Raphi Gold
5/7/24
Professor Haverals
Final Project

Proposing a New Tool for Student Journalists and Activists

"About 100 undergraduate and graduate students began a sit-in on McCosh Courtyard early Thursday morning, joining a wave of pro-Palestinian sit-ins across the country," begins the *Daily Princetonian*'s first coverage[1] of the "Gaza Solidarity Encampment" at Princeton. On the morning of April 25, 2024, when students bearing tents and supplies rushed into McCosh Courtyard at 7am to start their encampment, *Daily Princetonian* reporters arrived on the scene almost simultaneously, leaping into action alongside the activists. By noon that day, following an intense morning punctuated by two arrests, the protesters had established a sit-in complete with areas for art, prayer, supplies, food, and more. And adjacent to the protesters, student reporters had set up their own camp of sorts, a table labeled "Independent Press," stocked with snacks, electronic chargers, and at least two reporters at all times. Since then, the *Daily Princetonian* has been providing live updates detailing the status of Princeton's Gaza solidarity sit-in, taking on extra shifts and odd sleep schedules to thoroughly cover the protest. Universities across the country have established similar structures, with student journalists quickly becoming central voices in this outburst of protest. Reporters have put themselves in danger[2] to cover tense or violent moments, and many editorial boards have released statements[3] criticizing their administrations and defending students' rights to free expression.

Reporters and scholars have suggest that generally, activism and journalism sustain one another, and that the success of social movements significantly depends[4] on their coverage in the press. Activists spend time crafting media advisories and appointing press liaisons, while journalists keep close tabs on planned protests, often tipped by the activists themselves so they

[1] https://www.dailyprincetonian.com/article/2024/04/princeton-news-adpol-gaza-solidarity-encampment-launches-mccosh-courtyard (Accessed: 5 May 2024)
[2] https://www.washingtonpost.com/national/2024/05/02/campus-protests-student-journalists/1b0609b0-084\0-11ef-b186-090cb777107e_story.html?ref=upstract.com (Accessed: 5 May 2024)
[3] https://www.nytimes.com/2024/04/23/us/college-protest-editorial-voices.html (Accessed: 5 May 2024)
[4] https://www.jstor.org/stable/pdf/2675477.pdf?refreqid=fastly-default%3A38d4cf558ffc9eb75cb6106816a6fe20&ab_segments=&origin=&initiator=&acceptTC=1 (Accessed: 6 May 2024)

can prepare in advance. Student newspapers offer unique[5] time capsules into life on college campuses, and they have a particular flair for covering protest movements. When and where there are campus activists, there are likely campus journalists to go with them.

Using the archival databases of Ivy League daily newspapers as my primary dataset, I propose a project that prompts researchers to consider this storied relationship between press and protest through the lens of student journalism. Through this project, I compile and disseminate information about the history of activism on Ivy League campuses as described by their own daily student newspapers. This framework encompasses and engenders numerous sub questions. My project focuses on the following: What kind of language do journalists use to describe campus activism and how has that changed over time? What schools have had the highest concentration of protests? Can we track patterns, peaks, and troughs in college activism? What kinds of actions and tactics were most common, and which have received the most coverage?

To address these questions, I propose a cross-Ivy League daily newspaper distant reading-based investigation. I envision current campus journalists and activists using this tool to gain insights into their past that might shape their present and future. Activists can explore previous tactics used on their own campuses and whether they were successful in affecting change, which movements and strategies drew the most media attention, and how they were received by the student body and college administrations. Journalists might examine and critically analyze the language they have historically used to cover student protest and identify potential gaps in coverage, identifying instances of bias which they can work to eliminate in the future. These goals take on an extra layer of importance during this active moment for college protest and reporting across the country.

