, GaLiLeO becomes a site of discovery, not simply transmission of digitized documents. The hierarchies often embedded in digital tools through remediation of archives are brought onto the same plane via this underlying data structure. By changing identifiers related to documents (not alphabetical, chronological, or numerical), the texts are placed in an interconnected space of representation and understanding (thus they are actual relations rather than a relational table with connections assigned through inherited, post-hoc knowledge). The user can experience these texts with a perspective that is not her own - a browsable structure that provokes inquiry, allows for the selection of a virtual book from this network and reveals all of the ways that it nests within hierarchies or stands apart from them. Rather than browse by a post-hoc genre category, the adjacencies of texts based on a variety of features can be evaluated for their apparent lack of connection, the ambiguity of the language that binds them, and the serendipitous viewpoint that their similarity allows.

## Document Tools in GaLiLeO

The goal with the analytical tools for the text of the documents was to create a visual or numerical point of entry for asking questions about terminology, texts, and authors. While not as polished as the suite of widgets in Voyant Tools, the demonstration at Harvard in 2018 deployed 12 tools even though participants were unfamiliar with Jupyter notebooks and coding in R or other formal languages. Omitted here is a discussion of the Tag Exploration (developed by Hannah Marcus and Morgan MacLeod) that was part of a secondary Jupyter Notebook written in Python. Because the presentation was for a specialist audience of Italian Studies scholars, most of whom focus on Galileo, the pattern-based searches for an author or document were not fully built. These would have included addressing patterns based on an individual's frequency of appearance, role in a network of correspondents, co-mentions in a text (or lack thereof). Humanists are interested in central, peripheral, and eccentric figures; the popular and the obscure; those connected to figures of power (or not); those who cross boundaries or define them. All of these features could be calculated and queried (with the right data) to draw attention to people in the documents without ever needing to know a name to put in a search bar or select from a menu.

Ideally, users would be presented with a menu of attributes related to similarity and difference to help them arrive at a document of potential interest. For instance, such a pattern-based tool would surface the name of Margherita Sarrocchi (1560-1617), because she is mentioned more often by other authors than she is the author of surviving letters, a potentially interesting imbalance worthy of investigation. Given Sarrocchi's embeddedness in literary circles in Rome at the time, her correspondence with Galileo also offers a sense of the interdisciplinary nature of Galileo Studies. () For the sake of this demonstration, a letter that Sarrocchi wrote to the Perugian Guido Bettoli (active early-16th century), of which she sent a copy to Galileo, will serve as a guide through the document and lexical tools. Since Galileo read the letter (having written this contextual information on a surviving copy of it), and since it was not included in Ray's recent edition of the Sarrocchi-Galileo correspondence, it provides an alternative entry point to examining Sarrocchi's relationship with Galileo and his discoveries.

A function later in the prototype notebook, "read\_doc", prints the contents of the letter to the screen, but it is provided here for Italian readers to see the letter in question. The output also provides a sense of the quality of the OCR from the national edition. The letter is labeled EdNaz574 in the code cells for its location in the national edition.

read\_doc(text\_attributes, "EdNaz574")

IU.r0 Sigti mio Oss."10 Ho recivuto la di V.S. de di Giugno, che mi è parso un miracolo che me sii capitata così tardi, poiché ogni ordinario io mando alla posta del Papa, et la lettera non la recevetti prima di hieri; et però non se maravigli se ancora io tardi le rispondo. Gli ò vero che due mesi sono n’ hebb'i una da un frate, alla quale non resposi, per irritrovarmi in letto ammalata et perchè il Sigti Luca(2) scrisse a V. S. che per saper la mia opinione ne domandasse a Padre Innocentio del’ ordine di S.t0 Agostino, che sta costì in S.ta Maria Novella. Hora le dico a V. S., che tutto quello che se dice del ritrovamento delle stelle del Sigti Gallileo è vero, cioè che con Giove son quattro stelle erranti con moto proprio, sempre egualmente distante da Giove, ma non fra di loro; et io con li proprii occhi l’ho vedute mediante l’ ochiale del Sigti Gallileo, et fattele vedere a diversi amici: il che tutto il mondo il sa. Con Saturno sono due stelle, una da un lato et l’altra dal’ altro, che quasi lo, toccano. Venere, quando si congiunge col sole, si vede illuminare et diventar, come la lui?na, corniculata, infino a tanto che la si vede poi tutta piena; et mentre si va [e]m- piendo, appar minore, chiaro segno, anzi demostratione geometrica, che ella s’aggira intorno al sole; et quando è piena, gli è sopra, et per la gran distanza appar minore: questo, dico, si sa per demostratione geometrica, poi che non può apparir piena per oppo- sitione che habbia col sole. Molti matemateci grandi, et in particolare il P.re Claudio col P.re Gambergere W negavano questo da principio, et dipoi si sono disdetti, essendosene certificati, et ne hanno fatte publiche lettioni. Quanto poi che cotesti Signori dello Studio et Achademici non habbino scritto contra al Sigti Gallileo, io lo credo, et lo farò sapere al Sigti Gallileo; anzi gli mandato la lettera di V. S. In tanto V. S. gli assicuri che il Sigti Gallileo, oltre alla sublimità dello ingegno mirabile che ha, è di tanta buona conditione, che quando ancora eglino gli havessero scritto contra, s’acquetarebbe ad una minima loro scusa, essendo che egli non pretende altro che giovare al mondo; chè se fusse avido di haver fama, ne può haver molto maggiore da molte singolari compositioni che egli in diverse scientie ha fatto. Questo ò quanto mi occorre dire in risposta della sua domanda: del resto la rin- gratio del cortese affetto che ella dimostra verso di me, et delle lodi che, oltre al mio merito, mi dà; et così la prego a valersi di me in ogni sua occorrenza, chè mi troverà prontissima et grata alla sua buona voluntà. N. S. la guardi. Di Roma, a dì 27 di Agosto 1611.

Currently, the function “choose\_doc” is the primary contextualizer in the suite of tools and much more restricted than the pattern-based ideal described above. The function retrieves critical numeric and lexical information, presents those results as text, and saves images of graphs when appropriate. The output from choose\_doc is admittedly long, so it will be evaluated in phases. The visualizations that accompany the quantitative output are loaded here as images. Student users can refer to the output to prompt exploration of other documents and authors that are similar. Specialists will likely be drawn to the lexical and chronological features to start an investigatory path. All users will be presented with what could be called the document's perspective on the corpus: how it relates to other documents without imposition of siloes of periodization, nationality, genre, or form.

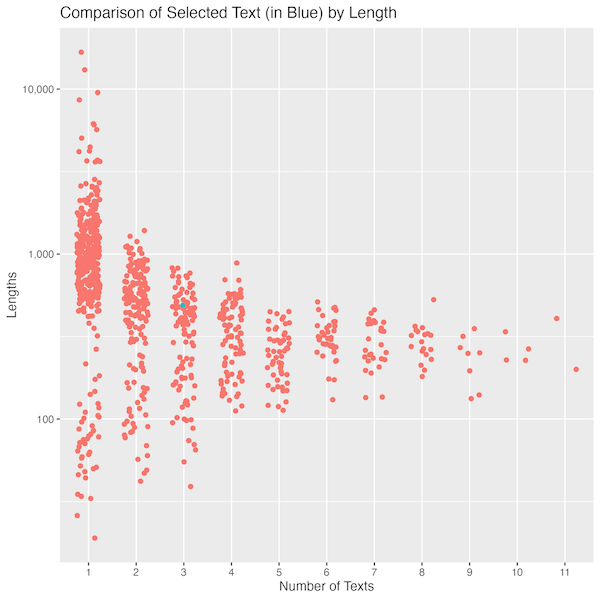
choose\_doc(text\_attributes, corpus, "EdNaz574")

This is 1 of 8 documents written by Sarrocchi, Margherita in the corpus.  
The title is: Carteggio   
  
It was written or published 1611 .  
104 document(s) was (were) dated (or published) in the same year.  
325 document(s) was (were) dated (or published) in the years before it.  
 1909 document(s) was (were) dated (or published) in years after it.  
  
The document has 489 tokens (unique strings of characters), not treating punctuation as a word.  
 659 document(s) is (are) longer.  
 1684 document(s) is (are) shorter.  
3 document(s) is (are) the same length.  
  
Here is some information about the type/token ratio, often a measure of linguistic complexity. A value closer to 1 indicates the frequent introduction of new words with little repetition.  
 The TTR (includes punctuation as a token) for the document is: 0.6160338 .  
 Documents in the same percentile in the corpus are: EdNaz2424 EdNaz4046 EdNaz1820 EdNaz3082 EdNaz3572 EdNaz3142 EdNaz526 EdNaz2649 EdNaz1770 EdNaz2995 EdNaz3324 EdNaz3355 EdNaz4040 EdNaz745 EdNaz68 EdNaz1360 EdNaz1028 EdNaz3206 EdNaz1192 EdNaz1197 151Letter1 675Letter1of1 999Letter2   
 The TTR (not including punctuation as a token) for the document is: 0.5439673 .  
 Documents in the same percentile in the corpus are: EdNaz2428 EdNaz3495 EdNaz1781 EdNaz2330 EdNaz896 EdNaz1476 EdNaz1008 EdNaz3850 EdNaz2497 EdNaz2547 EdNaz1805 EdNaz3292 EdNaz445 EdNaz291 EdNaz3008 EdNaz1421 EdNaz1359 EdNaz813 EdNaz563 EdNaz4008 1352Letter2 1033Letter2 1189Letter2   
  
The vocabulary in this document is notable for the following reasons:  
This is the only document in the corpus that uses the following words: achademici acquetarebbe aggira congiunge disdetti essendosene gambergere hebb irritrovarmi matemateci piendo recevetti resposi sitione .  
This document does not include 595 of the words that have the highest frequency (the words in the 99th percentile):  
To see these words (there may be hundreds of them), run the cell below.  
  
