# Introducing Digital Text Analysis in First-Year Writing: “Data” Evidence and Source Limitations

Keywords: Voyant; corpus; civics; general education

*This assignment describes a text-analysis project designed to aid critical data and research literacy in first-year writing and general education courses.*

David S. Busch, Cuyahoga Community College

## Context: A Historian Teaching First-Year Writing

Digital historians are increasingly paying more attention to the methodology behind their computational and quantitative analyses, including making decisions about how to deal with OCR (optical character recognition) errors, how to organize corpora text files, deciding which algorithms to use and why, and choosing how to interpret visualizations of those calculations (Robertson and Mullen 2018).

As a historian teaching in a first-year writing program, my challenge was to translate discussions within my field of history to a first-year writing course with students from mostly STEM (science, technology, engineering, and mathematics) disciplines. In preparation for the course, I asked myself: how could I introduce students to the important debates about evidence and methodology in my field without teaching a “digital history” course? To do so, I drew on the work of library and information scholar Johanna Drucker. Humanities scholars, she argues, need to reconceive “data as capta.” “Capta,” she continues, is the “situated, partial, and constitutive character of knowledge production, the recognition that knowledge is constructed, taken, not simply given as a natural representation of pre-existing fact” (Drucker 2011,2). This conceptual framework was especially important in teaching first-year writing at a STEM school, where many students hold the belief that “if it can be measured, it is true.” In this light, one of my main learning goals of the assignment was to teach students not only how to utilize digital tools to navigate a large corpus of sources, but also to understand the constructive and contextual nature of sources and, by extension, their limitations.

## Tool and Source: Using Voyant to Analyze Contemporary Reporting on Student Activism

Voyant is a free, web-based application for performing corpus and text analysis. Designed for general audience use, the platform provides a range of interpretive tools that allow users to analyze word frequency, proximity, and trends across a text or corpus, as well as produce different visualizations, including word clouds, graphs, and maps. Voyant’s user-friendly interface provides an accessible tool to introduce computer-assisted digital text analysis.

The theme of my course was “Civic Agency and the University,” which took inspiration from the theoretical field of civic studies and my work in the history of student activism in American higher education. The assignment was the students’ opportunity to contribute to the scholarly conversation on civics in American higher education. In this assignment, students critically analyzed a corpus of newspaper reporting on student activism from 2017 to 2020 as an entry point into exploring contemporary student activists’ political visions of education and the university.

### Pedagogical Strategy: Digital Research as “Lab” Research

The “Digital Civic Labs” assignment was a multi-session, group-based project. We held two introductory sessions with the digital resource center on my campus, where we introduced the students to the practice of digital scholarship and the set of decisions involved in the construction of databases and corpora files. The next five sessions were “digital labs,” which were broken down between a 15–20 minute overview of a particular tool within Voyant and group experimentation with that tool. The sessions were: Framing Your Problem and Purpose; Cleaning the Data (Stop Words); Search Terms and Close Reading; Experimenting with Visualizations; and The Data (or Visualizations) Suggests.

For each session, I also provided students with a set of guiding questions (with methodology):

* Compare and contrast the sentiment analysis corpora. What do you notice in comparing “negative reporting” versus “positive reporting”? What does comparing the two reveal about student activism?
* Compare and contrast the top publishers’ corpora. How did newspapers report on student activism? Were there differences across the newspapers? (NOTE: You can run this experiment with the sentiment analysis.)
* Compare and contrast the political groupings. What was the nature of student activism the past three years according to this database (that is, what were the political motivations, interests, and/or tactics)?
* Visualize each corpus using DreamScape and examine the links in context. Then, consider: To what extent and in what ways (how) was student activism linked over the past three years?
* Examine the corpora (chronology) using the trends tool. What trends emerge when exploring student activism from 2017–2020?

I framed the submission report as akin to a “lab report” for two reasons: a) it introduced students to a genre of writing that they would have to do in their respective fields; and b) it encouraged them to think critically about the relationship between method, evidence, and argument. The final report included the following sections: an introduction that framed the scholarly conversation (on activism and civics) and set up the analysis; a methods section where studentsdiscussed the set of methodological decisions they made, such as: what corpora they chose to focus on and what methods (stop words, search terms, visualizations) they utilized; a findings section that provided the visualizations and discussed what the visualizations demonstrated (based on their analyses); and a conclusion and limitations section, where they summarized their argument and discussed limitations of the digital experiments.

Reflections: Critical Insights and Next Steps

In reflecting with the students on their digital processes and reading their course evaluations, I was impressed by the students’ critical view of digital text analysis. They found the project intellectually challenging and intriguing, while also being skeptical of what could be argued based on distant reading and the sources. It is on the latter point (evidential skepticism) where I see success. Indeed, students come into the university with the assumption that evidential analysis is merely finding sources (primary or secondary) that support one’s argument. This assignment, especially the report sections that asked students to consider methodology (rationale) and limitations, challenges students to critically engage with the nature of evidence, in particular corpora data.

