Sample Project 3: Food and Civility 1650-1800

Overview

The second half of the seventeenth through the eighteenth century was a period of vast transition during which various new foods became symbols of changing notions of civil comportment and sociability. The accessibility and affordability of some of these new foods—whether from the East or the New World—challenged social conventions. For instance, perhaps one of the most contentious of the new offerings was coffee, which led to the rise of the coffeehouse as a place of open social interaction that transgressed centuries-old norms of social status and political discourse. Whether it was a food itself, or the social practices that accompanied it, new foods such as coffee became associated with different kinds of classes within early modern societies. While coffee itself was not as powerful a social factor as alcohol, it was accessible to anyone and consumed in environments where tongues wagged—resulting in its association with less-refined society.

Other foods, however, were less accessible and were initially viewed as more refined, either because they were seen more as medicine—as in the case of tea—or because they were simply far beyond the financial means of most Europeans—as in the case of the pineapple. Over the course of the period, however, increasing demand for tea and the challenge of growing a pineapple in Europe altered how these foods were seen. Tea, like coffee, became

a symbol of sociability, but it was held to be more refined and reserved, more civil. The pineapple remained a symbol of luxury and extreme wealth, and with it refinement.

1. Ideate

What are the core research questions?

How did the views on foods shift over the course of the long eighteenth century?

What are other more precise, relevant questions?

1. In texts dedicated to specific foods, are there any particular themes or topics, or places or types of items that stand out?

2. Are there recurring phrases or terms across numerous sources that might indicate changes in views on these foods and their relationship to civility?

Do the texts themselves change or, to put it another way, do discussions of foods shift from one type of text to another over the course of the period? What might such shifts tell us about the views of food and civility or manners?

Thinking about Methodology & Specific Tools

- **Topic Modeling**: this tool allows the user to see if there are any themes or topics that cut across a collection of texts.
- **Ngram**: this tool allows the user to track different kinds of phrases or terms that might occur around a food, essentially a kind of collocation.
- **Named Entity Recognition**: this tool allows the user to find references to additional items, such as other foods, places, and individuals.
- **Clustering**: this tool allows the user to see how texts discussing a food, or food and civility, might be similar with others or not.

2. Build

Steps

2.1 Searching

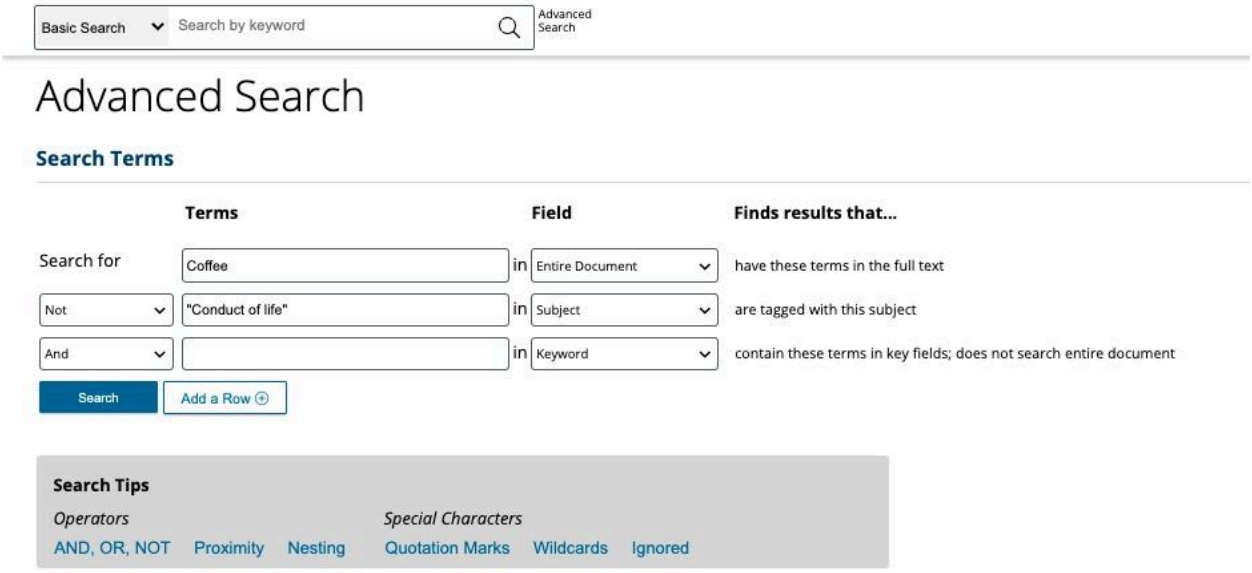
The content sets were constructed around three distinct 18th century food types: coffee, tea, and pineapples. The archives used differ between the three separate content sets. The documents come from a variety of collections within archives such as Eighteenth Century Collections Online (ECCO), British Library Newspapers, The Making of the Modern World, and Nineteenth Century Collections Online (NCCO), to name a few.