Patterns of use of high frequency words can identify similar styles using two measurements.  
 One measurement, Euclidean distance, describes direct correspondence of word use frequency:  
 The most similar document to EdNaz574 is EdNaz3104 . The least similar is EdNaz226 .  
The second type of measurement is Cosine similarity, which describes proportional correspondence of word use frequency:  
 The most similar document according to this measurement is EdNaz3104 . The least similar is EdNaz4180 .  
  
Topic modeling offers another way to see connections. This document is associated with the following topics in the corpus:  
Topic 1 73.15175 %: molto.e..piu..ill.ecc.mo.perche..p.ogni.tanto.ella.altro.galileo.qualche.cosa.mani.fatto.stato.pero..sigli.bene.poi.galilei.gia..sli   
Topic 2 16.34241 %: e..piu..sole.terra.cosi..parte.quali.sopra.altro.esser.moto.perche..poi.ne..tanto.quando.altra.essere.due.altri.puo..intorno.anco.si..solo   
Topic 3 7.782101 %: sti.sigti.eccti.ulti.serti.lincei.affti.ossti.cesi.luca.fed.apelle.cigoli.salviati.valerio.solari.line.velsero.machie.terza.margherita.ridolfi.avisi.peripatetici.demisiani

There are 8 documents written by Sarrocchi in the corpus overall, but only 4 appear in this subcorpus related to instruments and materials. That discrepancy should prompt a query for her other letters (and eventually the full text). She wrote to Bettoli in 1611, soon after Galileo had returned to Tuscany upon publishing his discoveries with the telescope. The barplot that is generated by the same function for contextualizing years (seen in Figure 5 above) visualizes that context in terms of quantities of letters that exist in the subcorpus. 1611 is in the top 5 years for which letters have survived in the various archives represented by the data. The chronological information indicates that Sarrocchi's enthusiastic confirmation of Galileo's observations was sent during a year of intensive correspondence. One might start to wonder how Sarrocchi became part of this first wave of interest in instruments and materials, particularly since she was not one of the earliest writers on these topics. This should also prompt an interative use of the code to contextualize her other letters chronologically. A convenient additional feature would be to plot all of the writer's documents against the backdrop of the corpus.

An embedded function within choose\_doc contextualizes the length of a document in comparison to the lengths of other documents in the subcorpus. Sarrocchi was writing longer letters than the average correspondent in this subcorpus. In addition to the prose readout from the code cell, prototype users saved a graph that plots lengths for the entire subcorpus.

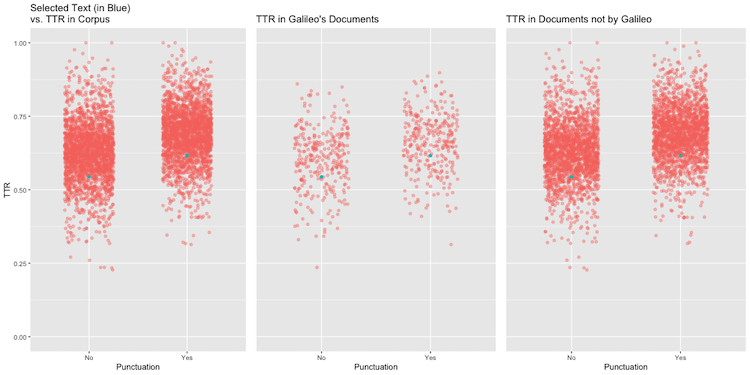
display\_png(file="./media/EdNaz574LengthComparison.png")



While perhaps initially disorienting, the plot in Figure 6 shows, for example, that 11 documents are ~250 words long (the dot farthest to the right). 11 are slightly longer than 500 (the dot to its left). At the left side of the graph, length variation increases. The y-axis is a logarithmic scale to account for the handful of documents that are around 10,000 words in length. The context shows how unusual in this subcorpus are documents of the same length. Red dots lower than Sarrocchi's blue dot are letters that are shorter, those higher are longer. It is not unusual for letters to have the same length. To assist with pattern-based inquiry, this function should be expanded to be able to reverse the query based on the location of a point on the plot. This would increase the juxtaposition circumstances for the texts and would help to identify Sarrocchi's variation in all of her letters (or not).

Beyond quantity of terms, the code also reveals proportionality of lexical choices. One of the visual outputs is a comparative type-token ratio scatterplot (Figure 7, below). This can be a way to understand repetition and complexity in documents by comparing unique word forms to the total number of words in a text. The figure compares the lexical complexity of a specific document (in blue), with or without punctuation, to works by Galileo and works by other authors. Importantly, the points with a TTR value of 1.0, meaning that no words are repeated in the document, are short excerpts from letters that were transcribed into the National Edition. They are sometimes so short as to be phrases or incomplete sentences. Sarrocchi's letter is not as lexically complex as most of the documents in the subcorpus. Here Sarrocchi's use of terms is seen in comparison to Galileo directly and against the backdrop of other authors in the subcorpus. The juxtaposition reveals more repetition in her letter than in those written by others.

display\_png(file="./media/EdNaz574TTRComparison.png")



Admittedly, the distinction between counting or not counting punctuation might be negligible. However, Galileo's prose is built on clauses and subclauses, so counting words followed by punctuation as a token helps to capture something about the prominence and variation within that stylistic element. For example, a document 1000 words long with 500 unique word types has a TTR of 50%. If the letter writer uses punctuation to establish rhythm and structure in her sentences, at a rate of one mark for every 5 words, the number of unique types could be as high as 700 (since *philosophy* and *philosophy,* would be treated as distinct strings of characters), while the length remains the same. Then the TTR would be 70%. If she has long sentences structured around connective words rather than punctuation, the number of types will only increase by a small percentage. Such measurements invite more opportunities to design pattern-based inquiry options.

The function “choose\_doc” also allows a user to see unique vocabulary and surprising omissions of words that are otherwise popular in the corpus. With Sarrocchi's letter, specialists will be interested in the output that reports in her apparently unique use of "achademici" and "matemateci", both of which are distinct spellings of the standard Italian terms academici (academicians) and matematici (mathematicians). Computational tools that rely on underlying use of dictionaries, language models, or lexica combine these variations into one result. In a corpus like the Galileo correspondence or his library, that would mean sifting through hundreds of results for "matematici" that might have instead used the variant.

Importantly, the function identifies what writers were *not* talking about alongside what they were. From what conversations were they excluded? What topics did they avoid even though they could have addressed them? The function "TopWordsNotInText" (so-named for transparency to novice coders) allows the user to probe the output from "choose\_doc" more deeply.

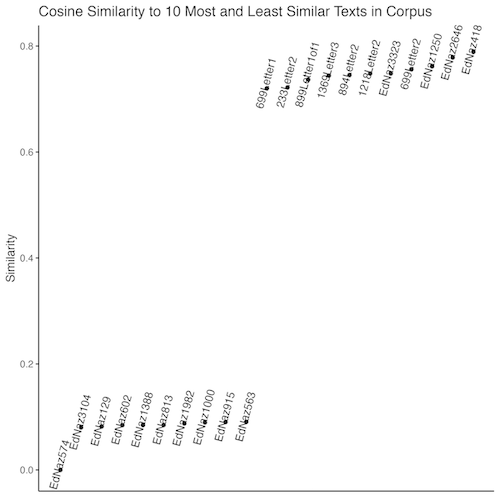
#Please put the Document ID in quotation marks to the left of this comment.  
ID <- "EdNaz574"  
# Call the TopWordsNotInText function and store the result in a dataframe  
result <- TopWordsNotInText(text\_attributes, ID)  
  
# Ensure the result is a dataframe  
result\_df <- as.data.frame(result)  
  
# Add an index column  
result\_df$Index <- seq.int(nrow(result\_df))  
  
# Reorder columns to have the index first  
result\_df <- result\_df[, c("Index", names(result\_df)[names(result\_df) != "Index"])]  
  