Of course, I want to be careful here and not overstate the apparent pedagogical “success” of the assignment. In many ways, it is in my post-reflection on the assignment that I, as the instructor, am seeing the need to make “capta” a more explicit framework to the project. For example, instead of sharing with students the already-made corpora files, I could involve the students in the process of developing the corpora files. In this scenario, students would design corpora files, with a reflective component akin to what Lindsay Poirier (2020) defines as an ethnography of datasets, that encourages them to consider the set of choices they make in building the corpora file. By designing the assignment in this way, I could help students become more familiar with the idea of “capta,” a familiarity they could apply in other courses where data is the primary object of analysis. Relatedly, as I plan to do in my next iteration of the course, I can also provide students with historical sources that are inherently limited. While the Google Alerts files were already digitized, having students engage with a set of primary sources will force them to experiment with OCR and transcribing—processes that also raise questions about the nature of data. Both scenarios enable students to further engage with “capta” while increasing active learning in the course.

In an increasingly digital world, first-year writing courses and related general education courses need to more effectively engage students with the “virtual stacks” (Blevins 2019, 265). This includes not only how to navigate those digital stacks, but to also critically engage their limitations, including text search and the overrepresentation of Anglophone material in digital archives (Putnam 2016). As teachers and scholars, we need to expand our toolkit to teaching the virtual stacks, especially in a first-year general education course that extends beyond simply learning how to conduct Boolean searches. I hope the project assignment provides a roadmap for those interested not only in incorporating digital inquiry into first-year writing or general education courses, but in doing so in a way that engages students as critical users of digital tools and data and as co-creators of knowledge.

References

Blevins, Cameron. 2019. “Intro to the Stacks: A Tour of the Virtual Stacks.” *Modern American*

*History* 2, no. 2: 265–268. https://doi.org/10.1017/mah.2019.16.

Drucker, Johanna. 2011. “Humanities Approaches to Graphical Display.” *Digital Humanities*

*Quarterly* 5, no. 1. http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html.

Poirier, Lindsay. 2020. “Ethnographies of Datasets: Teaching Critical Data Analysis through R

Notebooks.” *The Journal of Interactive Technology & Pedagogy* 18. https://jitp.commons.gc.cuny.edu/ethnographies-of-datasets-teaching-critical-data-analysis-through-r-notebooks/.

Putnam, Lara. 2016. “The Transnational and the Text-Searchable: Digitized Sources and the

Shadows They Cast.” *The American Historical Review* 121, no. 2: 377–402. https://doi.org/10.1093/ahr/121.2.377.

Robertson, Stephen and Lincoln Mullen. 2017. “Digital History and Argument.” White paper,Roy Rosenzweig Center for History and New Media. https://rrchnm.org/argument-white-paper/.

Appendices

[Appendix A – Digital Lab Assignment (Detailed)](https://docs.google.com/document/d/1fxzs_a7etdP7sfwwUNHL4nXgT2lwiJqFfYHJggGK72Y/edit?usp=sharing)

[embed https://docs.google.com/document/d/1fxzs\_a7etdP7sfwwUNHL4nXgT2lwiJqFfYHJggGK72Y/edit?usp=sharing]

Appendix B – Class Archive

[Student Activism and Protest: A Text Corpora Archive](http://dharchive.davidsbusch.com/)

Appendix C – Feedback and Reflections on Students’ Digital Lab Reports

In many of the introductory setups to the arguments, students were attuned to the nuances and limitations of their data and analysis.

Example 1: “Using Voyant, a text analysis tool, we examined a variety of movements and engagements proposed and carried out by students’ own accord in the past several years. Our initial findings were multifaceted and showed a strong **correlation** between student activists and their education and government and police involvement.”

Example 2: “Using Voyant, a text analysis tool, we examine how reporters talk about student activism, allowing us to **implicitly** determine how students define civics and find their role as civic agents.”

In the examples, I bolded two key words: correlation and implicitly. As rhetorical moves, they could be understood as hesitation in making arguments—a big conceptual leap that many composition instructors recognize. In the context of this assignment, they also demonstrate a critical take on what one can do with the digital text corpora as a form of evidence. Correlation, as we know, suggests relationship. As such, the use of “correlation” makes clear that students did not want to make a causal claim based on the digital analysis. Indeed, for many, their analysis was the basis for future research. In the second example, the word “implicitly” also suggests that students were also aware of the interpretive work of the text data—that to critically analyze the sources, they had to juxtapose the patterns and themes with the frameworks associated with the scholarship on civics. In both ways, student arguments rested upon a critical evaluation of the sources themselves—that is, what type of claims can and cannot be made based on the given corpora text files.

Indeed, on this point, students used their discussion of “stop words” to make clear that the distant reading of the text corpora, via Voyant, represented a range of methodological considerations. “Stop words” in Voyant is a set of words that a user must decide to exclude from the results of a tool. In the discussion below, students highlighted what they excluded and why, a process that encouraged them to recognize the context of the data itself or how the initial file was constructed.

