UNITED STATES COPYRIGHT OFFICE

Long Comment Regarding a Proposed Exemption Under 17 U.S.C. § 1201

ITEM A. COMMENTER INFORMATION

Quinn Dombrowski is the co-President of the Association of Computers and the Humanities, the US-based professional organization for digital humanities, and a digital humanities practitioner since 2004.

qad@stanford.edu

ITEM B. PROPOSED CLASS ADDRESSED

Proposed Class 3(b) Literary Works – Text and Data Mining is addressed specifically in this comment, although the general principles apply to Proposed Class 3(a) Motion Pictures – Text and Data Mining as well.

ITEM C. OVERVIEW

Digital humanities – particularly in its text and data mining variety – calls to mind complicated statistical methods and fancy data visualization… at least, for those who don't do this kind of work themselves. If you've ever done one of these research projects, especially if you had to go about building your own corpus, what you probably think of first is the amount of time-consuming, tedious labor that can easily take up more than 80% of the total project time. We do digital humanities because it lets us ask and answer different kinds of questions about history, culture, and society – not because it's fundamentally faster. In fact, considering the amount of training that it takes to develop both disciplinary and technical proficiency, it can take more time than traditional humanities scholarship, but there are no shortcuts to answering the kinds of questions that interest us.

The current legal landscape for working with already-digital media in a format that uses technological protection measures (TPM, or digital locks) imposes penalties that make the process *even slower* than it inherently is already. This has significant real-world consequences ranging from what scholars are able to teach, to what the next generation is writing their dissertations on, as the foundation for the future of humanities scholarship in the US. But I'll return to that. Let's start by following along, step by step, with a text and data mining research project to look at all the steps that go into doing this work. The best way to do this is to illustrate with a concrete example: we're going to look at the petition to expand the DMCA exemption for text and data mining, as well as the oppositions, to examine how they deploy a set of key terms

(using word searching), as well as their rhetoric more broadly (using topic modeling, a probabilistic method that may be considered a form of "AI").

1. Corpus and research question

Corpus selection – choosing which works to include in the corpus you're creating to answer a research question – is itself time-consuming, requiring a lot of thought and consideration before you even invest the time in trying to acquire the materials.

This is a nice corpus for exploring things like lexical choice. For starters, the petition is of very similar length to the oppositions combined (
30,103 vs 27,989). When you're comparing texts, you generally don't want them to be of radically different lengths, because a longer document, by definition, has more words – so it'll also have more distinctive words, because there are simply more of them.

You can get around this a little by looking at word *frequencies* vs. word *counts*, but it gets more dubious the bigger the difference is. In this translation of a haiku by Kobayashi Issa, "world", "dew", "of", "and" and "yet" are each 15% of the text, which is not a figure you're likely to see in any prose text, or even other forms of poetry, because of the unavoidable frequency of a core set of English prepositions, articles, and other function words:

> *This world of dew*
> *is a world of dew,*
> *and yet, and yet.*

This brings us to another perk of our corpus: these documents are basically the same genre. Sure, one is arguing for something and the other is arguing against something, but it's the same issue at stake. They were both written primarily by lawyers, in the same year, for the same audience. Unlike with literature, where one might debate, for instance, how many nods and winks to adults there may be in a children's book, these documents are inherently functional in nature with a clear and intentional audience of the Copyright Office. Even more conveniently, while they include some text that was not written by lawyers, *both documents* contain non-lawyerly text – and some of it is even from the same people!

Our research questions shape our corpora, and our corpora shape our questions: there are things you want to explore, and the things you can explore with what you're able to get, within the constraints of reality. Our question here could be formulated as something like, "How does the rhetoric of the petition and opposition depict the use of the exemption, its risks, and nature of the proposed expansion?"

This research question offers us a kind of "easy mode" with regard to our corpus, because we're looking at two documents that comprise 100% of what has been submitted to the copyright office about this petition, as of early March 2024. We often don't have the luxury of "everything" when it comes to these kinds of projects, sometimes because we don't even know what "everything" is (e.g. all youth fiction published in the US in 1952), and sometimes because it's not feasible to get everything. In those cases, we have to grapple with the question of *why* we can find / acquire the

things we *are* able to get. Is there anything that makes them meaningfully different from the "known unknowns" or the "unknown unknowns"? How will that impact the claims that we're able to make with our results? Happily, those issues aren't a problem here – but in most projects, we spend time thinking about it. There are lots of different but equally valid ways to make choices about corpus selection, as well as lots of methods for analyzing the resulting materials. While this example uses two different kinds of methods (word searching and topic modeling), digital humanities is a very broad and methodologically diverse field, and not all projects look like this one.

2. Text acquisition and transformation

All the documents are freely available for download as part of the US Copyright Office's Ninth Triennial Section 1201 Proceeding, 2024 Cycle. Each of the opposition filings is its own PDF. But since part of this exercise is to model the real-world conditions we often deal with when doing digital humanities projects, let's imagine that all four opposition filings come in a single PDF, ordered by filing number. This is not unlike trilogies published as a single volume, or anthologies that combine many different short stories on a single theme.

While there are a small number of tools that can work directly with PDFs (as long as there's an embedded, machine-readable text layer, and no digital locks involved), we almost never work directly with PDFs. What's most useful for text analysis is a plain text (.txt) file.

If you don't spend much time with plain text files, they're very much as advertised: a file with only plain text. "Plain" here means it's missing all the things that would make it enjoyable to read. You may not give it a lot of thought, but decent typography and layout contribute a lot to your reading experience. Plain text files have paragraph breaks, but that's about it in terms of affordances for you as a reader. Chapter or section headers are not set off in any way – the text is not bigger or bolder, and there's no application that can recognize them and let you jump between them easily if you're looking at a plain text file. You're also going to lose any meaning carried by the font (which can be important, for instance, to render unique dialogue by aliens or robots) since there's no font information. Also, any italics or bold? Gone. And this can have an impact on your ability to accurately read the text: a character's dismissive "Whatever" takes on a very different intonation when it's "What*ever*." All this information, lost in the conversion to plain text, is *gone* (deliberate emphasis in italics). Even if you take that plain text file and convert it back into an ePub or PDF or any other ebook format, it won't magically reinstate that formatting information. Project Gutenberg has a plain text option, but any human looking to read *Alice's Adventures in Wonderland* with their own eyes will choose any of the numerous alternatives.