- Search Terms: Coffee
- Selected Databases to Search: Amateur Newspapers from the American Antiquarian Society; American Fiction, 1774-1920; British Library Newspapers; Crime, Punishment, and Popular Culture, 1790-1920; Eighteenth Century Collections Online; Indigenous Peoples of North America; The Making of Modern Law: Foreign, Comparative, and International Law, 1600-1926; The Making of Modern Law: Primary Sources; The Making of Modern Law: Trials, 1600-1926; Sabin Americana: History of the Americas, 1500-1926; Seventeenth and Eighteenth Century Burney Newspapers Collection; Seventeenth and Eighteenth Century Nichols Newspapers Collection
- Search Limiters - by publication year(s): Between 1650-1800
- Search Limiters - by content type: *Monographs*
- Search Terms: tea, tay, tey, chai, poetry
- Selected Databases to Search: Eighteenth Century Collections Online, The Making of the Modern World, Archives Unbound, Sabin Americana: History of the Americas, 1500-1926, Nineteenth Century Collections Online
- Search Limiters - by publication year(s): Between 1650-1800
- Search Limiters - by content type: *Manuscripts, Monographs, Newspaper*
- Search Terms: pineapple, pine apple, pine-apple, anana, ananas, king apple, pineaple,pinapple, pineable, pyneable
- Selected Databases to Search: Eighteenth Century Collections Online, American Historical Periodicals from the American Antiquarian Society, British Library Newspapers, Nineteenth Century Collections Online, The Making of the Modern World, Seventeenth and Eighteenth Century Nichols Newspapers Collection, The Making of Modern Law: Legal Treatises, 1800-1926

- Search Limiters - by publication year(s): Between 1650-1800
- Search Limiters - by content type: *Manuscripts, Monographs, Newspaper, Periodical*