# Set the maximum number of rows to display to a large number  
options(repr.matrix.max.rows = nrow(result\_df))  
  
# Display the dataframe  
display(result\_df)

Index result   
1 1 più   
2 2 suo   
3 3 d   
4 4 sig   
5 5 o   
6 6 1   
7 7 r   
8 8 ill   
9 9 nel   
10 10 ecc   
11 11 questa   
12 12 sia   
13 13 mo   
14 14 dell   
15 15 0   
16 16 nella   
17 17 lei   
18 18 cosa   
19 19 quali   
20 20 qualche   
21 21 altri   
22 22 parte   
23 23 anco   
24 24 esser   
25 25 quella   
26 26 cose   
27 27 galileo   
28 28 nè   
29 29 quel   
30 30 bene   
31 31 qui   
32 32 tempo   
33 33 stato   
34 34 all   
35 35 già   
36 36 mani   
37 37 sì   
38 38 tutti   
39 39 galilei   
40 40 sarà   
41 41 senza   
42 42 ci   
43 43 sigli   
44 44 sli   
45 45 suoi   
46 46 poco   
47 47 essere   
48 48 ch   
49 49 far   
50 50 ser   
51 51 dalla   
52 52 à   
53 53 sue   
54 54 mano   
55 55 detto   
56 56 modo   
57 57 questi   
58 58 solo   
59 59 era   
60 60 dio   
61 61 fa   
62 62 eh   
63 63 ed   
64 64 vi   
65 65 fare   
66 66 so   
67 67 uno   
68 68 lettere   
69 69 ciò   
70 70 bacio   
71 71 nostro   
72 72 qua   
73 73 altre   
74 74 ben   
75 75 queste   
76 76 alle   
77 77 tutte   
78 78 quelle   
79 79 mai   
80 80 dove   
81 81 assai   
82 82 occasione   
83 83 3   
84 84 nelle   
85 85 nell   
86 86 libro   
87 87 tal   
88 88 qual   
89 89 giorni   
90 90 ut   
91 91 chi   
92 92 fine   
93 93 gl   
94 94 gusto   
95 95 ron   
96 96 terra   
97 97 tra   
98 98 signore   
99 99 insieme   
100 100 firenze   
101 101 fu   
102 102 possa   
103 103 mie   
104 104 noi   
105 105 circa   
106 106 quelli   
107 107 fvori   
108 108 desiderio   
109 109 alcuna   
110 110 onde   
111 111 pur   
112 112 medesimo   
113 113 posso   
114 114 c   
115 115 acciò   
116 116 volte   
117 117 molt   
118 118 1110   
119 119 casa   
120 120 nome   
121 121 tale   
122 122 havendo   
123 123 gratia   
124 124 est   
125 125 6   
126 126 punto   
127 127 opera   
128 128 illli   
129 129 anni   
130 130 fosse   
131 131 servitore   
132 132 4   
133 133 grande   
134 134 potrà   
135 135 pure   
136 136 pare   
137 137 presente   
138 138 maggior   
139 139 esso   
140 140 miei   
141 141 meno   
142 142 persona   
143 143 grazia   
144 144 sarebbe   
145 145 animo   
146 146 volta   
147 147 cielo   
148 148 fatta   
149 149 veramente   
150 150 stata   
151 151 ogo   
152 152 parti   
153 153 secondo   
154 154 materia   
155 155 alcuni   
156 156 giorno   
157 157 forse   
158 158 ce   
159 159 alli   
160 160 y   
161 161 subito   
162 162 potuto   
163 163 quod   
164 164 istesso   
165 165 tre   
166 166 delli   
167 167 f   
168 168 meglio   
169 169 appresso   
170 170 g   
171 171 primo   
172 172 anche   
173 173 faccia   
174 174 siano   
175 175 dall   
176 176 ro   
177 177 ex   
178 178 dato   
179 179 nvovo   
180 180 poter   
181 181 certo   
182 182 spero   
183 183 sanità   
184 184 libri   
185 185 dalle   
186 186 felicità   
187 187 esse   
188 188 havere   
189 189 cum   
190 190 quae   
191 191 acqua   
192 192 scrivere   
193 193 favore   
194 194 anno   
195 195 cui   
196 196 duca   
197 197 natura   
198 198 nvova   
199 199 cbe   
200 200 nostra   
201 201 te   
202 202 farà   
203 203 vita   
204 204 haveva   
205 205 alcuno   
206 206 quei   
207 207 alcune   
208 208 essa   
209 209 m0   
210 210 conto   
211 211 ora   
212 212 troppo   
213 213 principe   
214 214 virtù   
215 215 havuto   
216 216 sotto   
217 217 corpo   
218 218 dunque   
219 219 cuore   
220 220 ragione   
221 221 vista   
222 222 discorso   
223 223 verità   
224 224 via   
225 225 perciò   
226 226 pensiero   
227 227 passato   
228 228 proposito   
229 229 contro   
230 230 amico   
231 231 sè   
232 232 medesima   
233 233 scrive   
234 234 solamente   
235 235 voluto   
236 236 quam   
237 237 co   
238 238 desidero   
239 239 no   
240 240 fo   
241 241 8   
242 242 sin   
243 243 dar   
244 244 deve   
245 245 dopo   
246 246 voglio   
247 247 scrittura   
248 248 quest   
249 249 stesso   
250 250 riverenza   
251 251 benedetto   
252 252 avanti   
253 253 parole   
254 254 maniera   
255 255 presto   
256 256 sed   
257 257 aff   
258 258 opere   
259 259 voglia   
260 260 mezzo   
261 261 occhiale   
262 262 dei   
263 263 centro   
264 264 potesse   
265 265 5   
266 266 tanti   
267 267 grandissimo   
268 268 effetto   
269 269 clic   
270 270 nostri   
271 271 memoria   
272 272 doppo   
273 273 t   
274 274 tuttavia   
275 275 tali   
276 276 illli0   
277 277 buon   
278 278 etc   
279 279 mente   
280 280 finalmente   
281 281 veduto   
282 282 fuori   
283 283 somma   
284 284 intendere   
285 285 vorrei   
286 286 huomo   
287 287 padova   
288 288 sino   
289 289 bisogno   
290 290 haverà   
291 291 servirla   
292 292 male   
293 293 disse   
294 294 fiorenza   
295 295 havesse   
296 296 eli   
297 297 negozio   
298 298 contento   
299 299 signor   
300 300 dir   
301 301 causa   
302 302 pianeti   
303 303 adesso   
304 304 ancor   
305 305 ricevuto   
306 306 forza   
307 307 copia   
308 308 pisa   
309 309 macchie   
310 310 scrissi   
311 311 almeno   
312 312 par   
313 313 benché   
314 314 veder   
315 315 sicuro   
316 316 nvove   
317 317 lor   
318 318 ab   
319 319 11   
320 320 viene   
321 321 servire   
322 322 costà   
323 323 gio   
324 324 ri   
325 325 ss   
326 326 etiam   
327 327 huomini   
328 328 trattato   
329 329 inteso   
330 330 ngo   
331 331 fanno   
332 332 conforme   
333 333 card   
334 334 detta   
335 335 seco   
336 336 vera   
337 337 intanto   
338 338 tante   
339 339 fratello   
340 340 eccli   
341 341 degli   
342 342 iddio   
343 343 saranno   
344 344 maestro   
345 345 sti   
346 346 niente   
347 347 quell   
348 348 caro   
349 349 sto   
350 350 venire   
351 351 caso   
352 352 velocità   
353 353 tengo   
354 354 mihi   
355 355 dottrina   
356 356 dubbio   
357 357 figura   
358 358 simili   
359 359 speranza   
360 360 111   
361 361 hoc   
362 362 ragioni   
363 363 or   
364 364 fatti   
365 365 numero   
366 366 potrebbe   
367 367 stati   
368 368 intendo   
369 369 credere   
370 370 venetia   
371 371 ac   
372 372 aria   
373 373 visto   
374 374 b   
375 375 dare   
376 376 enim   
377 377 altezza   
378 378 possibile   
379 379 stima   
380 380 fin   
381 381 città   
382 382 essi   
383 383 cagione   
384 384 oblig   
385 385 cortesia   
386 386 stessa   
387 387 molta   
388 388 diligenza   
389 389 gloria   
390 390 parere   
391 391 linea   
392 392 rispetto   
393 393 trovato   
394 394 grandezza   
395 395 castelli   
396 396 dev   
397 397 tosto   
398 398 settimana   
399 399 cosi   
400 400 massime   
401 401 buono   
402 402 sigri   
403 403 tempi   
404 404 mezo   
405 405 qualità   
406 406 ultima   
407 407 simile   
408 408 nga   
409 409 forma   
410 410 sva   
411 411 bologna   
412 412 superficie   
413 413 alcun   
414 414 avviso   
415 415 fece   
416 416 trovo   
417 417 reverenza   
418 418 ricevuta   
419 419 difficoltà   
420 420 marzo   
421 421 vivo   
422 422 do   
423 423 dirò   
424 424 porta   
425 425 don   
426 426 rispondere   
427 427 necessario   
428 428 mille   
429 429 sento   
430 430 habbiamo   
431 431 resta   
432 432 età   
433 433 autore   
434 434 feci   
435 435 mostra   
436 436 poca   
437 437 sentito   
438 438 aspettando   
439 439 mese   
440 440 serli   
441 441 eccli0   
442 442 ove   
443 443 bisogna   
444 444 contrario   
445 445 erano   
446 446 atque   
447 447 là   
448 448 sola   
449 449 filosofo   
450 450 sera   
451 451 particolarmente  
452 452 quo   
453 453 rev   
454 454 ritorno   
455 455 senso   
456 456 dia   
457 457 francesco   
458 458 sentire   
459 459 forsi   
460 460 ecclia   
461 461 state   
462 462 20   
463 463 conservi   
464 464 havevo   
465 465 fatica   
466 466 stampa   
467 467 salute   
468 468 illustriss   
469 469 debito   
470 470 maggio   
471 471 pad   
472 472 die   
473 473 facendo   
474 474 ogn   
475 475 gratissima   
476 476 scrivo   
477 477 nello   
478 478 voi   
479 479 propria   
480 480 persone   
481 481 potere   
482 482 breve   
483 483 luglio   
484 484 ra   
485 485 poteva   
486 486 sei   
487 487 esperienza   
488 488 scudi   
489 489 tamen   
490 490 solito   
491 491 notte   
492 492 7   
493 493 uso   
494 494 supplico   
495 495 vetri   
496 496 benissimo   
497 497 ai   
498 498 negotio   
499 499 tiene   
500 500 elio   
501 501 oghi   
502 502 possi   
503 503 facilmente   
504 504 devo   
505 505 illli6   
506 506 padrone   
507 507 raggi   
508 508 possono   
509 509 spesso   
510 510 pochi   
511 511 servitù   
512 512 osservazioni   
513 513 meco   
514 514 grandissima   
515 515 vien   
516 516 sieno   
517 517 ricordo   
518 518 nondimeno   
519 519 manco   
520 520 svor   
521 521 maggiori   
522 522 vel   
523 523 ili   
524 524 faccio   
525 525 valore   
526 526 certa   
527 527 capo   
528 528 istessa   
529 529 quid   
530 530 amore   
531 531 desidera   
532 532 stimo   
533 533 siamo   
534 534 occhio   
535 535 passata   
536 536 farlo   
537 537 15   
538 538 vetro   
539 539 aprile   
540 540 hoggi   
541 541 sunt   
542 542 cerchio   
543 543 vedrà   
544 544 data   
545 545 infinito   
546 546 quantità   
547 547 12   
548 548 circonferenza   
549 549 trova   
550 550 passati   
551 551 seconda   
552 552 specchio   
553 553 ultimo   
554 554 sit   
555 555 siena   
556 556 voler   
557 557 peso   
558 558 rendo   
559 559 viaggio   
560 560 celeste   
561 561 leggere   
562 562 j   
563 563 nulla   
564 564 farmi   
565 565 fede   
566 566 fortuna   
567 567 aiuto   
568 568 saria   
569 569 dette   
570 570 potrò   
571 571 corpi   
572 572 perche   
573 573 fino   
574 574 spesa   
575 575 ancorché   
576 576 volentieri   
577 577 particolari   
578 578 osservationi   
579 579 stella   
580 580 honore   
581 581 mattina   
582 582 iam   
583 583 ultimamente   
584 584 piano   
585 585 bocca   
586 586 stampare   
587 587 9   
588 588 celesti   
589 589 autorità   
590 590 havessi   
591 591 farne   
592 592 grazie   
593 593 nova   
594 594 16   
595 595 dentro