I can think of all kinds of interesting research questions I'd love to ask about italics use and other typographic choices in different genres of literature across time, but every research method has its affordances and things it isn't suitable for, and computational text analysis as we usually do it is very poorly suited for those questions, because of how we work with plain text.

3. Documents include multiple things

Quite often, we start a project because we've read some of these documents (with our eyeballs, on paper or as a proper ebook) and something struck us – and we want to follow up on it by exploring it systematically. That's what happened for me here: I skimmed the petition and responses, and particularly in the responses, there were some words used with a frequency that surprised me: AI, Mellon, fair use. Less surprising was "security". But I wondered if they were actually as *frequent* as I was imagining them, and *who* was using them, *where*, and *how*. In any case, these were documents I'd read with my eyeballs, and I had a starting list of words I was interested in.

At this point, I have plain text files that will work with Python code I write to look for words. So… can I do that? There are no technical hurdles to doing so, but I have reason to suspect that the results aren't going to be interesting or useful if I try it on the plain text files that I currently have.

Let's take a closer look at the petition document, using the PDF page numbers:

- Pages 1-2: Item A: Commenter information. This has a description of the two organizations that are filing the petition, along with the names of the lawyers representing them. Page 1 also has some copyright office boilerplate text, including a Privacy Act Advisory Statement, and the address of the copyright office. That isn't text that was provided by the petitioners at all, but it was added after the fact as part of publishing this document.
- Pages 3-4: Table of contents. This text is intended as a finding aid for a human reading a paper or PDF copy with their eyeballs. There are some PDFs that have these sections set up in a digitally-actionable way, where you could click one to jump right to it; this is not such a PDF. All the text here is a *repetition* of header text that appears later on – but not for any semantically significant reason. This means that if any of the words I'm interested in appear in a header, they'll get an extra "hit" at the beginning of the document, just because there's a table of contents.
- Page 5: Item B: Proposed classes addressed. This has a formulaic description of two classes of copyrighted works, and a justification of why they're being addressed together.
- Pages 5-19: Item C: Overview. The overview boils the petition down to three main points, which have their own sub-headings.
- Page 19: Item D: Technological protection measure(s) and method(s) of circulation. A short paragraph.
- Page 19-34: Item E: Asserted adverse effects on noninfringing uses, which has multiple sub-points, some of which themselves have sub-points.
- Page 35: Cover page for documentary evidence
- Page 36: Secondary table of contents for the appendices
- Pages 37-90 are the set of letters that make up the appendices. Each letter is preceded by its own cover page, indicating the appendix letter and the title of the appendix. Some of

the letters are written on university letterhead, which may have additional text that is not related to the letter itself.

If we were looking at a corpus of tens of thousands of legal documents, we might conclude that it's not feasible to split and clean up all those files, or we might try a rough pass on doing this en masse – for instance, by writing code to automatically split the document before phrases like "ITEM B" and "ITEM C", as long as they're not followed by a lot of periods and a numeral, as you find in a table of contents. To check and see if this was successful, we would need to actually look at the automatically split text files with our own eyes, to make sure that they had the right contents – for instance, that Item E in some document doesn't have a reference to Item D in its text that accidentally got picked up and split into a spurious new file.

The proposed language in the AAP opposition to the petition would make this impossible while complying with the requirements of the exemption, since they state that anyone doing this work must view "the contents of the literary works in the corpus solely for the purpose of verification of the statistical research findings and no other type of analysis". Quality control of automatic text splitting is not a statistical research finding, but it is an essential prerequisite to make sure the files we're analyzing contain what we think they contain – which is fundamental to our ability to make any legitimate claims using the results.

One conceivable response may be to modify "statistical research findings" with "computational analysis", since this could accommodate data cleaning processes that involve searching the text for words and splitting the files accordingly, as described above. However, this kind of automated rough processing is not the only way we prepare text corpora. We work with large corpora when we are less interested in the contents of any one specific document than in the sum totality of the texts. In those cases, we can write off a handful – or even more – of specific texts if they prove to be formatted in a way that our scripts aren't able to successfully split, without it impacting the project overall. Other times, though, we *are very interested* in the contents of a small handful of documents, and throwing out one or two for being difficult to parse would make it impossible to proceed with the research project.

Writing a script to automatically split files is a fiddly task that can be time-consuming: you have to write very precise search parameters, typically using a syntax called *regular expressions*. There are lots of possible variations that might work to identify the things you're looking for, but other variations that seem near-identical may fail. What's more, it is much easier to write these search parameters looking at the *actual text you're splitting* rather than the copy optimized for human readers, because it's not always 100% clear how the plain-text conversion process will handle formatting and spacing. For instance, the AAP's opposition involves some strike-through text as one can read in the PDF, and each of the provisions they suggest modifying appears on its own line:

(B) The copy of each literary work is lawfully acquired and owned by the institution, or licensed to the institution without a time limitation on access;

(C) Any ~~The~~ person undertaking ~~the~~ circumvention or research activities views the contents of the literary works in the corpus solely for the purpose of verification of ~~the~~ statistical research findings and no other type of analysis; and

If I were trying to extract this kind of text using a script, one of the things I would be likely to try early on would be looking for single capitalized letters, in parentheses, at the start of the line. But if I look at the actual text file I'm working with, rather than the reader-friendly PDF, I would quickly see that the formatting is very different.

```
465  (B) The copy of each literary work is lawfully acquired and owned by the institution, or
     licensed to the institution without a time limitation on access; (C) Any The person undertaking
     the circumvention or research activities views
466
467  the contents of the literary works in the corpus solely for the purpose of verification of the
     statistical research findings and no other type of analysis; and (D) The institution uses
```

Those newlines before each letter are not there, but one has been inserted randomly in the middle of (C). If it were not permitted to *look at the actual text file*, I could easily have lost half a day to maddening trial and error with regular expressions trying to figure out something that would literally take a second of looking with my eyeballs. Also, depending on how strictly this proposed "no viewing the contents" rule is interpreted, it would preclude writing statements in the code that display portions of the text to check that you're getting what you think you're getting, which is a fundamental tool in our debugging toolkit when trying to find regular expressions that work for finding text.