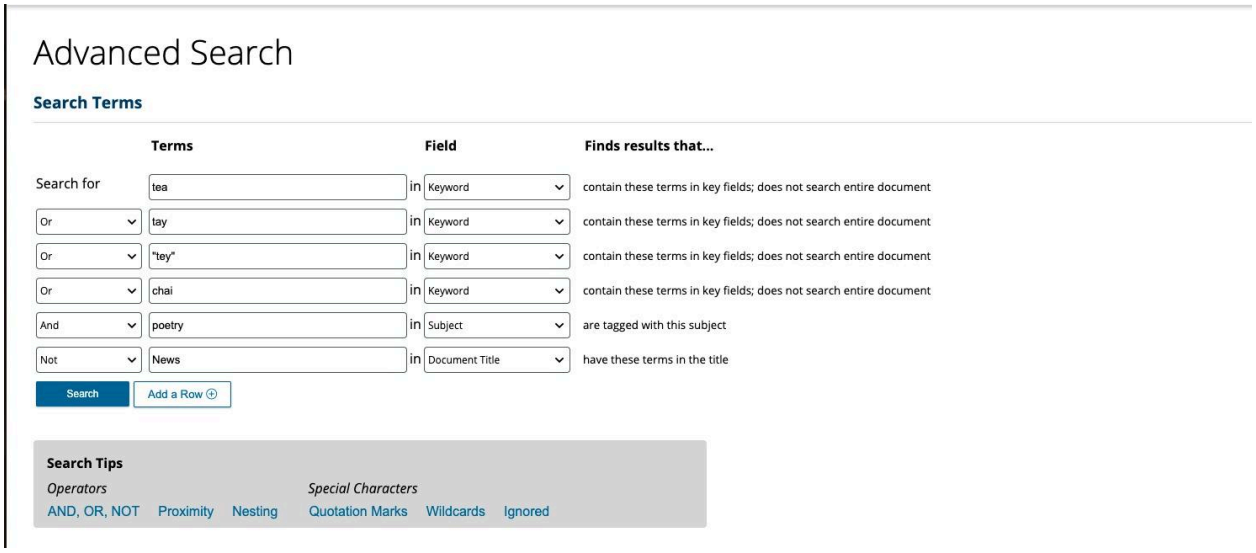
Statistics & Info

Coffee



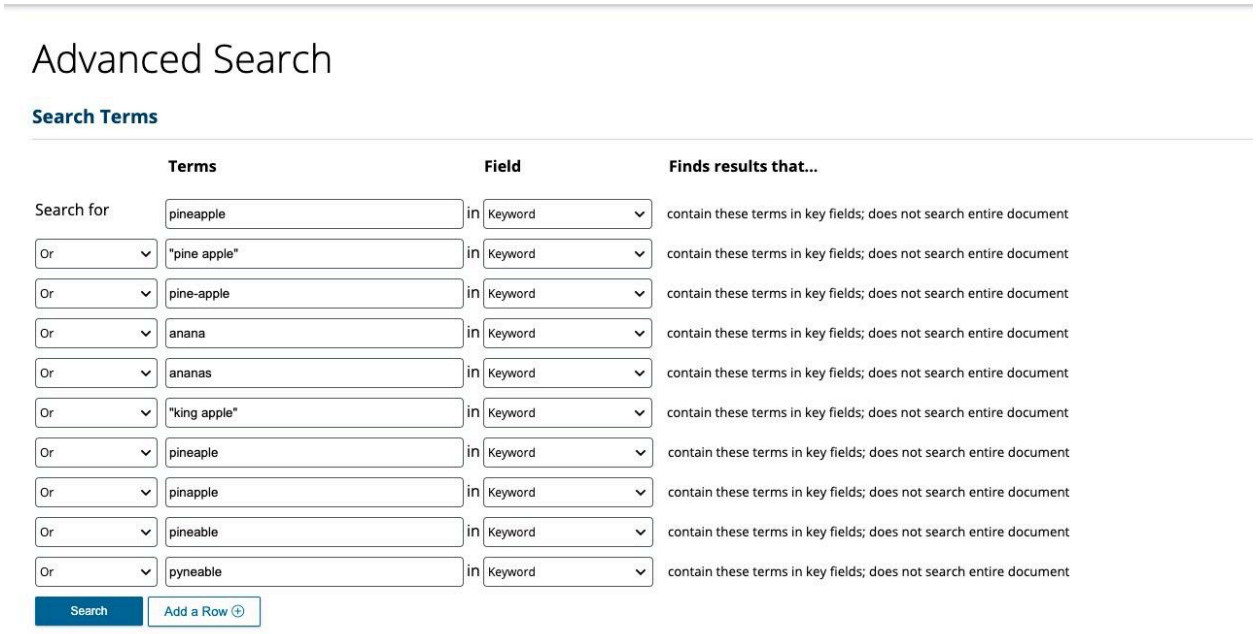
Content Set Name: 1650-1800 Coffee - Manners & Customs

Tea



Content Set Name: 1650-1800 Tea

Pineapple



Content Set Name: 1650-1800 Pineapple

Thinking about Methodology

- Topic Modeling: this tool allows the user to see if there are any themes or topics that cut across a collection of texts.
- Ngram: this tool allows the user to track different kinds of phrases or terms that might occur around a food, essentially a kind of collocation.
- Named Entity Recognition: this tool allows the user to find references to additional items, such as other foods, places, and individuals.
- Clustering: this tool allows the user to see how texts discussing a food, or food and civility, might be similar with others or not.