There are certainly more user-friendly ways to present this information, but specialists might be surprised to see the following: “galileo” (27); “mani” (hands; 36) and “galilei” (39); “bacio” (kiss, 70); “libro” (book; 86); “terra” (earth; 96); “opera” (work; 127); “cielo” (sky; 147); and “libri” (books; 184). Some of these omissions are simply the result of bad OCR (or OCR that is bad in ways different from errors found in other documents, also telling in and of itself). () Yet, they signal that Sarrocchi is expressing honor and flattery in different terms than her contemporaries. She is also spelling Galileo's name with 3 l's. She is not talking about one of his books (or her own), nor is she connecting the discoveries in the skies to the realities on the ground. Her letter also seems devoid of the most frequent modifiers used by other writers in the corpus. Curiosities piqued here can be addressed via further exploration using the word\_info function described in Section 1.6. Because humanistic analysis, until recently, has prioritized the evidence that is most apparent, it is challenging to understand what is not there, but could have been. (, 22) This function seeks to visualize these lacunae.

The penultimate output from the choose\_doc function reports on similarity of rates of use of the 100 most frequent words in the corpus. Users are offered analyses using both Euclidean and Cosine methods (since the lengths of the documents can vary considerably). This is similar to Voyant Tools, the ePistolarium, and other tools that highlight the quantitative distance between documents. The graph shows the 10 most similar documents and the 10 least similar documents. (Figure 8) Aside from OCR errors that could easily skew this result right now, prototype users speculated that this might show the effects of different epistolary best practices in the period. Since this is based on lower quality OCR than would be desirable for a corpus-wide study, the results are best used to point to letters that might be found to be similar after close reading.

display\_png(file="./media/EdNaz574CosineComparison.png")



Such results should prompt exploration of both the similarities and the differences. For example: Why is Sarrocchi's letter so similar to EdNaz3104, Niccolò Fabbri di Peiresc's 1635 letter to Galileo? His letter is nearly twice the length of hers. While not ground-breaking, the function read\_doc (seen earlier) allows for immediate exploration of the potential stylistic or other similarities between these documents.

read\_doc(text\_attributes, "EdNaz3104") # Replace XXXX with an ID, please keep the ""

Molto Ill.re et Excellent.mo Sigli’ mio et P.ron Colli110 Io non ho potuto fare in servicio di V. S. Illli’0 alcun officio che meriti una minima particella della gratitudine ch’ella ne mostra nella sua cortesissima lettera delli 22 Febraio(1); et quando ne potrei far al centuplo, sì come professo desiderarlo ardentissimamente, non potrei sodisfare al debito mio et obligatione che tengo alla somma virtù et amorevolezza di V. S. Illli6; dispiacendomi di non saperlene esprimere condegnamente li sentimenti interni, per la poca prattica di cotesta longua volgare e per la debolezza dell’ingegno. Ma poi che veggo ch’ella s’appagga del cuore, m’assicuro ch’ella rim aneto sempre sodisfatta della mia fedele corrispondenza et del mio devoto ossequio, et ch’ella non sarà per rivocar in dubbio eh’ io non mi muova a far sempre ogni tentativo a me possibile por finir l’impresa, la qual, se Domendidio ci degna aiutare, doverebbe riuscire un giorno conforme a i voti et all’ oppinione eh’ io n’ haveva presa quando viddi la risposta dell’ Era,"10 S.r Carelli Padrone in una lettera scritta tutta di suo pugno, et non di mano o del concetto d’un secretano, havendo provato più volte che quando S. Emlia non gustava qualche proposta si è sempre contenuta nel silentio, senz’alcune scliuse nè altri complimenti; di maniera che quando viddi la sua risposta, se ben in poche parolle (21, presi grand’ animo et ardire di raddopiar l’officio nelli termini che V. S. Illli'0 haverà poi veduti(3), alli quali veramente S. Emlia non m’ ha replicato, se ben m’ha fatto risposta, di suo pugno ancora, sotto elli 2 Marzo, a diversi articoli della medesima mia lettera dov’ era inserito il secondo officio per V. S. Bili®: ma poi che son certo che n’ haveva fatto lettura per responder a gli articoli d’ella, mi giova credere ciò ch’ella m’accenna, elio non sarà stato senza qualche puntura et rimorzo d’immanità, et che il tempo et la patienza potranno far maggior operatione ch’ella non si persvade, massime concorrendovi gli officii potentissimi dell’Excellen.mo Sli Conte di Noailles; ot secondo la riuscita deU’ambasciata dell’Einli10 Sli Cardi0 di Lione(4) forzi elio vi si potrà un giorno far intervenire qualche suo officio ancora, sapendo eh’ onquella Corte, quando una grazia è risoluta privatamente, hanno a caro che ne sia fatto istanza da diverse persone, alle quali insieme se ne faccia la concessione publica: il che aspettando, non ho voluto per hora replicar altro in proposito della persona e negotio di V. S. Illli0 Y altr’ hieri, che passò qui l’ordinario d’Avignone per Roma, poi che S. Emlia non me ne faceva piti altra mentione. Ma per mantenere il negotio vivo, havendomi S. Emlia scritto che il P. Sylvestro di Pietra Slia gli havova presentato un suo libro De symbolìs heroìcis (4), che S. Pli\* m’ haveva fatto veder qui, passandovi questo Natale con Mgli Caraffa(2), Nuntio di Colonia, presi occasion di ricordare a S. Emli\*, che se la pressa deili altre maggiori et più degne occupationi non gli haveva permesso di leggere o scorrere detto libro, si degnasse vedere nel libro IV, al capli0 V, ciò che dice l’authore d’un liorologio hydraulico deli’inventione del P. Lino(3), del quale vederà qui V. S. Illli'6 il dissegno et la descrittione, che è cosa mirabile, se pur l’effetto può riuscire(4); et perciò che liauthore del libro non dice haver veduto la machina istessa nè nomina alcuni che li habbiano veduta, ho preggato S. Em.za di far chiamare detto Pli9 Sylvestro, et interrogarlo sopra la reai verità di questa madrina et d’intenderne ancora il parere di detto Mgli Caraffa, che ne doveva esser consapevole non solamente per haverne veduto qualche cosa, ma forzi anco per haverne penetrato li secreto. Anzi scrissi io ancora, sotto coperta di S. Emlitt, non solo ai detto Pli6 Sylvestro, che sta hora in Roma nel Collegio Romano, ma al detto Mgli Nuncio (li quale, passando qui incognito, volle venire a trattenersi due hore nel mio studiolo col detto Plie Sylvestro), per testificare all’uno et all’altro il dispiacere che mi rimase, doppo la lor partenza, d’essermi scordato di parlargli di quella macliina elei P. Lino per intenderne da loro medesimi ciò che se ne poteva credere, acciò di porgli in obligo non solamente di renderne conto a S. Emlia-, ma darmene qualche participatione et intervento in ciò che n’ haveranno da trattare con S. Emlia D.a onde io spero di prendere a suo tempo occasione di riparlare del negocio di V. S. Illli® con maggior vehemenza et forzi efficacia di prima, già che se la riuscita di questa machina è vera (si come mi scrive il Sli’ Pietro Paulo Rubenùri5’ d’Anverza con una sua lettera deli 16 Marzo, che ricevei hier sera, essergli stata testificata dal detto Pli® Sylvestro et da altri, che affermano esser tale come si rappresenta, havendogli aggionto detto Pli® Sylvestro che li havea veduta a belli aggio, et che Mgli’ Caraffa la fece portare a casa sua per essaniinarla con commodità, et ch’havendola osservata qualche giorno la trovò essattissima), par che sia una prvova et testificatione caduta dal Cielo inmano d’un Padre Giesvita, più tosto che d’un’altra professione, per non lasciar alcun lui?ogo di suspicione contra il testimonio di quel Padre inventore et di quell’altro che V ha publicata, per convincere il torto di quelli trovavano tanta repugnanza nella dottrina Copernicana et in ciò che V. S. n’ liaveva proposto per scherzo problematico. Anzi mi promette detto Sli Rubenio, grand’ ammiratore del genio di V. S. Illli0, di far un viaggio a posta in Liege per andare a visitar il P.re Lino et la sua madrina, il che non sarà senza darmene relatione; et io ce lo spingerò quanto più mi sarà possibile: et cercato qualche prattica et corrispondenza con detto Pli6 Lino per mezzo delli detti Sig.1'1 Caraffa et Plie Sylvestro o altri, poi che P hanno conosciuto: più tosto procurrerò di farlo chiamare in Roma et trattar che prendi la sua strada per questi paesi, per goderlo al suo passaggio et cavarne quel maggior costrutto che si potrà darne vivete vocis oraculo, s’ egli non porta seco l’ horologio hydraulico, in maniera che possiamo haver la vista qui nelle sue mani: il tutto per haver sempre nvovi argomenti di rammemorare V. S. Illli0 a qUe» c]ie ia possono affittare meglio di me. Nè tacerò mai che non mi sia imposto silencio, non pretendendo interessi alcuni in Roma nè altrove, per essere io pienamente contento della mia sorte, et per non considerare que’ clic sonno sopra di me che per haver compassione dell’amaritudini che patiscono, maggiori di me al centuplo, nò quelli che sonno sotto di me che per rendere grazie alla Di.na Mli\*1 dello stato dove mi ritrvovo, clic tanti altri più degni di me stimarebbono un paradiso terrestre, il qual mi par dover godere pacificamente, senza uscirne per andar cercar la malliora nelli maggior impieghi: et questo mi dà la libertà di parlare, dove gli altri restano muti, cornine ella dice, senza timore di perdere la fortuna et l’accesso di quelli alli quali io son pronto a continuare la servitù lecita, mentre non l’Laveranno discara, et non più; havendo imparato questa buona prattica dalla b. m. del Sli Gio. Vincenzo Pincllo, già 35 anni sonno, mentre V. S. Ill.re stava ella ancora nello Studio di Padoa. Di maniera che non mi è parso stranno ciò che mi scrivo V. S. Illli®, eh’ olla s’aftìige meno delli suoi disaggii di quel eh’ altri può credere, poi che gli rimangono tanti conforti et tante degne occasioni di esserci tare la vera philosophia, la quale è troppo facile et troppo indegna di grande raccomm and azione mentre si sta in prosperità, et al contrario si rende più splendida et rilucente al centuplo nell’adversità che gli porgi la fragilità humana; sì come li più generosi del mondo passarebbono vva vita ignava et indegna di memoria, se gli mancassero nemici et occasioni de guerra e di vittoria da cssercitar il lor valore, la sola adversità prencipalmente havendo fatto celeberrimo il buon Giobbo et tanti Sli' Padri et philosophi delli maggiori dell’ antiquità, la cui constanza et magnanimità gli ha fatti degni d’ammiratione alli postori, come sarà anche V. S. Illli", nonostante qual si voglia morsura dell’ invidia. Et quel voto solo che con tanta gentilezza et gravissima prudenza V. S. Illli0 si degna fare, che trvovino pur nvove machine li suoi nemici, ch’ella gliene renderà grazie se le hanno da fruttare le dolcezze ch’ella sente negli officli de compassione ch’ella riceve da gli amici et servitori, non merita meno appresso la posterità, a mio giudicio ben che debole, che gli apotliegmi più celebri di tutti gli savii della Grecia antiqua. Et la confidenza con la quale ella degna discredersi meco, mi rapisce il cuore del tutto: di che rendendole quelle maggior grazie che posso, le fo humilissima riverenza, et preggo dal Signore la continvata felicità interna et ^acquisto deli l’esterna, quando piacerà alla D. Mali1'1 Di Aix, alli 1° Aprile 1635. Di V. S. molto Ill.re Devotissli et Oblig.mo Sei\*.1’0 Di Peiresc.