The effort of writing and debugging and testing and fixing code to split documents computationally is worth it when you're faced with hundreds or thousands of them. If you've only got a single document with a given format, it doesn't make sense to do it that way. What's more, not all documents even follow a reliable and consistent structure. While the original petition split each submitted letter into its own appendix, the AAP letter has a set of "Exhibits", and exhibit 3 groups all the letters together. Any code I wrote to split the letters in the petition would not work for the same purpose in the AAP letter.

4. The power of human eyeballs

Given a corpus the size of these copyright office filings (which is to say, very small), with heterogenous formatting (so it'd be challenging to write code that would reliably work), it is much more realistic to split them up manually: opening the plain text files, and using eyeballs aided by some degree of more nimble text searching, creating files that have meaningful portions of the text. With human eyes and brain involved in the process, the searching doesn't have to be as precise. All you need to do is find *somewhere roughly around* the boundary between things

you want to separate, and then you can use your own human knowledge to select the correct range of text to move to a new file. Keep in mind that these constraints (very small corpus with formatting all over the place) do not apply to every text and data mining DH project, or even most of them! For projects on the scale of what I'm doing here, eyeballs are far and beyond the most feasible way to split up the text files.

So that's what I did here, manually creating files that are labeled with both their origin and contents. But crucially, this requires not only looking at the text files, but also modifying them – creating files with the content separations needed for the analysis I want to do. I'm not reading them any more than I have to; again, plain text is not conducive to a satisfying reading experience. I'm skimming, recognizing the boundaries of segments, splitting the text and moving on. This is, frankly, tedious and unpleasant work. No one enjoys doing this, or even retains much about the content of the texts after having done it. It is no substitute for a reasonable human reading experience; it's just putting yourself in the place of a computational algorithm, since you're far better than the tools at hand.

Human eyeballs are much better than algorithms for a handful of things, and one of those is immediately recognizing problems. In this case, I realized that most of the letters the AAP included as its exhibits were simply screenshots (or perhaps, un-OCR'd scans), rather than computer-readable text, leading to blank sections in the plain text files I'd saved from the filed PDFs. The original petition did this as well with some of its appendix letters, and others exported but with errors (e.g. the letter from Rachael Samburg and Tim Vollmer replaced every instance of a double l with a single l and a space; this is a problem since one of the words I'm searching for is *Mellon*.) I've seen errors like this before in my multiple decades of scanning and OCR-ing books to build research corpora, but it's not so common that I'd ordinarily check for it if I were testing the quality of the text computationally, which overall is quite good. What's more, I Looking at it with my eyeballs, though, I immediately noticed this and realized its impact on my search terms. Crucially, though, I never would have noticed that there was a problem if I hadn't consulted the plain text version of this document, because the human-friendly PDF showed no sign of trouble at all: it was perfectly legible and correctly formatted.

In the end, I ran the affected PDFs through the ABBYY FineReader software to obtain an OCR'd version of those documents, which I used only where there was a blank in my original export, since the OCR introduced a small number of errors.

Manually splitting the files allowed me to curate the corpus a little bit as well. To my surprise, the export from the PDF to plain text included the text in the screenshots from Kinolab, from the DVD-CCA response. As I understood the DVD-CCA response, the point of including these was not fundamentally the text (mostly metadata about films) but rather, the interface itself and its affordances. The fact that *Bicycle Thieves / Ladri di biciclette (1948)* is filed under "Drama" doesn't contribute to the discussion, and contributes additional lexical items to this overall set of documents that aren't actually meaningful to the discussion at hand. As a result, I made the call to exclude it from the final set of split text files.

5. The analysis

I wanted to visualize the number of occurrences of a set of key terms that had jumped out at me as I first read these documents (in PDF form) with my eyeballs:

- AI
- collaborat-
- fair
- Hathi
- mellon
- pira-
- security

I have some code that I often run to find words in documents. I had to rework it a bit to accommodate partial words (since I wanted to capture variants like collaborat[ion/or/es] or pira[te/ted/cy]).

I would not generally claim that the output of that code is a *statistical research finding*, but rather *another kind of analysis*. Which is exactly what the AAP would like to exclude from legitimacy when it comes to the question of the circumstances under which researchers can look at the plain text file. It seems deeply unjust that simply because my research method is straightforward, it should be denied the same verification access as more sophisticated computation – especially since the opponents also make an effort to conflate all machine learning or stochastic methods with not only the vague but threatening specter of "AI", but *generative AI* in particular. This is as reasonable as arguing that basketball should be off limits for vegetarians, because the game involves a ball, and everyone knows that meatballs are made out of dead animals. In fact, consulting the original texts here was important for understanding the usage and context of some of these terms, particularly AI.

7 of the 34 references to AI come from the original petition. Even splitting that document into its constituent pieces paints a clearer picture of that term's use: it comes entirely from the letters in the appendix. Three come from Matthew Sag's letter, and two each from Allison Cooper's and Lauren Tilton/Taylor Arnold's letters. Sag's letter notes that he "testified to before the U.S. Senate Committee on the Judiciary Subcommittee on Intellectual Property about Copyright and AI." and twice references a forthcoming paper entitled "Copyright Safety for Generative AI". Cooper's letter notes that "Kinolab would benefit from an exploration of the ways in which AI might enable further research on film language" and "If the Library of Congress expands the TOM [sic] exemption to allow for broader corpora sharing, Kinolab will pursue partnerships with researchers like Bamman at other institutions of higher learning to explore the development of ethical AI-based tools for searching moving images". Tilton/Arnold say "Designed to support media and AI literacy, the [Distant Viewing] toolkit is being designed to help a broader public understand how computer vision can help them analyze images" and at the end add that "scholars [in the EU] are positioned to innovate in AI and machine learning while scholars in the United States would be barred from this kind of research if this expansion is not granted."