None of the tools required specific content sets. It was easier to create Cleaning Configurations that removed all punctuation, set all characters to lowercase, and also removed extended ASCII characters. Numbers were also removed. It is also recommended to add letters A-Z to the stop word list to remove random characters that would appear in visualizations like ngrams.

3. Clean

The content sets were both cleaned to remove punctuation, as well as numbers, and special characters, including extended ASCII characters.

4. Analyze

Tools Used

4.1 Topic Modeling

It seemed best to cast a wider net in part to see what kinds of words appeared in the models created by the MALLET software that powers the tool. Requesting more words than the default, and double the topics produces finer grained topics, in reflection of the size of the content set. This sample project utilized 15 word topics and 20 topics.

4.2 Sentiment Analysis

This tool has no settings other than selection of the cleaning configuration.

4.3 Ngrams

Like Topic Modeling, it seemed worthwhile to go beyond the default settings given the size of the content set. Therefore, the threshold for the number of times an Ngram had to appear to be considered useful was raised to 4. Similarly, the minimum Ngram size was set to 2 (bigram) and the maximum size to 5 so that the search would find collocates rather than just single words. These settings translate into a search for “Ngrams of between 2 to 5 words that appear in documents at least 4 or more times”.

Understanding Results

This project’s results are messy and full of considerable noise because of the highly variable spellings or orthography inherent in early modern English. The results highlight two issues to address: how to build content sets effectively when the terms being examined are extremely variable; and how does that variability shape the ability to run meaningful analysis.

An important indicator of the kinds of problems faced by this project is the Optical Character Recognition (OCR) error where the early modern “s” is actually recognized as a lower case f; in this instance it is quite problematic, as a related concern of this era is the rise of “fame” or celebrity. When the word “fame” appears in the topic models and Ngrams, there is no clear way of distinguishing between the legitimate presence of the word “fame” from an erroneous OCR “same” with a long “s”.

Topic Modeling

Some of the topics that arise from the analysis match themes of interest, or suggest some proximity. Most notably, none of the Coffee topics mention Coffee, so there is clearly an issue with the content set or the Cleaning Configuration. It will require revision.

Tea

- great, life, man, good,fame, men, virtue, god, reason, fee
- tea, leaves, chinese, use, green, fame, plant, china, tobacco, virtues

Pineapple

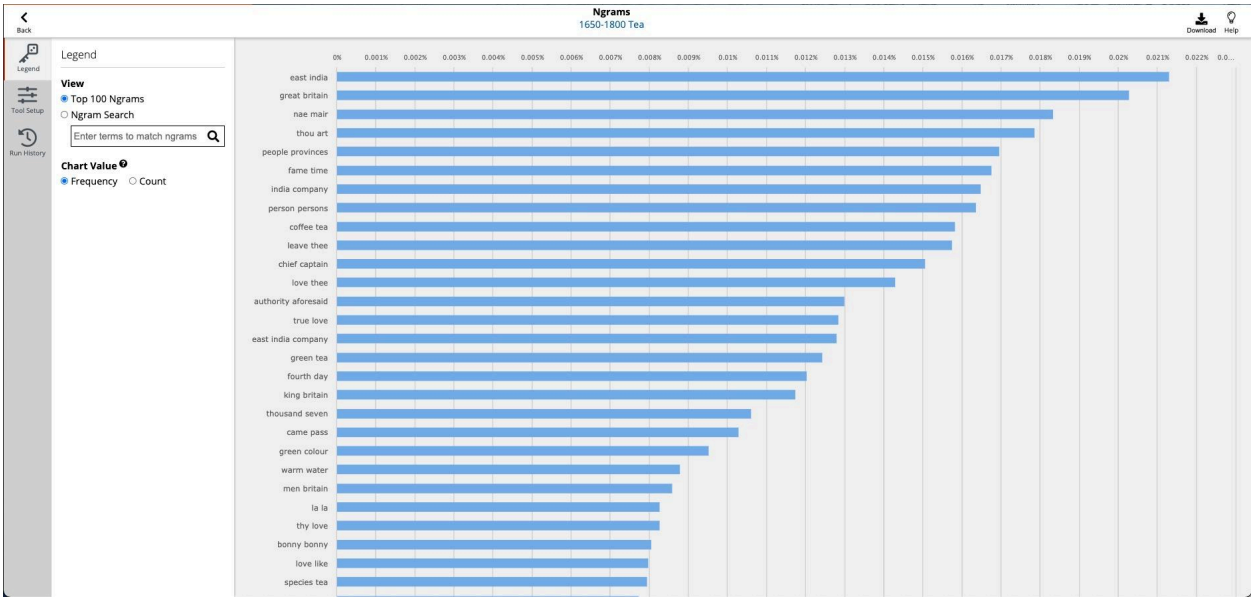
- pine, heat, tan, generally, hothouse, fruit, plants, time, pots, great
- plants, earth, roots, ground, fruit, plant, time, planted, leaves, fame