A close reading reveals that both writers are establishing a similar rhythm of *captatio benevolentiae* (the rhetorical practice of establishing good will in the reader through obsequiousness and flattery). They are also following a well-established formula of pleasantries around replying to mail. The function breaks Sarrocchi from the post-hoc categories that have followed her frequently (but not always) in scholarship. Instead of grouping her with women to study gender differences in style, this approach uses the larger-scale quantitative information to suggest small-scale pathways of exploration. In some cases these results might encourage biographical research, in others rhetorical and literary study.

Seen through this computational and quantitative lens, Sarrocchi's correspondence and literary works are not being considered from the typical scholarly contextualizations within works by other women writers, literary authors, and those outside learned academies. Yes, these are all categories to which she belonged, but they were not necessarily the definitive boundaries for the development or deployment of her writing style. This tool and the others that follow are meant to identify further potential communities and similarities. This is not meant to exclude or supercede the results of late-19th and 20th century humanistic epistemologies, but to admit space for permeability across these boundaries that reflects recent scholarship.

The final contextual information in "choose\_doc" reports results from topic modeling. GaLiLeO aspires to providing pathways of interest, not interpretations. Much like the ARTFL Encyclopédie (), that uses multiple digital models for users to explore, a preliminary topic model was made from each corpus and subcorpus for proof-of-concept purposes. Like the stylometric output that is also included, the model was not optimized given the condition of the OCRed text, so after removing the stop words used for Italian in Voyant Tools, all remaining words (even those occurring in one document only) were considered to infer just 25 topics. The spirit of applying LDA to the corpus was similar to that of the ARTFL team with the Encyclopédie, in which topic modeling was used to explore the ways in which lexical choices of entries cut across the categorization imposed by the structure of the encyclopedia. () Here, the categories are the ones imposed by the other quantitative methods, and LDA is one of the tools to explore their permeability. The model is created in advance to create run-time efficiencies for the demonstration (which itself was slow at times). Information about whether or not a term is a top word in a topic is preserved and reported by the word\_info function (described in more detail below). At the document level, the output merely identifies the top 3 topics from a preliminary model of the subcorpus.

The challenge of LDA as a navigational tool is that all of the terms that have a high likelihood of indicating a topic in a document are not necessarily present in the documents that have a likelihood of representing the topic. This last output of choose\_doc reports word co-occurrence at a corpus level, not a document level, by indicating the topics likely present in the letter. For instance, if an author uses “sole” (sun) he or she is likely to use “terra” (earth), “parte” (part), or “sopra” (above) as seen in Topic 2 in the choose\_doc output final lines, but those words are not all present in any given document, nor are other contextual words of interest. Lexical tools described below try to account for this gap between the probabilities of the models and the actualities of the documents.

Despite the completely rudimentary nature of this proof-of-concept topic model, the results do point to aspects of the text and its context that were revealed by the other output from choose\_doc. Sarrocchi may not use all of the words in a topic, but enough of them to be associated with the pleasantries of letter writing (most prominently), a discussion of natural philosophy, and networking within or by means of members of learned academies. (NB: The topic numbers listed are not the topic numbers in the model, but the ranking of the topics likely present in the document. Their order will rarely, if ever, match the order of the list provided in the lexical contextualization section below.)

A specialist can add dimensions of contextualization here, while a novice can use the output to continue to explore. With both this and the final output of choose\_doc, the low quality of some of the OCR data undermines the strength of any emerging argument about this specific letter by Sarrocchi, but encourages a return to the materials, physical or their digital surrogates, to test new hypotheses. These kinds of questions could be expanded to other corpora because they are based on the kinds of patterns, differences, and absences that are common to intellectual pursuits in history and literature.

## Lexical Tools in GaLiLeO

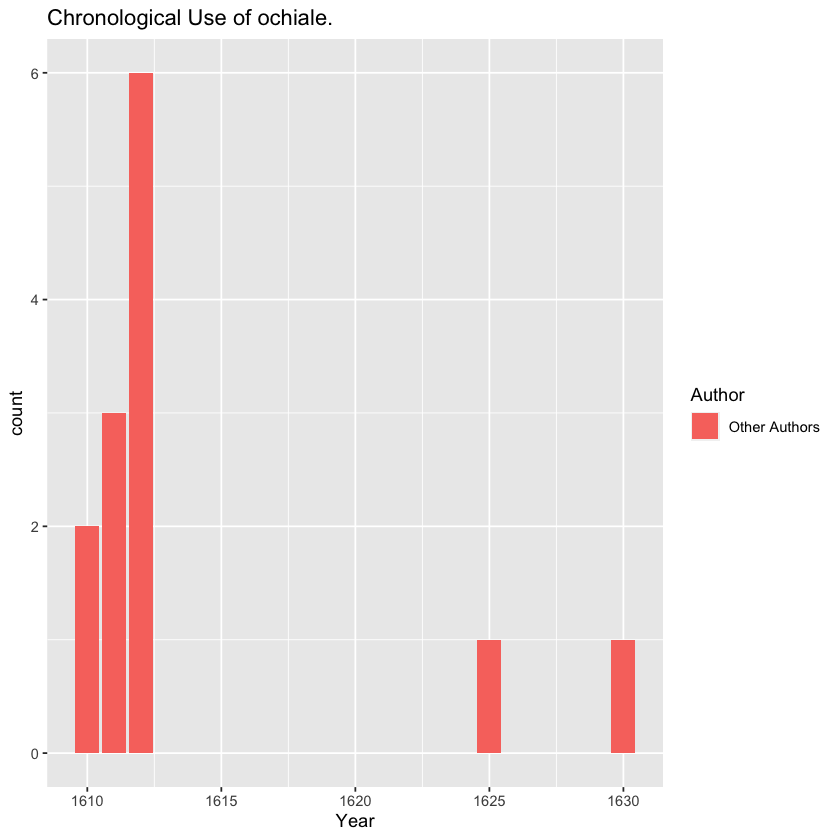
The second half of the GaLiLeO prototype notebook was dedicated to learning more about specific words of interest. To assist with usability, the function “word\_info” outputs features of word types based on embedded functions. Users see the term contextualized chronologically, within author subsets (currently limited to Galileo and a group that includes all other authors), and within top words for the topic model (when applicable).

To demonstrate and evaluate the methodological priorities that guided the development of these functions, the same letter will serve as the example. Sarrocchi's expressions of enthusiasm and engagement with natural philosophy might draw attention to her orthography and whether or not it was truly unusual. Upon close reading of the letter, the term *ochiale* would look unusual to even beginner students of Italian, since its root *occhio* (eye) is the now-standard spelling with the double-c. The term *ochiale* is an alternative spelling for the Italian word for spyglass or telescope *occhiale*. The GaLiLeO tools brought this unusual orthography to light, rather than obscuring it behind editorial practices and database elements of lemmatization and orthographical standardization. Lemmatization and certain editorial practices collapse these variants into one form for searching for all of the letters that talk about the telescope such that a search for *ochiale* produces no results in the Galileoteca.