I could have gone looking for the locations of these AI references in the original PDF document (where I may have had to literally skim using my eyes, and approximate percentage within the

document – recall, not all the letters include machine-readable or -searchable text, if I were to try to search within the PDF document), which could have taken a fair bit of time to track down. Instead, I easily modified my code to print out the full sentences where each term occurs. This tiny context window should reasonably be covered under fair use, as it is essential to text and data mining for scholarly research and teaching. But it also involves extracting sentences from the plain text files, rather than the PDF directly, which is what the opponents object to.

The rhetoric around AI in the opposition documents is fundamentally different. There is one example from the STM opposition letter (in the TPM section), 7 from the MPA-RIAA (in the adverse effects section), and 19 in the AAP response (mostly in the adverse effects section, but also under the TPM and overview). Examples include "Even more concerning is the fact that researchers operating under the exemption are seeking to exploit their DRM-free corpora to train or develop generative AI systems." (No researcher said anything remotely to that effect.) and "the letters of support submitted by petitioners suggest that the corpora and/or results of TDM research could also be (and seemingly are being) used for their expressive content, including for the development and training of generative AI systems." (This indicates to me that the opponents did not understand what the letters are proposing, or are choosing to misunderstand in bad faith.) One of my goals for this research project is to visualize the use of my identified set of terms across these documents, but because "AI" in particular is being used so differently – especially with regard to its baseless juxtaposition with "generative" in the responses – I want to make sure that that distinction is noticeable in the final visualization.

6. The visualization

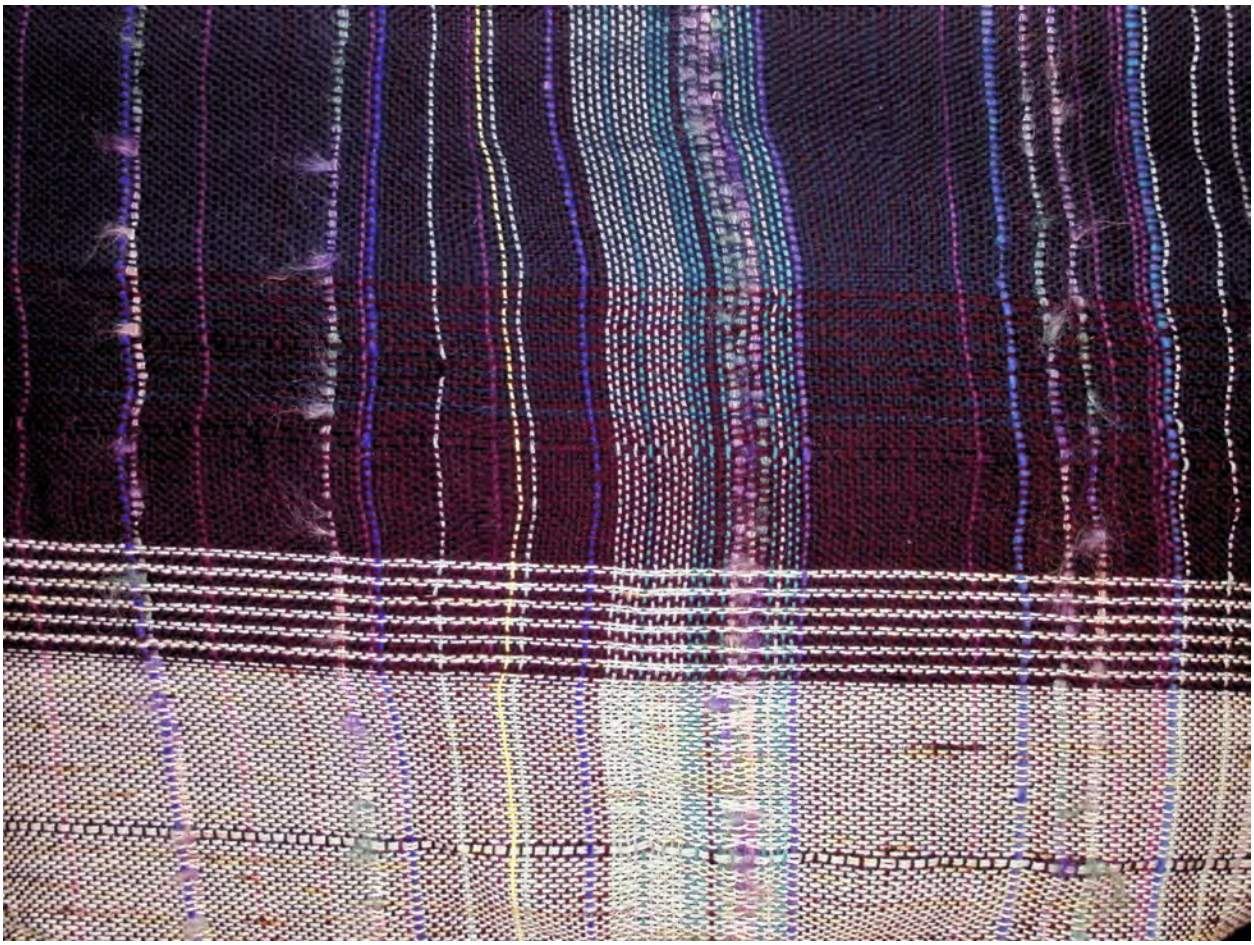
One unusual thing about digital humanities compared to many other disciplines is its embrace of creativity and physical making as a community praxis. Many digital humanities scholars also run makerspaces, print shops, or other spaces that grapple with materiality in a way that may be surprising for scholars who are also enmeshed in the digital. Rather than doing a simple computer visualization of the distribution of my seven key terms, I wove one. The warp (horizontal yarn, in the image) is split between the petition (blue) and opposition (different white yarn for each opposition document), with collaboration (red) in the middle. The weft (vertical) yarns are color-coded as follows, with each yarn representing one occurrence of the term:

- AI: fluffy purple/blue/green (because AI is a "fuzzy" concept)
- Collaborat-: red
- Fair: light blue-green
- Hathi: turquoise
- Mellon: dark purple
- Pira-: gray
- Security: yellow

Additional blue or white yarn is used to indicate the presence of documents with no key terms, or to fill out the overall word count for documents that had a lower percentage of the key words. Letters are marked with a thick blue-purple yarn.

