



SCHOOL OF INFORMATION
102 SOUTH HALL #4600
BERKELEY, CALIFORNIA 94720-4600

November 18, 2020

I am writing this letter in support of the Authors Alliance’s petition to the Copyright Office for an exemption to §1201. I am an assistant professor in the School of Information at UC Berkeley (with an affiliated appointment in the Department of Electrical Engineering and Computer Sciences), a senior fellow at the Berkeley Institute of Data Science, and faculty member of the Berkeley Artificial Intelligence Research Lab (BAIR). My research is centered on the areas of natural language processing and cultural analytics, where I focus on two complementary goals: improving the state of the art for computational methods for literary and cultural objects¹ and applying NLP and machine learning to empirical questions in the humanities and social sciences.² My work predominantly explores the affordances of empirical methods for the study of literature and culture, and has been recognized by the National Endowment for the Humanities, the National Science Foundation, and an NSF CAREER award. I offer these views in my individual capacity as a researcher working in text data mining and cultural analytics, and not on behalf of any organization.

At the core of all work in text data mining is access to data; the ability to access data shapes the research questions we are able to ask, the methods that we select to answer them, and the ways in which our findings are disseminated to the broader public. For a long time, work in cultural analytics was focused on texts in the public domain, such as those accessible through open resources like Project Gutenberg; public domain texts provide a proving ground for analytical methods in text data mining and facilitate the important scientific goal of reproducibility: by providing a stable source of data that *everyone* can access, it enables researchers to verify claims made by others, thereby strengthening trust in the scientific process and encouraging innovation. Part of my research group’s work over the past few years has focused on improving the state of the art in NLP for literary texts; in order to create a benchmark that others can use to evaluate their own systems, we purposely selected 100 public domain works from Project Gutenberg (Bamman et al. (2019), “An Annotated Dataset of Literary Entities,” <https://github.com/dbamman/litbank>).

At the same time, however, public domain resources are necessarily limited. At the time of writing, the public domain in the United States largely spans materials created before 1925. While this body of material includes important works in the 19th-century literary canon (such as Jane Austen’s *Pride and Prejudice* and Mark Twain’s *Tom Sawyer*), it still represents works nearly a century removed from the present day, limiting its ability to answer research questions that are relevant to a contemporary audience. These include not only 21st-century questions on the influence of the internet and social media on literary forms and reading behavior—but even much older questions including the rise of the Harlem Renaissance in the 1920s and 30s. Indeed, public domain works published on Project Gutenberg systematically overrepresent white, male

¹See, for example: David Bamman, Olivia Lewke and Anya Mansoor (2020), “An Annotated Dataset of Coreference in English Literature,” LREC 2020; Matthew Sims, Jong Ho Park and David Bamman (2019), “Literary Event Detection,” ACL 2019; David Bamman, Sejal Popat and Sheng Shen (2019), “An Annotated Dataset of Literary Entities,” NAACL 2019; and Lara McConnaughey, Jennifer Dai and David Bamman (2017), “The Labeled Segmentation of Printed Books,” EMNLP 2017.

²See: Matthew Sims and David Bamman (2020), “Measuring Information Propagation in Literary Social Networks,” EMNLP 2020; and Ted Underwood, David Bamman, and Sabrina Lee (2018), “The Transformation of Gender in English-Language Fiction,” *Cultural Analytics*.

authors, and so the research questions it is able to answer again privilege that social group over others.

There are two primary ways that researchers carry out work on in-copyright texts. The first is through the use of large-scale digital libraries like the HathiTrust, which enable non-consumptive research access to a vast collection of in-copyright materials (17.4 million total works at the time of writing). The HathiTrust Digital Library is a trailblazer in facilitating transformative research by hundreds of researchers by providing access to in-copyright materials, enabling researchers to answer questions at the scale of thousands of texts that simply could not be answered otherwise, but it is not a solution for all research questions. In order to carry out research in a secure environment, all computing is carried out on servers at the HathiTrust through the use of a secure “data capsule” which allows researchers to computationally process texts without being able to directly transfer any material outside of the secure environment. This limits computational processing to the capacity of the HathiTrust’s resources, which is occasionally outside the demands of contemporary state-of-the-art models in NLP—which, for example, may require the use of graphics processing units (GPUs) common in NLP research labs, but not in large-scale conventional compute clusters.