Coffee

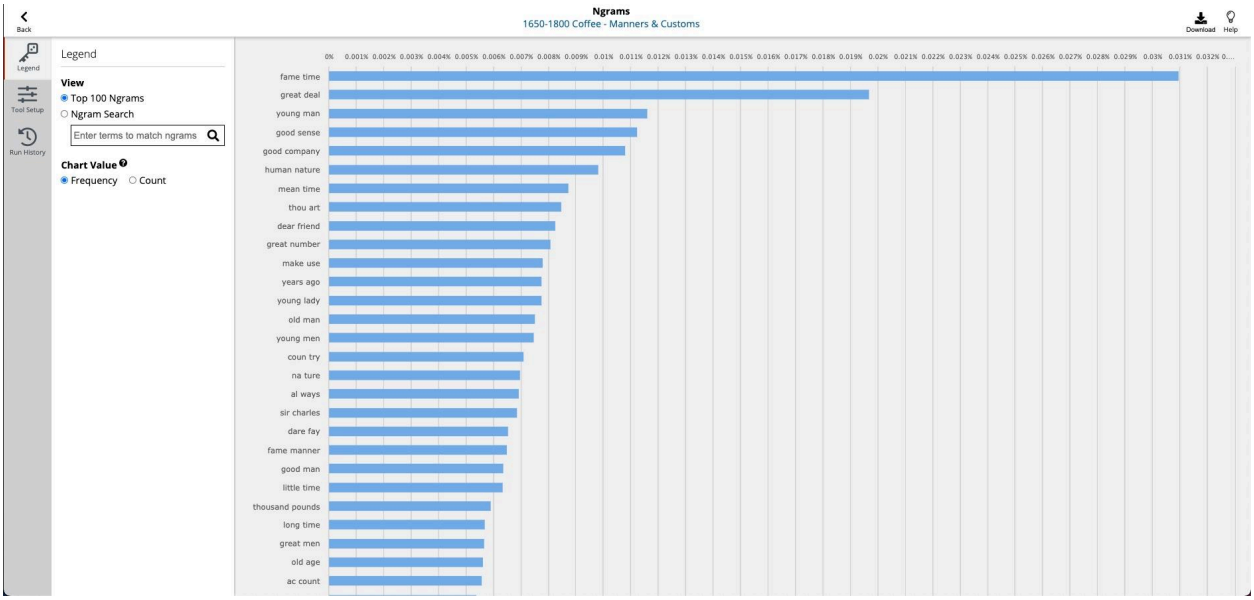
- god, hath, lord, great, things, good, unto, evil, men, power
- great, paper, fame, lady, ladies, town, letter, author, time, gentleman