References to AI are woven in two ways: if AI is being used as a generic umbrella term for probabilistic computational methods, it is woven using the same plain weave as the rest of the visualization. When AI is modified (or implicitly modified) by “generative” in the source document, it is woven with a technique called leno, which involves twisting warp threads over each other before weaving, resulting in a lacy texture.

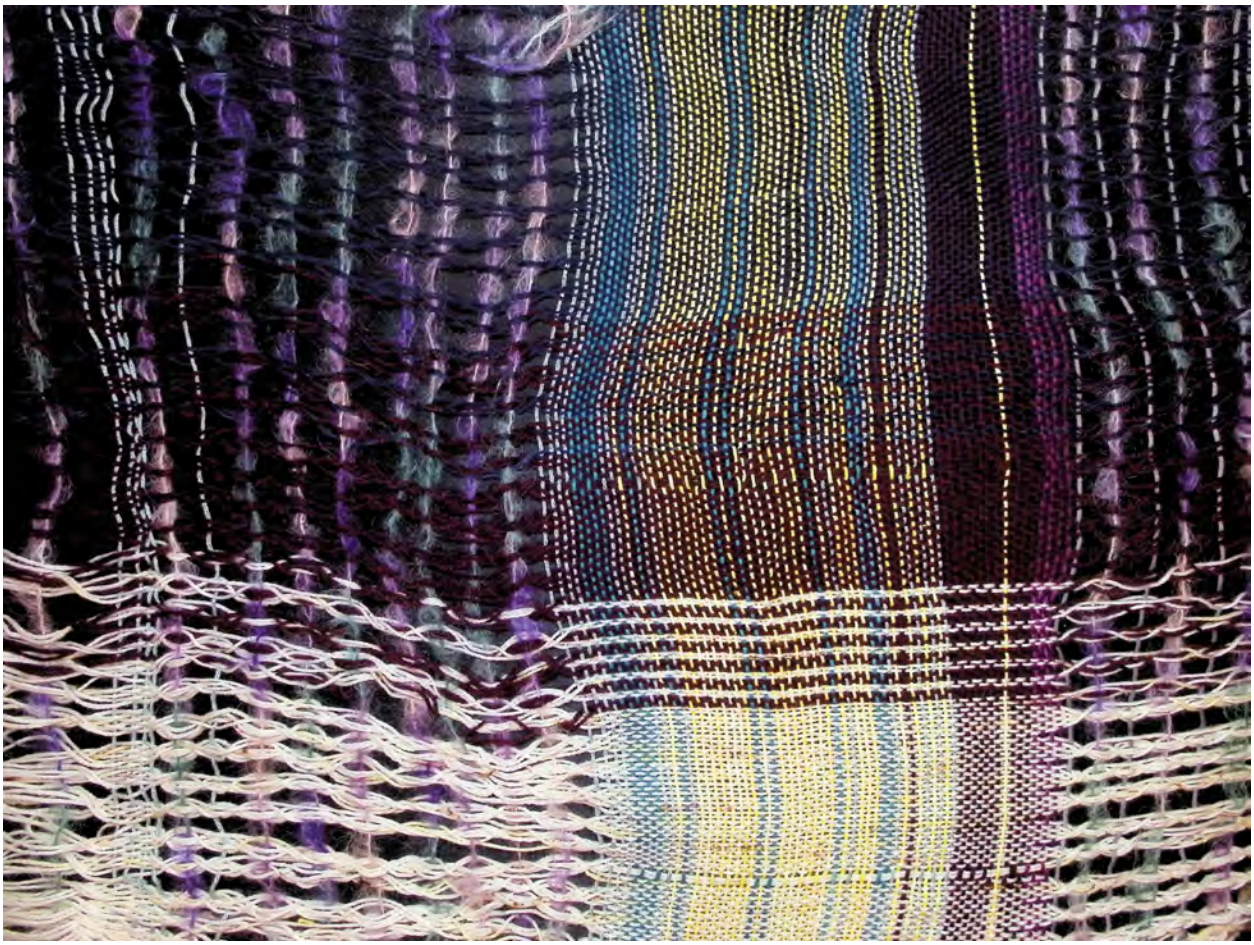
The beginning of the petition; most visible terms are security (yellow), Mellon (purple), Hathi (turquoise) and fair (light blue).