Instead, the function word\_info retrieves chronological, authorial, and textual context. The chronological information, if it were based on more accurate OCR output, would prompt consideration of how long it took for one of Galileo's Italian terms for the telescope, occhiale, to become standardized. The preliminary numeric results in the top lines of the output suggest about 20 years. The accompanying bar chart (output last) presents more granular information. This year-by-year information points to a more rapid standardization, within three years. Calling the function for different variants used to describe the telescope would better test the theory.

word\_info(type\_attributes, "ochiale")

The earliest occurrence of this word is 1610 .  
The latest occurrence of this word is 1630 .  
It appears 16 times in the corpus overall:  
Not present in works written by Galileo.   
16 time(s) in documents by others.  
The word appears in the following topics:  
 ochiale  
nera\_nero\_bianco\_bianca\_he 0  
tabacco\_servit\_tose\_cred\_antecedentemente 0  
insti\_lattantio\_rassegno\_stan\_nomo 0  
duos\_invenire\_numeros\_quesito\_ritira 0  
a\_\_o\_\_perche 0  
sag\_ducati\_germini\_vetri\_vetro 0  
est\_quod\_quae\_quam\_sed 0  
que\_el\_en\_para\_forzi 0  
zugmesser\_chepplero\_kepplero\_brahe\_toccare 0  
e\_\_piu\_\_sole 0  
riformagioni\_vapores\_stupore\_esporra\_ 1  
molto\_e\_\_piu\_ 0  
fratte\_bagno\_ludovisi\_contrarie\_permette 0  
a\_\_ma\_\_havea 0  
sti\_sigti\_eccti\_ulti\_serti 11  
negozio\_illustriss\_invenzione\_longitudine\_ss 0  
vene\_sangue\_spiriti\_arterie\_bx 0  
svor\_celeste\_amatiss\_fig\_vino 1  
tedeschi\_prohibire\_fiore\_chela\_gb 0  
aqqua\_fiume\_canale\_diritto\_ta 0  
afforismo\_fait\_scrivi\_praestantissima\_raccommandazioni 3  
diodati\_lione\_reverentia\_tengho\_piegho 0



While the results have an immediate use for exploring other documents and terms, the underlying data about the use of *ochiale* (and all terms in the corpus) provide a foundation for pattern-based inquiry in which the specific word is unknown, but a user would like to search for words with known features. A series of logical tests could identify which words came into vogue in a certain period or with a specific relative frequency in the corpus. Section 1.6 will experiment more with terms that were or were not used by Galileo in high or low frequency compared to other authors in the collection.

The second set of output from word\_info relates the relationship of the searched term to the topic model of the corpus. For the prototype, the topic information stems from the same preliminary model of 25 topics that informed the choose\_doc output. The numeric column indicates how many of the 16 total instances of ochiale are assigned to different topics. The most frequent designation is a topic that might be labeled "excessive noble forms of address." Since this model is not optimized for the corpus, the result points to further investigation. Were this a reliably descriptive model of the letters, a scholar could speculate about correspondents outside Galileo's immediate circle adopting a common style of communication for approaching a second-order (or higher) member of the larger social network. One might also investigate if and how Galileo's preferred variant, occhiale, appeared in topics. At this point the model *of* the documents becomes a model *for* an exploration of naming conventions.

The related function “get\_KWIC” is modified from Matthew Jockers’ code in *Text Analysis with R for Students of Literature* (). While a platform like Voyant can also provide contextual information for term usage, often the title of the document is the only metadata available for sorting the results. The contextual tools in Voyant also draw attention to high-frequency terms moreso than unusual ones. The tools in GaLiLeO add year of publication (or date sent) and the simplistic label of whether or not the text in question was written by Galileo (GG) or someone else (NotGG), in keeping with the theme of the other functions. The context column contains a helpful artifact of the data's creation: an indication of the occurrence number.

With a term like the related *telescopio*, which has several hundred results, even the list needs to be parsed to make sense of the kinds of use of the word. Computation can help. For example, in the output of the code cell below, 13 documents use ochiale a total of 16 times. The functions described above permit further exploration of the community of users with similar lexical habits. At the same time, the output indicates where *ochiale* occurs in each text. While it is visualized in a redundant fashion, in the second line of the output below, the artist Ludovico Cardi da Cigoli's letter to Galileo in March of 1610 used this variant spelling twice in a relatively short letter (322 words total): once as the 139th word and again as the 242nd word. This contextualizes the term in such a way as to see authors who make consistent use of a term and those who open a letter with it (the first line of the results, for instance), but never mention it again.

get\_KWIC(type\_attributes, "ochiale")

ID GG Year  
1 EdNaz299 NotGG 1610  
2 EdNaz273 NotGG 1610  
3 EdNaz574 NotGG 1611  
4 EdNaz601 NotGG 1611  
5 EdNaz632 NotGG 1611  
6 EdNaz651 NotGG 1612  
7 EdNaz666 NotGG 1612  
8 EdNaz718 NotGG 1612  
9 EdNaz729 NotGG 1612  
10 EdNaz736 NotGG 1612  
11 EdNaz753 NotGG 1612  
12 EdNaz1719 NotGG 1625  
13 EdNaz2015 NotGG 1630  
 Context   
1 ochiale 12 1110 mio padrone sono arivato qua e gli mando 1 ochiale e quando v s vedessi che sia appallato la cavi   
2 ochiale 139 grami con lei avendo ridotto a tale xierfezione il suo ochiale ch ella à potuto scorgiere et osservare nel cielo cose ochiale 242 a venezzia ad un suo ministro ne procurassi il detto ochiale et se il libro era stanpato gniene mandassi ora la  
3 ochiale 168 io con li proprii occhi l ho vedute mediante l ochiale del sigti gallileo et fattele vedere a diversi amici il   
4 ochiale 73 mancherò tornarla a suplicare che si vogli aricordare del mio ochiale ciò è di inviarmelo al che fare la ne prego   
5 ochiale 96 il sti lodovico li averà scritto come con un mio ochiale ho fatto alcune osservationi di nobi nel sole li quali   
6 ochiale 250 che giove lo vede montuoso vidi bene con il suo ochiale nel dintorno della lui na due merlature assai evidenti et   
7 ochiale 103 credo avere scritto a y s come io ò uno ochiale et è assai buono tanto che veggo da santa maria ochiale 530 fatte vedere ad altri ancora nè credo sia imperfezione dello ochiale poi che le veggo varie et delle tonde et delle   
8 ochiale 303 che è cosa ridicola et che si fa fare uno ochiale a venezia che sarà lui ngho tre braccia con il   
9 ochiale 358 ero scordato di dire che le machie sono cavate dal ochiale così dentro la stanza però credo tutte venghino da rovescio   
10 ochiale 83 vedrò se si potrà acomodare un regolo unito con l ochiale che si possino fare pili giuste s ella à modo   
11 ochiale 69 fino che si è ocultata per quanto mostra il mio ochiale et se bene è andata variando sempre di forma e   
12 ochiale 52 suo et perchè io fo anche mentione di questo novo ochiale di veder le cose minute et lo chiamo microscopio veda   
13 ochiale 43 li disse che mi domandassi se li havevo portato un ochiale di quelli del s r galileo conforme a una memoria ochiale 155 gratia di far chiedere al s l galileo il detto ochiale con altro strumento che haveva preparato e v s ill

While 16 results are small enough to compare manually, more common terms create an organizational challenge for finding contextual patterns. For digital humanists, topic modeling, seeded by keywords, can alleviate this challenge, although with the same caveat that the terms that predict a topic do not all co-occur together in every document that has been inferred to represent the topic. For non-digital humanists, another tool was needed.

The underlying function in GaLiLeO for this contextual lexical comparison is see\_KWIC, which visualizes terms that co-occur with the keyword over time. This function allows for significant customization for users who do not mind typing up lists of words to exclude. The initial stop word list is quite restricted: definite and indefinite articles, Galileo's name, prepositions, some pronouns, and a handful of terms that are artifacts of the OCR process and data creation. This is intentional. Depending on the type of query, certain co-occurrences might be of tremendous interest. A scholar might be interested in modifiers (even otherwise banal terms such as *assai*/rather, *molto* /many or very, or *poco*/few). A different user might want to exclude verbs, while someone else might want to analyze them. With that in mind, users can currently modify the stop words list (similarly to Voyant). A future option would define only the words to count as context, establishing a white list, which could also be modified.