Ngrams

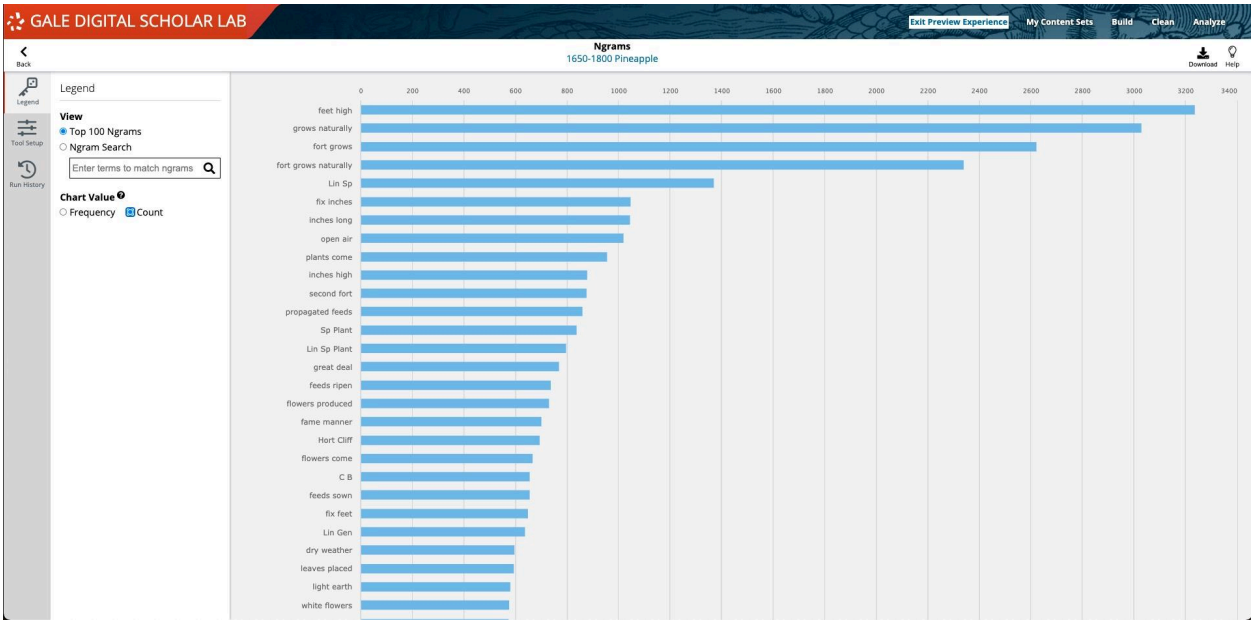
Tea



Coffee



Pineapple



5. Interpret

Read more about ways you can expand this project with iteration, research questions, and analysis.

Research Outcomes

Without a clearer approach to cleaning, it is very difficult to answer the original research questions. There are some hints, of course; for instance, the presence of the word “young” and types of people (men, women, lady, etc.) occur with greater frequency in the Ngram results of the Coffee content set than in any of the other content sets. Both Tea and Coffee mention ethics and morality. Coffee’s Ngrams note both good and evil, liquor, and common sense, while Tea’s Ngrams focus on the otherness of the product itself. Similarly, the Pineapple content set is focused heavily on gardening questions, which corroborates the interest among the ultra-rich of the early modern era in growing the fruit in Europe.

New Questions

How can interest in new foods be examined through advertisements and shipping notices? While these texts would rarely discuss civility, the growing accessibility of tea and coffee and other goods like chocolate are an important indicator of their acceptance and prevalence during the period. Building a content to trace how frequently these goods appear in these genres might be a way of providing some context to the other content sets.

The questions that arise out of this project also focus more on method than on intellectual content.

1. How, and to what extent, does orthographic normalization affect computational text analysis?
2. How can Gale Digital Scholar Lab be used in cases where analysis requires deep and extensive re-editing and reshaping of texts to get to a point where its tools are useful?

Thinking Critically About Research

Content Set Building

The content-set-building for this comparative project began with a broad timeline, as the shifts to be traced occurred over the long term. The year 1650 is something of a watershed moment in European history. In the English context, it follows the execution of Charles I by the English Parliament following the English Civil War; on the continent, it follows the end of the Thirty Years' War in 1648. At the other end of the time frame, the year 1800 comes a decade after the start of the French Revolution and coincides with the rise of Napoleon as the First Consul of the French Republic. These boundaries, in a cultural context, are porous, only outer limits for building a content set that focuses on publications of the late 17th through the end of the 18th centuries. The Eighteenth Century Collections Online (ECCO) also ends at 1800.