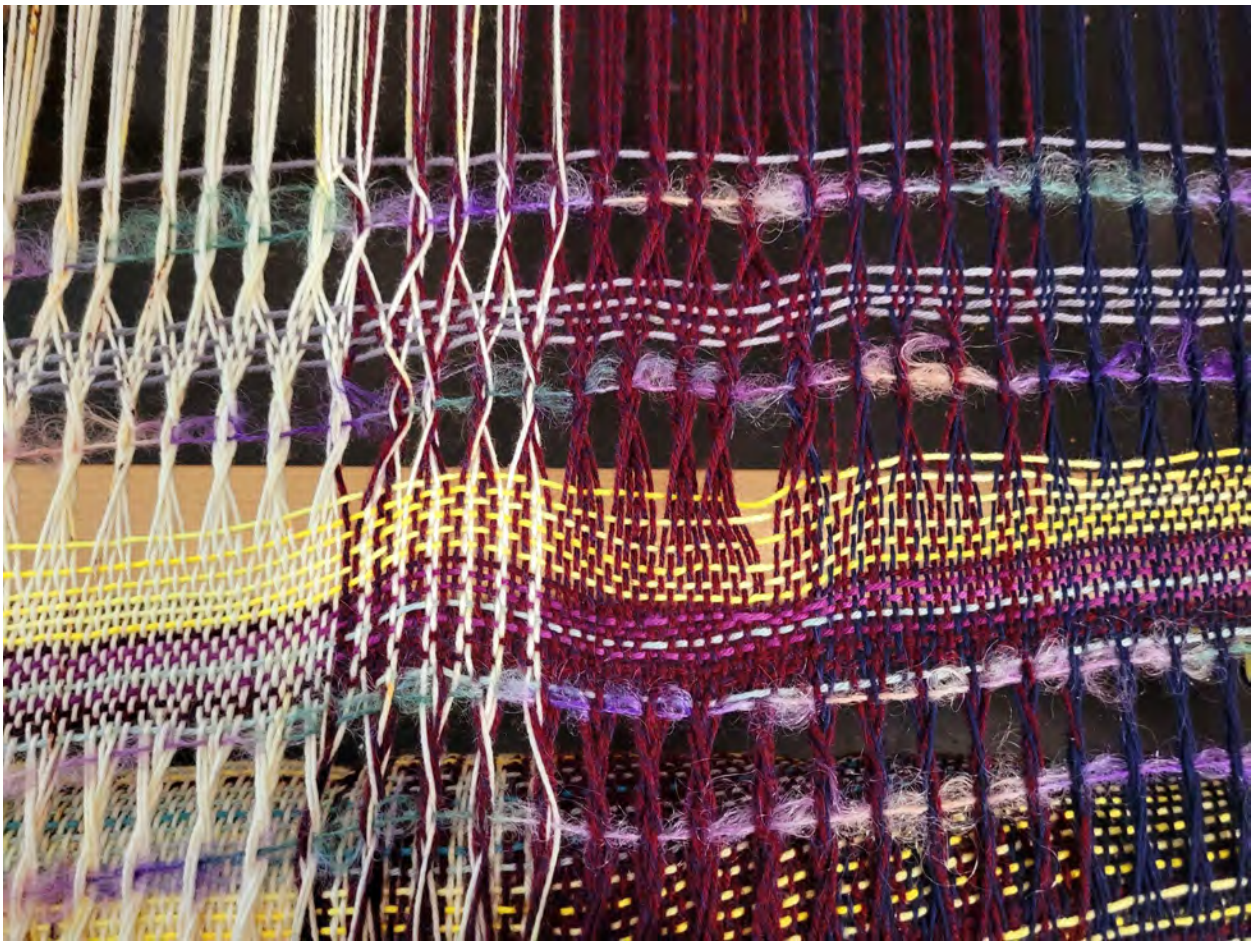
The beginning of the petition; most visible terms are security (yellow), Mellon (purple), Hathi (turquoise) and fair (light blue).

Left to right: Matthew Sag, Rachael Samberg & Timothy Vollmer, and Lauren Tilton & Taylor Arnold's letters of support for the petition. The term "AI" is particularly visible as a fluffy yarn.

Part of the opposition filings. "Security" in yellow is particularly visible, as is "fair" (in light blue) and "AI" in fluffy yarn. The leno twists visible in the white warp section represent references to generative AI.



Part of the AAP's filing about adverse effects. Note the repeated references to generative AI (as visible through the twisted warp yarn), and the section referencing security, HathiTrust, and fair use in the middle.



Close-up of a section from the AAP's opposition filing, in the section on TPM. Terms used are [generative] AI, piracy (gray yarn towards the top), security (yellow), Mellon (purple), fair (light blue), and collaboration (maroon).

7. An "AI" powered analysis

Prior to the release of ChatGPT, I tried hard to never use the term "AI", opting for something more precise instead. Machine learning. Natural-language processing. Large language models. That rhetorical war has been lost, and many kinds of methods and processes are now lumped together under that label, with its power to impress and/or terrify (sometimes both at once). I've seen it defined so broadly as to include anything non-deterministic, such as the MALLET topic modeling tool originally developed in 2002. The way topic modeling works is, basically, you give it a set of texts (which can – and usually should – be smaller than an entire document, think more like paragraphs or sections than novels or legal filings), you tell it how many "topics" (or groupings of words that tend to co-occur) you want it to generate, and it shuffles all the words around into different buckets until it's finished. Unless you very specifically fix some parameters, you'll never get the exact same results twice (again, because it's a machine learning model, non-deterministic, and therefore "AI"), but if it's working well, you should get similar results if you run it over the same corpus multiple times.

I broke up the files I'd split into paragraph chunks, combining paragraphs together if they were under 100 words. Then I ran a 20-topic model. Running it a few times, the clusters of words were fairly consistent. Below are the list of the 50 words most associated with each topic, with the topics numbered 0-19.

0 0.25 works literary circumvention proposed exemptions measures c.f.r section motion technological protection pictures purposes office dissemination language proposal i)(d replication distribution controls rule prevent circumventing adverse copy educational protected tpms place corpus identified downloading librarian relevant unauthorized time includes involved infringing effects nonprofit limit requires statutory asserted adopted employ terms facilitate
1 0.25 proponents fair copyrighted register rulemaking class noninfringing comments recommendation initial amount scope act permit oct proposed sharing argument portion broad constitute evidence cited permitted fed reg prior activity determine quoting rec performances infringement relation e.g argue limitations term place underlying notwithstanding reply wrote alleged raised ground determination order permits notice
2 0.25 berkeley mining authors knowledge states scholarly data united program copyright public alliance digital projects december advance communication millennium california scholar area demonstrating months number networks applications award content dataset thousands archives supporting collaborative million strategy guiding grantmaking demonstration grants global purchased length date location grantee organization social funds alliance's source
3 0.25 research scholars working project time materials process data set in-copyright significant stanford share collaboration small additional preparing shows computational part members expensive breaking library groups understand break cases issues resources drm cfr effort collaborative staff wil larger teams decryption context similar taking complete films dvd concerns computing digitization painful slow
4 0.25 works databases pursue emphasis expanded set questions copies fiction ways standards added single requests appears vast received risk wide novels range potentially question majority distribute open additional international recipients writers optics receive instance published ownership explore arguments space african-american simply built people circulation setting ensure stm's impossible increasing unable downstream
5 0.25 text tdm mining data copyright research law legal support exemptions issues matthew sag case scholars building register digital office institute individual expand vol computational literacies guidance policy national samberg n.d applying lltdm-x cross-border richard challenges face renewal times paper community technology mer jean empirical lltdm filed patterns librarians science navigate
6 0.25 access content library rights make aacs members system libraries discs protect format blu-ray audiovisual e.g include academic journal materials llc property intellectual video ebooks millions create music services streaming and/or companies disc **pirated** online service **piracy** management articles distributed major efforts case enable developed advanced software control limited measure(s desired
7 0.25 university professor information director request counsel letter response office hart general activities virginia associate temple butler sincerely subject chicago department college circumvention february emory lab stanford english january legal behalf dear terrence richmond policies respond school writing copyrights act continue graduate ucb email brandon teach direct write studies corpus law