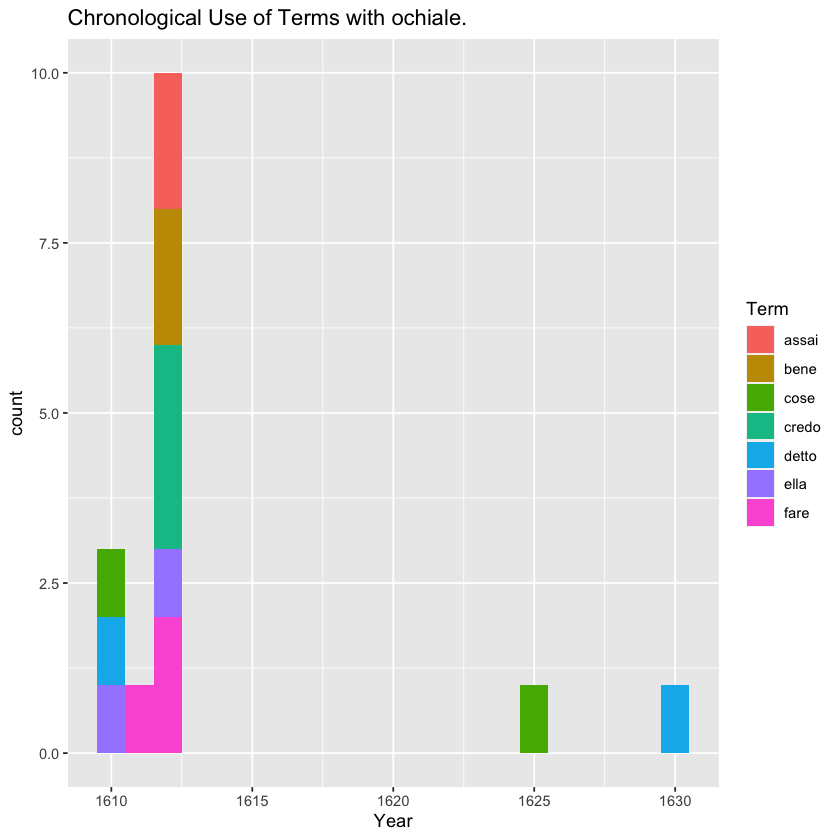
sort(stop\_words, decreasing =F)

[1] "10" "100" "4" "40" "5" "a"   
 [7] "à" "ab" "ad" "agl" "agli" "ai"   
 [13] "al" "all" "alla" "alle" "alli" "allo"   
 [19] "anche" "and" "aut" "avrÖ" "avrš" "c"   
 [25] "che" "chi" "ci" "cioè" "coi" "col"   
 [31] "come" "con" "contro" "così" "cui" "cuius"   
 [37] "cum" "d" "da" "dagl" "dagli" "dai"   
 [43] "dal" "dall" "dalla" "dalle" "dallo" "de"   
 [49] "degl" "degli" "dei" "del" "dell" "della"   
 [55] "delle" "dello" "di" "dov" "dove" "e"   
 [61] "ed" "elsewhere" "et" "ex" "farÖ" "farš"   
 [67] "fra" "galilei" "galileo" "già" "gl" "gli"   
 [73] "hoc" "i" "il" "ills" "in" "io"   
 [79] "ista" "l" "la" "le" "li" "lo"   
 [85] "ma" "mo" "ne" "negl" "negli" "nei"   
 [91] "nel" "nell" "nella" "nelle" "nello" "non"   
 [97] "o" "of" "one" "per" "perch\_" "pi\u009d"   
[103] "qua" "quale" "quanta" "quante" "quanti" "quanto"   
[109] "quei" "quel" "quella" "quelle" "quelli" "quello"   
[115] "questa" "queste" "questi" "questo" "qui" "r"   
[121] "re" "sarÖ" "sarš" "se" "seu" "si"   
[127] "sì" "sig" "starÖ" "starš" "su" "sugl"   
[133] "sugli" "sui" "sul" "sull" "sulla" "sulle"   
[139] "sullo" "sum" "text" "the" "to" "tra"   
[145] "tres" "tutti" "tutto" "tuum" "un" "una"   
[151] "under" "unknown" "uno" "unum" "ut" "with"   
[157] "x" "x8f" "y"

expanded\_stop\_words <- c("mi", "m", "me", "mio", "mia", "mie", "miei", "ti", "t", "te", "tuo", "tua", "tuoi", "tue",   
 "si", "s", "suo", "sua", "suoi", "sue", "nostro", "nostri", "nostre", "nostra",   
 "vostra", "vostro", "v", "vostri", "vostre", "loro", "lui",  
 "quali", "quale",   
 "ho", "hai", "ha", "hà", "abbiamo", "habbiamo", "avete", "havete", "hanno",   
 "è", "sono", "siete", "siamo", "sei",  
 "ill")  
stop\_words <- c(stop\_words, expanded\_stop\_words)

For the case of Sarrocchi's alternate spelling of *ochiale* (and the restricted sample of texts used for this article), the graph helps to reveal that authors who use this variant are not consistent in their use of content-bearing contextual terms (at least with a context of 20 words). They use the formal term of address (*ella*) and the phrase or terms *assai bene* (rather well). In contrast, letter writers who use Galileo's preferred *occhiale* (not pictured) over time seem to consistently have done or made something (*fatto*), rationalize when they speak about it (*perchè* - because), and discuss a gain or excess (*molto* and *più*).

see\_KWIC(type\_attributes, "ochiale", stop\_words)



The importance of a word is characterized by these functions as more than a highest frequency value or as a marker of distinction (such as tf-idf). By contextualizing the term within textual and metadata information, a user can ask questions about patterns without knowing the words that fit those patterns. The visual and quantitative display create a critical distance from the text that can provide an alternative perspective on the lexical relationships embedded in the text. This invites surprise, questions, and potentially evaluation of previous claims of importance in a collection. The final tools in GaLiLeO attempt to make this explicit, but are ultimately quite simplistic in the prototype.

## Pattern Tools in GaLiLeO

Where the document and lexical tools can help to identify a potentially fruitful avenue of inquiry, they are aided significantly by some contextual knowledge on the part of the user. Alternatively, the more ambitious pattern-based tools attempt to foreground the kinds of questions and structures that can motivate historical and literary analyses. These final tools still need significant development and refinement.

One function allows direct comparison of authors' vocabulary. The eventual goal is to better understand lexical variety, breadth, and overlap. The first step, which was intended to provoke conversation at the prototype workshop, simply reports on the size of the vocabulary in the texts in the subcorpus by a given author. To be more helpful, it should also output the number of documents and the length of those documents. In the example below, the function is called three times to compare Galileo's lexicon to that of his daughter (Suor Maria Celeste) and Margherita Sarrocchi.

compare\_author\_vocabulary("Galilei, Galileo", corpus, text\_attributes, corpus\_attributes)

Galilei, Galileo uses 13349 different words (not including punctuation).  
The overall corpus contains 75798 different words (not including punctuation.

compare\_author\_vocabulary("Galilei, Maria Celeste", corpus, text\_attributes, corpus\_attributes)

Galilei, Maria Celeste uses 4970 different words (not including punctuation).  
The overall corpus contains 75798 different words (not including punctuation.

compare\_author\_vocabulary("Sarrocchi, Margherita", corpus, text\_attributes, corpus\_attributes)   
#Replace the Xxxx with Last name, First name, please leave the ""

Sarrocchi, Margherita uses 1258 different words (not including punctuation).  
The overall corpus contains 75798 different words (not including punctuation.

In this subcorpus Galileo has 5 times as many letters as his daughter; Maria Celeste's corpus of 64 letters is already 16 times greater than Sarrocchi's. Their associated vocabularies reflect slightly different proportions: Maria Celeste's is nearly one quarter of Galileo's; Sarrocchi's is one seventh of hers. As with nearly every other feature, the focus is on contextualizing Galileo against the backdrop of other authors in the corpus or subcorpus. A version of this function is already provided in the document-level information, which showed that Sarrocchi used an alternate orthography for the Italian terms for academics and mathematics in letter 574 in the National Edition of Galileo's work. This function expands those results to all documents by the author and, at least at this preliminary stage, allows for comparison by range of use, not simply single occurrences in a document.

The final tool perhaps holds the most promise as a template for query building. The function “find\_types\_by\_range” responds to questions of the type: What are the words that Galileo uses in high frequency that the other authors in the corpus do not use at all? Again, this is a proof of concept, which needs more customization. For now, high is considered to be the 75th percentile or higher in frequency of usage; low is the 25th percentile. Users can also input a simple yes or no to indicate words that Galileo or other authors use (yes) or not (no).

#Takes a few seconds to run  
find\_types\_by\_range(corpus\_attributes, GG="high", NotGG = "no", size =25)

For context, here is the overall range of relative frequencies in the corpus:  
The lowest frequency is 0. (No document uses all words in the corpus.)  
The highest is:  
[1] 0.1282051  
A numeric summary of the means for each frequency describes this context in more detail:  
 Min. 1st Qu. Median Mean 3rd Qu. Max.   
0.00000003 0.00000040 0.00000103 0.00001319 0.00000249 0.03233686   
The range of mean vocabulary frequencies in Galileo's documents is the following.  
 High is considered higher than the third quartile (75th percentile).  
 Min. 1st Qu. Median Mean 3rd Qu. Max.   
0.000005484 0.000005484 0.000005484 0.000052840 0.000016451 0.030873506   
Other authors do not use 5717 words that Galileo does use. Here are the words that match your query:  
[1] "1689, afflizzione, aggiugnesi, alterazioni, alternatamente, ancorette, animella, applicano, assegnar, attaccate, averse, bernengero, capellatura, cercata, circonvicine, comandino, commozioni, concludenti, concludenza, contenente, contenuti, contradizioni, costare, crini, dedurre, deputare, descrizioni, diffiderei, dintorni, direzzioni, distendono, durante, egualissime, eretto, essenziale, estendono, fabbricato, finzione, firlie, giustificazioni, gp, ibe, ilev, illuminante, illustrate, ingrandimento, inibi, irradiazione, laboriose, languida, languido, maj, mantenne, manuscritte, mazzei, megli, meridiane, mettergli, mitre, ns, ooo, osservabili, osservargli, padlie, piccolissimi, piccolissimo, pisis, posatamente, pot, pulitissima, purgarmi, regii, remote, restandogli, reud, reulia, rodomonte, rupi, scorcio, scultura, settore, sfuggimento, sincerarmi, siviglia, sporgono, sx, tenebroso, trattone, ugr, vassalli, vertici, vibrazioni, vicinissime, visibil"  
[1] 50 2  
[1] 50 2  
 GGtokens Freq  
101 1689 0.000021935  
682 afflizzione 0.000021935  
955 alternatamente 0.000021935  
1149 ancorette 0.000021935  
1227 animella 0.000021935  
1402 applicano 0.000021935  
1880 attaccate 0.000021935  
2031 averse 0.000021935  
2874 cercata 0.000021935  
3068 circonvicine 0.000021935  
3257 comandino 0.000021935  
3324 commozioni 0.000021935  
3543 concludenza 0.000021935  
4277 costare 0.000021935  
4387 crini 0.000021935  
4611 dedurre 0.000021935  
4761 deputare 0.000021935  
4804 descrizioni 0.000021935  
5054 diffiderei 0.000021935  
5167 dintorni 0.000021935  
5189 direzzioni 0.000021935  
5729 durante 0.000021935  
6300 essenziale 0.000021935  
6333 estendono 0.000021935  
6877 firlie 0.000021935  
 GGtokens Freq  
16276 siviglia 0.00002741874  
16817 sporgono 0.00002741874  
18598 vicinissime 0.00002741874  
4041 contradizioni 0.00003290249  
5389 distendono 0.00003290249  
10531 mitre 0.00003290249  
11190 ns 0.00003290249  
12835 posatamente 0.00003290249  
14209 remote 0.00003290249  
15561 scorcio 0.00003290249  
18543 vertici 0.00003290249  
714 aggiugnesi 0.00003838624  
11769 osservabili 0.00003838624  
18119 ugr 0.00003838624  
3541 concludenti 0.00004386999  
3976 contenuti 0.00004386999  
18682 visibil 0.00004386999  
3949 contenente 0.00004935374  
9188 irradiazione 0.00004935374  
17447 tenebroso 0.00004935374  
950 alterazioni 0.00005483749  
17929 trattone 0.00006032124  
18331 vassalli 0.00006032124  
18586 vibrazioni 0.00008225623  
15664 scultura 0.00012064248