Given the variable spellings of the keywords themselves, it took several searches of each content set to dig through and find the right documents. Each of these foods was new; and in the case of the pineapple, their names weren't necessarily stable. The content sets were built using not only variable spellings but also different names for each food, in the case of Tea and Pineapple. Tea was spelled "te," "tey," "the" (which caused obvious problems both with the search and with text cleaning, and so had to be left out as it's not an effective search term), but also could be referred to as "cha," "chai," among other terms. Pineapple often appeared as two words, but also was known as a king fruit, and especially early on in the period, known by the name "ananas," which, having indigenous origins, became the preferred name in many European languages, like French and Spanish.

Given the highly variable nature of the ECCO archive (which contains every significant English-language and foreign-language title printed in the United Kingdom between 1701 and 1800), content-set-building also had to account for a wide array of genres. It quickly became clear that certain texts would likely be unsuitable, even though highly relevant. While advertisements and shipping arrivals contain a wealth of information on new goods such as tea and coffee, they are not a good fit for questions of civility other than to suggest that such goods were becoming more accessible. These texts aren't commensurate or comparable with those that discussed civil comportment or other kinds of sociability, making their inclusion in the content sets problematic. They're shorter, contain sales information, and are more numerous than the comportment texts, causing them to

overwhelm the longer, usually denser discussions of new foods as signs or instruments of social finesse.

Poetry was another highly problematic genre. Whether in refined poems or lowbrow lyrics of ballads, new foods received metrical treatment during this era, often because of the sort of people consuming them: the tea and coffee drinkers or the ultrarich aristocrat or businessman pining for the fruit in his hothouse. Such works, as ideal as they are for research questions, can't be isolated from their larger collections in the Digital Scholar Lab. The inclusion of a document containing several poems because one might discuss the topic would be highly problematic, as the other poems would alter the results of the analysis.

For these reasons, the content-set-building focused on texts that discussed new foods specifically and were stand-alone or monograph works. In the case of Pineapples, this resulted in a content set that focused almost exclusively on gardening and the problems of hothouse pineapple production.

Thinking about Limitations and Iteration

As useful as these results might be, there are limitations to the kinds of cleaning and analysis that can be done with the Gale Digital Scholar Lab.

• Currently, there is no method within the Gale Digital Scholar Lab to compare all words against an English dictionary in order to identify problematic optical character recognition (OCR). Iterating through Cleaning Configurations using Topic Modeling is a sound method for finding problematic words, but it's a time-consuming process. Replacements can be made easily, allowing problematic OCR to be fixed, but there's no method for finding all instances of misspelled words.

• This project didn't build content sets using actual cases, nor were they built following a close reading of the documents included in the sets. A more precise content set could be built by determining (following examination of each document) whether or not it was appropriate to include in a content set focused on the specific parameters of the project.

• The project didn't fully consider the difference between raw numbers, or "counts," and statistical measures as distinct ways of thinking about significance. Although the Topic Modeling output allows the researcher to examine counts, the Latent Dirichlet Allocation

Method used by MALLET is a kind of prediction of the likelihood of words appearing with each other. It's suggestive, in other words, of something significant. The Ngrams, in contrast, are raw counts across the content set. Having more documents or longer documents (i.e., documents with more words) would increase those counts. Numerical presence, however, doesn't always translate into intellectual significance or meaningfulness.

These limitations raise a fundamental question about how a platform like Gale Digital Scholar Lab transforms and alters texts. The Cleaning Configurations within the Lab are designed to ready the content set for analysis by making minor edits to texts that remove features and content that might obscure analytical results or cause problems for the tools themselves. This work can be considered "corrective" or "preparatory," rather than "editorial" or as "revisions." Making most of the ECCO texts functional for computational analysis, however, doesn't involve "corrective" or "preparatory" changes, but rather "editorial" work. Conceptually, this is quite distinct, as it means much more engagement with the text by a researcher who needs to create a new version rather than just prepare an existing one. Importantly, preparatory cleaning will still need to take place before analysis of new versions of the texts.