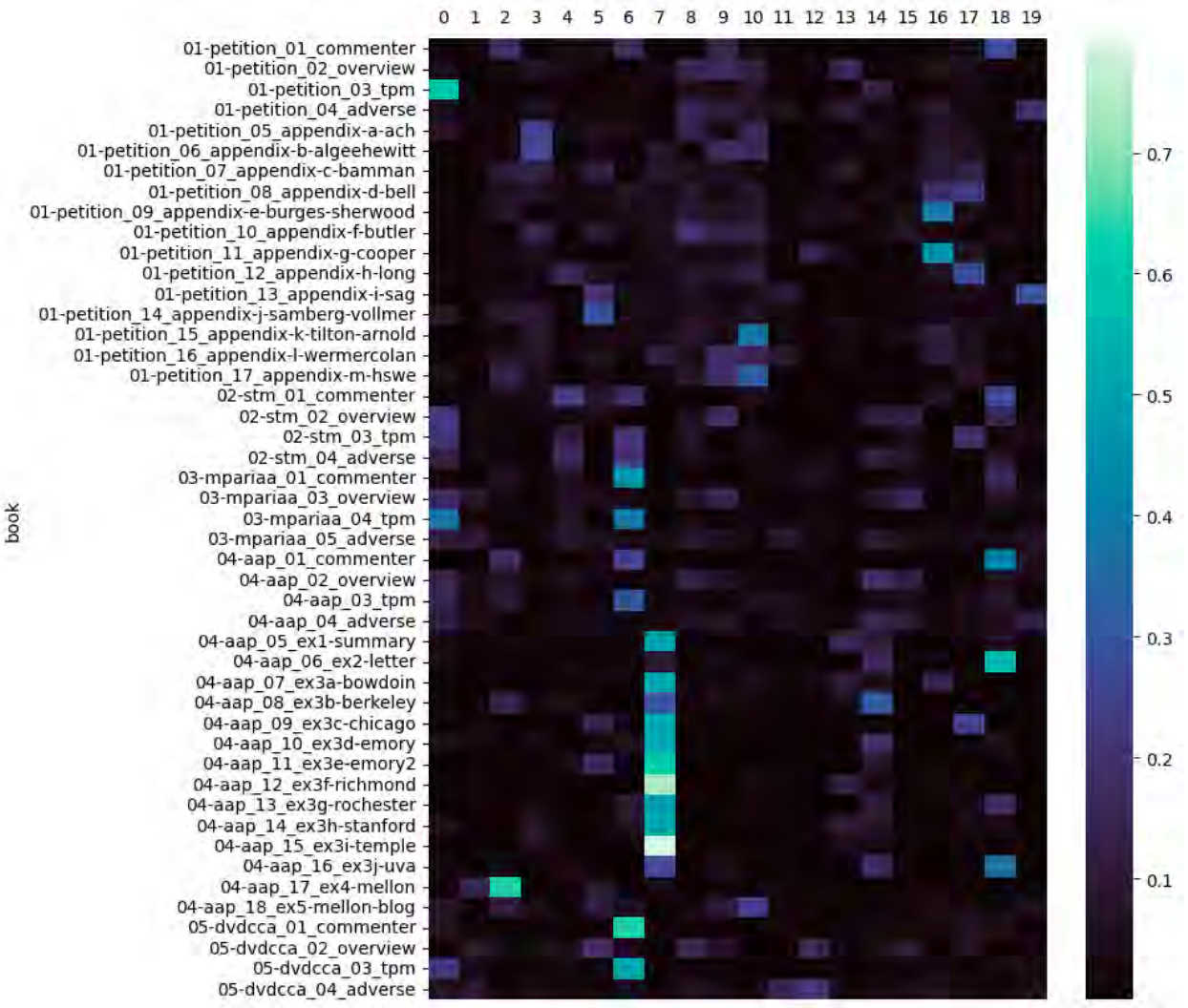
8 0.25 researchers corpora institutions expansion proposed corpus exemption sharing works work institution research costs share ability copies existing researcher limits underlying comply inability lawfully decrypted limitations requirements lack **collaborate** required clear simply labor study create shared practice affect perspectives obtain acquired allowing issue fact compiled reduce exemption's academic ultimately limitation undertake
9 0.25 research exemption current researchers tdm education teaching higher projects valuable access limited students affiliated existing independent academic corpora findings conducting contemporary enable ensure work noted full effectively enabled scholarship conduct result develop support staff purposes material relevant provide compliance order independently ambiguity prevent interest conducted limiting impact solely degree requirement
10 0.25 humanities data research digital methods **mellon** build foundation field questions culture understanding study technical barriers important science quality arts scholars history notes increase projects texts grant specific collaborations society efforts key experience public benefit complex diverse funding literary including resources broader engage granting addition expand number change sources expanding dmca
11 0.25 books google work hathitrust court cir digital book search guild view fair news copies text snippet copying authors found libraries short creation space-shifting precedent public portions significant supp activity considered copy limited made provided snippets copied function google's reveal term engaging register's defendant u.s.c involved feature sufficiently searcher conclusion substitute
12 0.25 kinolab clips motion corpus users researchers entire pictures bowdoin close clip viewing picture feb curators work annotating platform user collection college annotation fact site authorized annotate exhibit download present i.e visited watch original scenes requirement kinolab's film submitted readily movie hours making pose sharing identify run minutes schema offers examples
13 0.25 letter app algee-hewitt tilton bamman sherwood joel david arnold lauren bell humanities emily mark taylor allison john foundation burges computers mellon appendix professor association long tdm wermer-colan henry alexander timothy brandon butler hoyt vollmer samberg explains tpm fields well-funded rachael cooper redundancy engage dollars creates similarly analyzing screen kind tpms
14 0.25 **security** exemption copyright measures institution information owners activities circumvented institutions protect researcher including provide required highly provided owner circumventing specific reasonable confidential regulation contained created based requirements records books respect ii)(b apply physical identify persons members letter request title publisher risks requested matter made identification responses safeguards secure relied failure
15 0.25 petitioners current viewing proponents collaboration comment support prohibition access proceeding record results allowed distant longer letters third-party due verification petrs supporting generally long rely effective impact answers give i)(a techniques petition led close concerns unprotected petitioner ambiguous basis seek form legitimate seeking subject finally core times prevents discussed demonstrate failed
16 0.25 film project media research television digital studies analysis cooper methods language mediate narrative team building data professor student close-up dvds burges gender scale representation metadata shared collaboration students grant moving rochester bias films image faculty race years sharing review annotated quantitative funded cultural questions identity diversity dmca scholarship visual made