As members of a fast-moving field which prioritizes public visibility and accessibility, Digital Humanities scholars are well-positioned to respond to these times by compiling and presenting data that will be currently meaningful. University of Maryland English Professor Richard Kirschenbaum captures this sentiment in a chapter[6] of his book entitled, "What is Digital Humanities and What's it Doing in English Departments?". Kirschenbaum writes, "Records drawn from a database and arranged in a particular order become a picture of modern life – but

[5] https://www.thefire.org/research-learn/role-student-publications-campus#:~:text=you%27re%20covering.-,Conclusion,larger%20community%20about%20relevant%20events. (Accessed: 4 May 2024)
[6] https://app.perusall.com/courses/introdh24/kirschenbaum_2012_what-is-digital-humanities-and-what-s-it-doing-in-english-departments (Accessed: 5 May 2024)

simultaneously an argument about this life, an interpretation of what these images, which we encounter every day, every second, actually mean." Through this project, I hope to draw an image of modern life through sketches of the past.

More specifically, my project uses digital humanities tools to compile a comprehensive database of Ivy League newspaper articles covering protests throughout history and store them in a contained website with advanced search options, graphs, and data visualizations. Within the database, a user can control their own experience by filtering for particular dates, words, topics, and colleges. For example, a user can find results for Vietnam in the late 1960s at only Harvard and Yale in order to more closely compare those particular colleges' anti-Vietnam War activism movements. Data visualizations such as word clouds help identify common protest tactics and the language used to describe them, and graphs indicate peaks and troughs in student activism or media coverage thereof. This allows viewers to go beyond the typical limits of an archival database, which often requires prior knowledge and a good deal of guesswork to discover important information. Here, viewers are presented with information about particularly active times in history or patterns to look out for, and can structure their searches based on that knowledge. I use the *Daily Princetonian*'s recent activism data piece[7] as a model to carry out these more technically complex aspects of my project.

One challenge in enacting this project is that some archival databases are more advanced and complete than others. Using *The Daily Princetonian*'s comprehensive Larry Dupraz Archives[8] as a model, I conducted a review of the archival databases of all eight Ivy League daily student publications. The Larry Dupraz archives includes scanned copies of all volumes dating back to its origin in 1876, accompanied by transcribed Optical Character Recognition (OCR) text on the side, with options to browse by title or date. It also allows for advanced search options, allowing interfacers to filter by date or phrase, conduct a boolean search, use advanced query syntax, and apply further search filters after the fact. Users can also download issues as PDFs and create their own accounts in order to save articles or volumes as named lists.

Using these Larry Dupraz criteria, I evaluated the archival websites of other daily Ivy League publications. I found that the papers of Yale, Columbia, Penn, and Cornell fulfilled the criteria, while Brown, Dartmouth, and Harvard's archives were lacking. While the *Brown Daily*

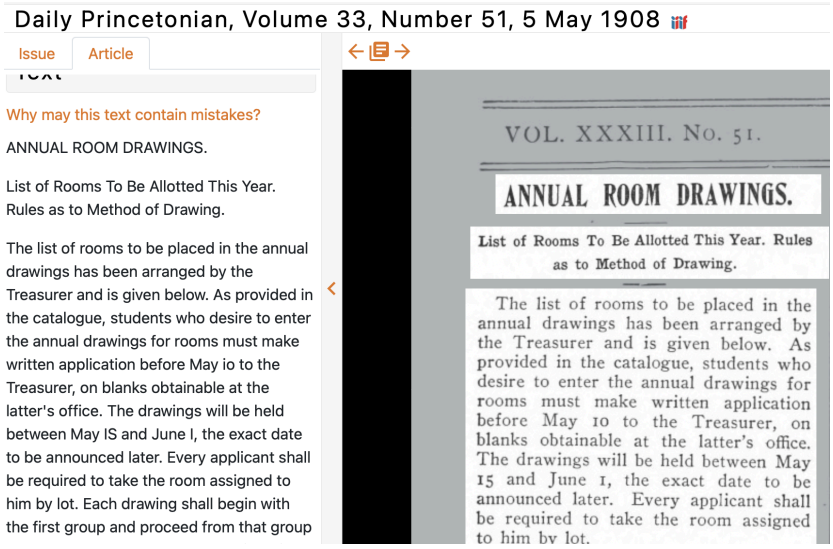
[7] https://projects.dailyprincetonian.com/activism-archive-project/ (Accessed: 5 May 2024)
[8] https://theprince.princeton.edu (Accessed: 6 May 2024)

Herald does provide scanned images of its volumes, and users can easily view the full metadata on any given volume, many advanced search options are unavailable, and OCR appears as code rather than text, posing a potential barrier to non-coders. The *Harvard Crimson*'s archives can only be found through general searches on the *Harvard Crimson*'s website and lack scanned raw images of the original volumes, which can only be accessed in-person at Harvard's Widener Library. Additionally, certain volumes are missing[9], and only issues from the current year are kept on-site, while earlier volumes must be requested through the Harvard Depository. The *Dartmouth* archives are similarly inadequate, with no scanned copies and only limited issues available.