Beyond the Lab

All of the tool outputs can be downloaded as images to use in PowerPoint slides or embedded in web pages or other ways for presentation.

New Visualizations

It is also possible to download the data that powers the visualizations as comma-separated values (CSV) or JavaScript Object Notation (JSON) files, allowing researchers to create and format their own visualizations. With the proper skills, it's possible to collate or create new visualizations that may combine outputs from similar visualizations into one, allowing the researcher to compare and contrast in new ways not available in the Digital Scholar Lab tool.

The Topic Modeling tool downloads are especially rich with possibilities for new visualization. The Topic View download is large and contains results for each document and

measure for the tool—much more data than the Topic Model visualizations can currently display. Programmers can treat this as the ideal place to start exploring the data created by the Digital Scholar Lab using other tools and visualization designs.

Refining the Content Sets

This project is limited in its development because of the challenges of cleaning and the limitations of OCR for historical text. The Gale Digital Scholar Lab, however, provides researchers with the means of downloading entire content sets so they can be used elsewhere. This functionality allows the user to resolve some of the impediments related to this project, as early modern English is one of the ”hot spots” of computational text analysis when it comes to resolution of OCR quality, editing OCR-created text, and handling highly variable spelling or orthography.

Two tools exist for helping to standardize the spellings in early modern English texts: MorphAdorner and VARD 2. Both use similar statistical measures to determine whether a word is, in fact, a different word. For instance, these tools can change double “vv” into a “w,” adjust spellings of words like “againste,” “agaynst,” or “ageinst” to a standardized “against.” They also permit custom replacements if the selected texts have consistent errors due to misrecognized characters or gaps. The end result is the creation of new versions of the texts—essentially modern editions—that have more consistent, standardized spelling. As outlined in Gale Digital Scholar Lab and Computational Text Analysis guide, computational text analysis isn’t necessarily dependent on standardized spelling per se, but most literary and historical research questions and methods relating computational text analysis in distant reading require it.

To put this another way, standardized spelling allows platforms like the Gale Digital Scholar Lab to count or identify the same words and provide the basis for the statistical analytics that forms the usual ways of distant reading a large corpus of texts. By turning words like “againste,” “agaynst,” or “ageinst” into the standardized ”against,” it allows computational analysis to proceed in a meaningful way. Of course, the tools will work fine if the spelling isn’t changed or standardized, but the software will treat the different spellings as different words. This highlights how important it is to consider whether standardized spelling will add or obscure what the researcher is interested in investigating. For instance, if a corpus retains texts with ”againste,” ”agaynst,” ”ageinst,” and ”against,” tools like Topic Modeling

and Ngrams, which examine how many times (in various ways) a distinct word appears and in what context, will treat these as separate words, even though human readers will know they're the same word.

The end result might be that certain topics won't emerge from the analysis, and Ngrams will be different. There are scenarios where this might be a desired outcome; for instance, the word "peece" might mean "piece" or "peace." Standardization would require contextualization, and it might be useful to leave the original spelling as a way of tracing how this spelling itself appears in a corpus. If the research questions focus on different spellings and the nature of a language itself, standardization of spelling offered by these tools would remove the very thing being researched, and so it wouldn't be appropriate. In such instances, the Digital Scholar Lab's tools are ideal for this sort of research. Most text analysis, however, focuses on recognition of features across a corpus, which requires a certain level of spelling standardization to make the texts themselves comparable or "commensurate" (to use a technical term).

Similar Projects

These kinds of issues extend well beyond the Lab and are areas of research in their own right. Other projects include:

• The Early Modern OCR Project (eMOP), based out of Texas A&M University, which is focused on producing versions of early modern texts that can be used in computational text analysis. Gale is one of the partners in this work, which involves the Eighteenth Century Collections Online (ECCO), the source archive for the content sets for the example project on food and civility.

• Another project worth examining is the Distance Reading Early Modernity subproject of the now-concluded Early Modern Conversions project, based at McGill University. (DREaM) used VARD 2 to standardize the texts from the Early English Books Online (EEBO) archive from the University of Michigan's Text Creation Partnership project, of which Gale is also a contributing partner.