While this mismatch between computing demands and available resources can of course be alleviated as GPUs make their way into compute clusters, one issue that also arises in the use of digital collections compiled by a third party is the presence of gaps in the collection needed to answer a specific research question. The materials in the HathiTrust originate in university libraries, and so are necessarily biased toward academic monographs and away from, for example, mass-market romances and science fiction more commonly found in city public libraries. This gap is a common impetus for the second way that researchers carry out work on in-copyright texts: by digitizing a collection themselves. In my group’s own work on creating annotated resources to improve the state of the art for NLP, we did just that: we bought 500 in-copyright books, scanned them, and carried out OCR on those page scans to recognize the text printed on them. OCR is an errorful process; on a sample of books we scanned, we measured the word error rate to be 0.305% (i.e., roughly one incorrectly recognized word every page); and this process of scanning each one of 500 books is also very labor intensive, consuming the better part of four months. A much faster and more accurate way that we could have selected would have been to buy digital versions of those texts as eBooks; but our concern over violating §1201 dissuaded us from that route, committing our efforts to the slower, more error-prone process and consuming research time that could have been more productively applied elsewhere.

While text certainly has the longest history as the subject of research in data mining and cultural analytics, the rise of computer vision and video processing techniques have also enabled film to arise as a meaningful object of computational inquiry. However, while the existence of public-domain datasets of texts (such as Project Gutenberg) and in-copyright secure environments (like the HathiTrust data capsule) allow researchers to explore text data mining methods without risk of implicating §1201, no such pre-existing resource exists for movies or television. Researchers need to create such datasets themselves.

In early 2018, I decided to create such a dataset in order to explore several questions around film: can we measure directorial *style* in movies? What is it that allows us to immediately recognize that a movie is directed by Wes Anderson and another by David Lynch? While computational methods have shed light on the field of authorship attribution—predicting the author of text either in order to deanonymize them (such as the authorship of the *Federalist Papers*) or to simply characterize what makes them distinct—no such work exists for using computational methods to characterize the visual properties of a movie that uniquely make it recognizable as the style of a particular director. We might hypothesize that “style” in this case can be decomposed into a number of aspects that *could* be measured—including pacing variables such as average shot length, proportions of shot types (close-up vs. long shot), and the color palette used over the course of

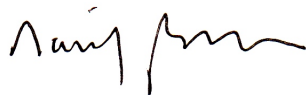
a movie. I decided to begin assembling a dataset to examine these questions, ideally creating a dataset of ca. 10,000 movies; a dataset this large would not only be sufficient to answer questions of directorial style, but would revolutionize the computational study of film by putting it on the same scale as work in text analysis.

The barrier, however, was in creating such a dataset. One fast and accurate way for a researcher to create such a dataset would be to buy DVDs and use software to rip the film from that medium. Given my own low tolerance for §1201 risk, however, I decided to digitize movies in a way that accords with fair use: buying DVDs of the movies, playing them on a computer, and using screen-capture tools to record the screen while the movie is playing in real time (note the time required to digitize a movie using this method is exactly the original runtime of the movie itself). While this digitization method allows for movie data to be used in computational analysis, it is an imperfect process that necessarily loses important information—the subsequent data is of lower resolution than the original medium, and important structural markers like chapter boundaries are lost. However, after digitizing roughly 200 movies in this way, it became clear that this was an infeasible path forward. If a human operator were present for the duration of the screen capture for each movie with an average runtime of 2 hours (and worked 8 hours a day, 5 days a week, 50 weeks per year) it would take 10 years to complete the act of digitization alone.

Two years later, I have still not taken up this original line of research, which I expect will be transformative once it is able to be carried out. If an exemption to §1201 were granted, I would certainly pick up this line of research and begin examining other ways in which movies can function as an object of study in their own right—questions worth examining include historical trends in film over the past century (has the depiction of violence within movies become more or less predominant?), imitation and influence (can we trace which directors have had the greatest impact on the visual style of directors after them?) and reception (which specific aspects of film do audiences, critics, and the box office most respond to?).

All of these questions have been examined within text data mining given the existence of large digitized text collections, but so far remain outside the limits of our knowledge for film.

Sincerely,

A handwritten signature in black ink, appearing to read "David Bamman". The signature is fluid and cursive, with a prominent initial "D" and a long, sweeping tail.

David Bamman
Assistant Professor
School of Information
University of California, Berkeley