Thus, the first step of my project entails ensuring all of the archives meet the Larry Dupraz criteria. I scan images of all past newspaper volumes, then transcribe text using an OCR tool such as Transkribus and run the transcription through CERberus to detect potential errors. In his article[10] "'QI-JTB The Raven': Taking Dirty OCR Seriously," Ryan Cordell uses a sample of OCR-derived text for Edgar Allen Poe's "The Raven" to illustrate the shortcomings of OCR and advocate for the importance of maintaining access to raw image data to avoid errors and ensure that scholars maintain relationships with original texts. Cordell believes that scholars are required to "Appropriately account for the sources they use" and acknowledge that, "Massive, errorful OCR archives necessitate close bibliographic and book-historical attention that both leverages their powers while historicizing their creation and use." While it would be unreasonable to independently verify this massive corpus of text against the raw image data in the archives, I propose offering users the option of correcting OCR errors on their own, as the *Yale Daily News* archives[11] allow for. The Larry Dupraz archives are also transparent regarding transcribed text, explaining "searchable text and titles in this collection have been automatically generated using OCR software. They have not been manually reviewed or corrected." I would include a similar disclaimer to ensure transparency and traceability. I would also clean my compiled data using Open Refine to ensure that the information is as organized and accurate as possible before proceeding.

---

[9] https://ask.library.harvard.edu/faq/82192 (Accessed: 3 May 2024)
[10] https://app.perusall.com/courses/introdh24/cordell_2017_q-i-jtb-the-raven (Accessed: 5 May 2024)
[11] https://ydnhistorical.library.yale.edu/?a=p&p=help&e=-------en-20--1--txt-txIN------#correcttext(Accessed: 5 May 2024)



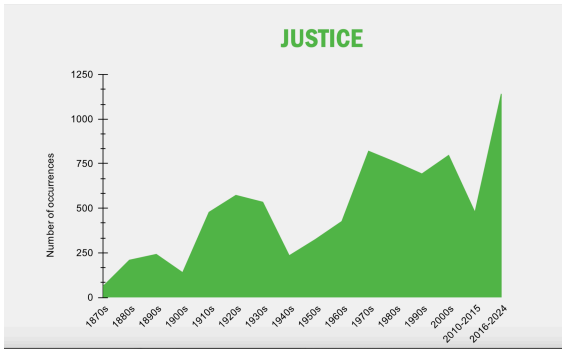
Example of original article (right) and OCR transcribed text (left), Larry Dupraz Archives.

Once the data is scanned, transcribed and cleaned, I extract all articles related to activism and protest using Gensim topic modeling tools to ensure that the articles remain focused on the subject and that words with double-meanings like "protest" and "movement" do not result in misleading articles being included in the database. Still, I recognize that it is impossible to entirely avoid these potential linguistic errors and will account for that in datasheets accompanying my dataset. These protest-related articles and their associated images and texts comprise my final dataset. Next, I use Voyant tools to conduct distant reading on the articles, pulling out topics such as "Vietnam" "Ethnic Studies" or "South African apartheid," and organize the data so that users can search by subject, date, or tactic. Through topic modeling, I also create groupings of words which comprise a common subject. I then use Gephi to form graphs mapping word frequency, and Cirrus to create word clouds to address my research questions.