For users of the prototype, this aspect of the tool simply provoked a sense of the magnitude of the unexplored. The results are cumbersome, given the state of the underlying textual data. Similarly, Patrick Juola and Ashley Bernola's Conjecturator () compares textual features such as the use of a word within two groups of texts that are differentiated by a structural aspect such as decade. With over 87,000 statistically significant observations, these results are "mind-numbing, even demoralizing." (, 135) As instantiated, the find\_types\_by\_range output still requires specialist knowledge to select a potentially useful path forward. Further interactivity and use of visual cues would help to address some of these concerns.

With a tool such as this an author from the period would not need to fit or fight a corpus-scale model of a collection of texts; she would have her own. These permutations of terms provoke questions about individual models of language use rather than a best-fit model for all authors represented. Sarrocchi's document can be compared to the authors' corpus and the corpus at large. The topic model of the overall corpus would be deprioritized in order to see and compare individual authors and texts. In this way, GaLiLeO might capture some of Underwood's perspectival modeling () without risking the bias identified by Katherine Bode (). Perpetual reference and comparison to a model that is heavily influenced by the prolific authors found within it continues to disenfranchise the already marginalized figures who were nevertheless writing and creating in the period. Continuing the example of Sarrocchi, a future instance of GaLiLeO would facilitate a comparison of her letters and the text of her epic poem (with which Galileo was familiar) to Galileo's corpus or to the library corpus overall. Her Latin geometrical and theological texts (should they be found) could be distinguished from her letters and text in Italian for a more nuanced comparison of stylistic, thematic, lexical, and contextual similarities or differences.

## Implications and Next Steps

Current use of GaLiLeO is limited to the two principle investigators (and the exploration facilitated by mybinder functionality with a sample of the data for this article), it has always been intended for a broader audience of Galileo scholars, interested peers in history and literature, and students. Thanks to financial support from Harvard University's Dean's Competitive Fund for Promising Scholarship, Galileo scholars were able to convene to test this alpha prototype and brainstorm next steps. In spite of being able to publish new scholarship on Galileo as a result of these tools (), this project is on hold since August 2022 due to two primary obstacles. First is the question of the corpus and whether or not to unedit Galileo and his colleagues to create new digital materials from manuscripts. () Second, the developers are searching for an outlet to publish the metadata associated with the correspondence, the full texts of materials from the library, and the code for running analysis. At the time of the most intensive project development, the project leads' local institutions were not in a place to support long-term scalable hosting of the project. GaLiLeO is thus a project in search of scholarly and institutional collaborators for building a user-friendly interface, a robust underlying data set, and hosting.

Despite these limitations, the preliminary tools in GaLiLeO do facilitate the kind of question making that has guided humanist investigation. Margherita Sarrocchi's letter has been uncovered for its unusual lexical choices and contextual variety. Her style, lexical richness, and document complexity were seen in comparison to contemporaries without the siloes of gender, genre, affiliations, or chronology that often shape corpus-scale document organization and analysis. The tools also documented the popular subjects not mentioned by Sarrocchi. The computational and quantitative results become a window onto the collection and indication of fruitful pathways, rather than a declaration of large-scale patterns that often obscure variations. By providing a way to enter the corpus without inherited interpretive lenses, GaLiLeO invites the continued work of discovery, uncovering, recreating, differentiating, evaluating, contextualizing, and documenting representations and expressions.

The day-long workshop also generated exciting ideas about what a project like this could be. The energy and interest were palpable. The diverse range of digital community needs that were discussed went far beyond the scope of GaLiLeO. The conversations also identified immediate and persistent obstacles. The project directors realized immediately at the workshop that the ability to see words contextualized at multiple levels of address both identified points where a collaboration was necessary and how the tools facilitated that collaboration. The literary scholar and digital humanist could see the poetic and dramatic points of contact with the texts and narratives that were most familiar to the historian of science collaborator for other attributes. With someone like Sarrocchi, this reveals that the historian might consult with an art historian to better understand the similarity with the painter Cigoli, a social historian or historian of science specializing in the creation and reception of institutions for validating knowledge in early modern Italy, and a literary scholar for the connections between her letters to Galileo and the types of expression found in her epic poem.

The other general pieces of feedback included the following features, listed here to help with the planning of future DH projects:

* More tools would be welcome: an API, historical text reuse detection, reporting on avoided topics, named entities, and a “humanistic suggestion tool” that offers models for queries along with rationales.
* More documents would be welcome: marginalia, more library texts, better OCR, images, and machine readable text based on manuscripts instead of 19th century editorial versions of letters.
* Transmedia linguistic analysis would be a valuable further step.
* GaLiLeO represents an opportunity to bring new perspective on digital projects that are considered “complete” by linking from its terms, people, years, and texts to the lemmario, biographies, chronologies, and bibliographies that have been developed by colleagues in Italy and the U.S. These include the Galileo//thek@, The Galileo Project (), and even the authority file of 8,000 names connected to Galilean materials created by the Museo Galileo (<http://www.museogalileo.it/esplora/biblioteche/biblioteca/archivionomi.html>){:target="\_blank"}
* The project will benefit from phased roll-outs of the underlying data and metadata, particularly as more texts from the library are added and additional metadata tags are created.
* This is an opportunity to involve undergraduate and graduate students of Italian in an exploration of the history of Italy as well as the ways in which Italian content can influence the development of predominantly Anglophone technology. It could also be a platform to put specialist, informed readers in contact with novice, curious readers.

The field of Galileo studies is moving in directions that still require the attentive philological study of what Galileo and his contemporaries wrote, but with the integration of other disciplinary perspectives. Some of the new schoarly work on Galileo focuses on visual culture, both in terms of diagrams and illustrations but also the arts. Renewed attention has been brought to Galileo's marginalia and reading practices. Researchers are also pushing back against claims of exceptionalism to highlight linguistic traditions in which Galileo participated and also the popular topics that he avoided. Ultimately, embedding the project within the digital Galileo ecosystem would benefit the field more than creating a new digital history tool with claims to being something definitive that incorporates all materials. Using this tool in conjunction with the lemmatized resources in the Galileo//thek@ or the names in The Galileo Project would still maintain the experimental, exploratory spirit of the GaLiLeO tools. The Archilet - Reti Epistolari project has already created a model for this with metadata related to literary correspondence in Italy 1600-1800. () While GaLiLeO cannot yet address these aspects of embeddedness, the motivated user will be able to connect the output from these tools as search queries in this larger ecosystem.

## Project Credits

It is important to note that the successful demonstration of the GaLiLeO prototype was also the result of several groups of collaborators. A special note of thanks is owed to Stephen Houser, Director of Academic Technology and Consulting at Bowdoin College, for his assistance with the preparation of this article. Other collaborators included:

*Galileo’s Correspondence*

* Hannah Marcus and Paula Findlen: metadata and tagging schema of letters
* Data curators: Hannah Marcus, Daniele Macuglia, Rachel Midura, Mackenzie Cooley, Paolo Savoia, Demetrius Loufas, Brian Brege, Padraic Rohan, Chris Bacich, Julia Roever, Charlotte Thun-Hohenstein
* Undergraduate assistants: Kyle Lee-Crossett, Max Morales
* Morgan MacLeod: Python tools for tag analysis

*Galileo’s Library*

* Crystal Hall: GaLiLeO infrastructure & tool development, corpus integration, library and correspondence document gathering, text cleaning, analysis, documentation, coordination with Harvard IT support for the workshop in September 2018.
* Aaron Gilbreath, Academic Tecnology & Consulting, Bowdoin College: experimental shiny app design for visualizing descriptions of the corpus (no longer supported)
* Additional transcribers and text cleaners: Gabriella Papper ’18, Dean Zucconi ’19, Ingrid Horton (MAI Services)

*Workshop Support*

* William Barthelemy, Academic Technology, Harvard University Institutional Technology

## Bibliography

# Check R version  
R.version

\_   
platform x86\_64-apple-darwin17.0   
arch x86\_64   
os darwin17.0   
system x86\_64, darwin17.0   
status   
major 4   
minor 2.3   
year 2023   
month 03   
day 15   
svn rev 83980   
language R   
version.string R version 4.2.3 (2023-03-15)  
nickname Shortstop Beagle