If we go straight into analysis with the corpora as is, there would be many instances of ineffective data cluttering the most frequently used word cloud, the links visual, and the collocates table. To distill out the noise, we implemented a stop word list, which removes all instances of select words from the data. In determining whether a word should be on the stop word list, we first discussed whether the word is inherent to the nature of the articles selected with the google alert (words such as “protest”, “activism”, and “students”), and then we discussed whether the word is useful for analysis (words such as “says”, “reuters”, and “mr”). Though we considered focusing on other corpora categorized by top publishers, sentiment, and politics, we decided on chronology because we were interested in how articles talked about student protests as they happened and in the years following. Specifically, we observed reporting on protests in Hong Kong, India, and Afghanistan while they happened, and as years passed.

This same group also presented a rich discussion on their methods; specifically the ways they used certain tools to understand the text data from different analytical angles.

*Procedure*

To begin analyzing the chronology corpora we generated a list of stop words using prior knowledge of the corpora and the Cirrus word cloud. The first set of words generated by the cirrus included words and variations like “student”, “protest”, and “say”. Since this corpus was specifically constructed to be about student activism and words pertaining to students (e.g. campus, youth, etc.) can be removed as stop words. Put another way, since the subject of the corpora has been established, therefore, any words pertaining to the subject add no new information and cannot help further the extrapolated knowledge. The same explanation can be given for variations of the word protest: since we know the corpora is regarding activism, it is obvious that that will involve protesting and so it can also be removed. Labeling “say” as a stop word is a more mechanical reason: since the corpora consist of news articles, it will contain action words to utilize quotes and other opinions and so is also relatively useless to the lab.

The next step was to use Links to form connections between the remaining words to understand the context they are mostly being used in without having to scan through multiple articles– as it would be nearly impossible to go through most of the 1080 documents. The Links tool made associations between the most common words and other smaller words and also gave a visual representation of how often those connections were made in the documents relative to other connections. From these observations, we began to derive inferences about the content of the articles.

The Collocates tool was then used to ensure that these inferences can actually be made and are substantial. The tool shows how many times high-frequency words are associated with other words. This is particularly useful because it provides context on how the high-frequency words are being used in the documents. Some words may have either a positive or a negative connotation which can be determined based on adverbs and adjectives used with those words

To address some of the limitations, I asked students to substantiate their distant reading with “close readings,” or what Voyant defines as “drill downs.”

*Drill Down*

We utilized the drill down tool to explore the corpora’s articles with a close reading of phrases within them. After analyzing our cirri of word maps for each group, we chose to focus on the words “right(s)” and “government,” since they were present in all four groups. Other words were larger and, thus, more prominent within one or two of the groups (such as specific names of political figures and parties), but these were not consistently of interest in all four cirri.

From there, we noticed phrases such as “right to speech,” “First Amendment right,” “right to protest,” all of which concerned the reactions of universities to student protest efforts. When students “feel powerless—locked out of [the] decision-making process, bypassed by real governance—they turn to protest. [Protest] is often the only way to exert power and affect policies and practices that impact their lives and communities

Just as importantly, especially given my emphasis on the limitations of the data, students demonstrated a nuanced take on both components with a particular focus on the construction of the corpora text files.

*Limitations*

Distant reading, by nature, has a great many limitations. These limitations do not start with the concept of distant reading, however. They begin with the data collection and preprocessing before any analysis can be done. It is intrinsic to all data collection that some details will be missed, or that factors unbeknownst to anyone would twist the outcome. In addition to those, the initial dataset was limited by its search terms. To create the initial dataset, google alerts were used to scrape news articles containing the terms “young people protest” and “student protest.” Obviously, having only two terms limited the breadth of the net that was cast, but google’s system also seemed to have interpreted the terms to scrape a wide variety of articles, which adds another layer of uncertainty. Google’s search algorithm, and thus the Alerts scraping algorithm, is a “black box”: it is hidden from us by design. However, we also know the algorithm is iterated on constantly and thus likely changed to some degree during the collection period. This implies that the selection method and criteria changed, which means some articles may have been scraped or ignored at one point where they would not have been at another.

Of course, some students still fell into the trap of confirmation bias. For example, one group framed their process as one of seeking words to confirm their argument.

*High Frequency Terms*

The following word cloud images are examples of the visualization we utilized to select notable words for analysis. The most prominent terms are displayed in the largest font, but rather than simply picking any of the most common words, we chose those specific to developing our argument around civic engagement. So, while “Hong Kong” was present in each publisher’s collection, its significance lies only in that it is a current location for many student protests. However, words such as police and government are representative of student activism as a whole, regardless of subject matter or location. Additionally, education, though not a prominent word in these visuals, is an underlying concept related to student activism and civic engagement as a whole, tying all three terms together.

Text

Description automatically generated with medium confidence

Figure 1. Word clouds developed by students using Voyant, showing high-frequency terms in the corpora examined for coverage of Hong Kong protests.

## 

## About the Author

**David S. Busch** is a community-engaged historian and digital humanist. He currently works at the Jack, Joseph, and Morton Mandel Humanities Center at Cuyahoga Community College, where he is developing a new summer humanities / leadership program. You can learn more about Busch’s scholarly background, research and teaching on his personal website: [www.davidsbusch.com](http://www.davidsbusch.com)