17 0.25 corpus machine work learning computational large literary models collections long literature techniques films cultural analysis make lab text artificial train hoyt textual intelligence human languages datasets deep directly study time japanese related technology serve critical examine corpora inquiry potential model collection trained made capacity begin extracting creators translation algorithmic investigations
18 0.25 american association university publishers press tdm hci aap publishing library information alliance provide c.f.r systems seeking persons activities professors applicable procedures written generative society u.s comment copyright requesting group relevant including aap's explanation comments engage book submitted foia freedom address engaged house policies authors data state technologies stm company uva
19 0.25 works fair copyrighted copyright purpose transformative proceeding triennial factor market work original expressive office analysis u.s highly expression harm copying purposes content warhol information database factors scholarly availability decision sag matthew iparadigms nature creative character non-expressive courts verify solely current plaintiffs substitution statistics interest favor concluded limitation explained applies makes

It's a very different view on these documents than my list of seven words that jumped out at me using my own eyeballs and intuition.

- *Piracy* and *pirated* both appear in topic 6.
- *Mellon* appears in topic 10, as does *collaborations*
- *AI* is not in the 50 terms most associated with any of these topics.
- *Collaborative* clusters with words like "projects", "grantmaking", and "demonstration" in topic 2
- *Collaborative* is also one of the top words in topic 3 along with "research", "scholars", and "library", as well as *collaboration*
- *Collaborate* appears with "researchers", "requirements", "limitations", and "costs" in topic 8
- *Collaboration* also appears in the top words for topics 15 and 16

The fact that different forms of *collaborat-* appear in different topics highlights something noteworthy about topic modeling: it's literally clustering word forms, not concepts. From the perspective of the algorithm, "collaborative" and "collaboration" are completely distinct words – no less distinct than "text" and "statutory". The fact that topic 3 has multiple forms of *collaborat-* is a good sign (having multiple forms of the same word appear together suggests the topic model is picking up on actual coherent trends in the text, since you would expect different forms of the same word to be used in similar contexts). Having forms of *collaborat-* in so many topics suggests that it's a theme that permeates these documents, but this also makes it difficult to use this topic model to look at collaboration in the documents, because you can't clearly point to one or two "collaboration" topics to track. One might be tempted to point to topic 3 (which has multiple forms of *collaborat-*, but if we create a heat map showing how much of a document is made up of a particular topic (where brighter colors indicate a higher percentage of the document's words are associated with that topic), we can see that topic 3 is a poor proxy for the prevalence of collaboration in these documents, as illustrated by our simple word search analysis earlier:



This visualization does highlight some clear associations between certain topics and groups of documents. Topic 7 (top words including: professor, information, director, request, counsel, letter, response, office, sincerely, circumvention) is very strongly associated with responses to the AAP letter. These responses are also, on the whole, quite short, which is why they appear so bright on the visualization: it's easy for a very high percentage of the words to be from a topic when we're only looking at a few sentences. Because of the length difference compared to some of the longer documents which can run many pages, I'd remove these short responses if I were digging into this in more depth, which would better show the amount of variation in topic distribution across the longer documents. Nonetheless, there are other trends we can see here: topic 6 (top words including: rights, property, millions, piracy) appears primarily in the opposition documents. Topic 14 (security, exemption, measures, circumvented, regulation) appears visibly in the TPM section of the petition, and in the opposition.

Many of these topics could be a jumping-off point for further analysis of the discourse in these documents: what are the words that appear together? How are they being used? For instance, you could use NLP tools like spaCy (which use probabilistic language models for identifying things like part of speech, named entities, etc. – meaning it's another thing under the broad umbrella of

"AI") to look at subjects and objects in these text: *who* is asserted to be doing *what*? You could also feed some of the interesting words surfaced through the topic modeling back into the earlier simple word-search algorithm. And there are numerous other computational methods that we commonly use to try to get a handle on how language is being used in texts, some probabilistic, others deterministic.

Crucially, though, doing this kind of work well is slow, painstaking, and requires care and attention to detail. You have to prepare the texts carefully. You have to factor in things like different document length with many kinds of analysis. You have to be able to look at the actual text files you're analyzing – not just the more pleasant human-readable variants – to make sure you're not getting tripped up by conversion errors, bad OCR, or other technical details that can derail your whole analysis. This type of scholarship, done responsibly, is literally the opposite of the ChatGPT-like "AI" that is the implicit boogeyman in multiple opposition documents. We're using tools and our own intuition to closely and carefully analyze documents, not producing text with plausible vibes but a fundamental disconnect from reality.

ITEM D. TECHNOLOGICAL PROTECTION MEASURE(S) AND METHOD(S) OF CIRCUMVENTION

See the original petition for expanding the exemption for text and data mining for a description of the TDM situation.

ITEM E. ASSERTED ADVERSE EFFECTS ON NONINFRINGING USES

See above, as well as the original petition for expanding the exemption for text and data mining.