These graphs and word clouds add a supervised learning element to this otherwise unsupervised project, serving to enrich the unsupervised learning experience by pointing users directly to data they might find interesting. In doing so, I combine supervised and unsupervised learning approaches, as scholars Dong Nguyen et al. suggest is a common strategy for "more elaborate research designs."[12] My model, which learns from labeled and unlabeled data alike,

[12] https://app.perusall.com/courses/introdh24/how-we-do-things-with-words-analyzing-text-as-social-and-cultural-data (Accessed: 5 May 2024)

includes both so that the two might inform one another. As Boyd-Graber et al. point out in their book *Applications of Topic Models*, "The observational nature of topic models is both a weakness and a strength." Researchers cannot always control the production of these documents, and in recognizing that, observational models can increase the potential for unexpected discovery. As Boyd-Graber et al. put it, "Topic models complement this type of carefully designed survey by unearthing issues or factors important to participants whether or not those issues were anticipated." Therefore, I include topic modeling and combine supervised and unsupervised approaches to ensure that even those questions I forget to ask are identified and considered.



Examples of word clouding and graphing, from the *Daily Princetonian* activism supplement.

In addition to topics and patterns, users might also be interested in further exploring the journalistic rhetoric historically used to describe activism. For this rhetorical aspect of the database, I use Stylo to conduct emotional stylometric analysis. I draw on Cynthia Whissell's method[13] of analyzing the emotional patterns of Beatles songs, employing stylometry to explore both the denotative and connotative meanings of words. Did journalists describe a particular protest negatively or positively? Was it a "riot," a "demonstration," or a "rally"? Was it peaceful or disruptive?

Of course, I do not suggest that stylometry is an entirely precise or reliable science. As Whissell herself points out, "Stylometry has been criticized for being a cold and impoverished method of textual analysis – one which studies words without studying their meaning." But by using emotional stylometry, I imbue meaning into this method of word-counting, combining

[13] https://app.perusall.com/courses/introdh24/whissell-1996-traditional-and-emotional-stylometric-analysis-of (Accessed: 4 May 2024)

quantitative and qualitative analysis. Journalists and activists alike can use this stylometric component for guidance when they conduct deeper archival research.

In order to ensure transparency about the expansiveness and limitations of the project, I also create publicly accessible datasheets for each dataset created along the way, allowing users to understand the process and think critically about my methodology. These datasheets work to identify patterns and bias in my research and align with data feminism researchers Catherine D'Ignazio and Lauren Klein's principle[14] of "make labor visible." Through making my labor visible, I call attention to the often invisible labor that goes into data projects, emphasizing process. As Todd Presner[15] writes in his Digital Humanities Manifesto, "Process [rather than product] is the new god," adding that Digital Humanities is a form of iterative scholarship which honors the quality of results, but also, "honors the steps by means of which results are obtained as a form of publication of comparable value." In providing easy access to datasheets and being transparent about potential errors and biases, I aim to highlight process in addition to results. As Presner maintains, "Untapped gold mines of knowledge are to be found in the realm of process." Providing users with easy access to my process allows for a greater understanding of how and why this dataset came to be, and can provide a model for digital humanists who wish to expand this project or create similar ones.

As with all digital humanities projects, it is also important to consider the practical matters of copyright and dissemination. Because this site stores vast tomes of intellectual property, I attribute authorship of all articles in the database to the original author, and credit all those who have worked to craft the archival newspaper datasets I use. I publish these attributions on the website, and they can also be tracked through my accompanying datasheets. As for dissemination, I store this project as a publicly accessible website for anyone to use. In order to reach a wider audience, however, I invite each Ivy League newspaper publication to publish a link to the project on their websites or within their own archives so that journalists and activists at those schools could easily find my project.

This project will serve student journalists and activists for decades to come, as well as any individuals curious about the history of activism on Ivy League campuses as portrayed through their student newspapers. Moreover, the depth and breadth of this project will allow it to

[14] https://www.youtube.com/watch?v=gulxU_hK4aY (Accessed: 5 May 2024)
[15] https://app.perusall.com/courses/introdh24/presner-et-al-2009-digital-humanities-manifesto-2-0 (Accessed: 5 May 2024)

extend far beyond the original website. On a smaller scale, the website can be periodically updated as new campus protest movements form in order to maintain relevance. More broadly, the project can inspire a more concerted effort to track journalism coverage of student protest movements across all college campuses and not just Ivy Leagues. I encourage future digital humanists to carry out similar projects which capture the history of social movements across local and national media in the United States, and even globally, transforming the way we see the relationship between journalism and activism.

Honor Pledge: I pledge on my honor that I have not violated the honor code on this assignment.

Signed: *Raphaela Gold*