Learning from Copyright Rejects:
Topic Modeling and Applied Copyright Law

Jamaica Jones, PhD Student
University of Pittsburgh, School of Computing and Information

Introduction

In July of 2021, Southern Illinois University law professor Zvi S. Rosen presented a paper at the annual workshop of the International Society for the History and Theory of Intellectual Property. In it, he reported the results of his research into the history of copyright registration decisions made by the various entities tasked with such responsibilities throughout U.S. history.[1] His was a painstaking manual effort, individually knitting together data in a variety of heterogeneous legacy formats to identify broad themes spanning the history of copyright registration in the U.S.

The study described in this paper was inspired by a curiosity about what additional context could be afforded surrounding copyright registration were a specific computational method known as topic modeling applied to related data sets. The study recounted here was performed over a collection of rejection letters written by the U.S. Copyright Office Review Board to communicate the reasons why particular works were deemed unsuitable for copyright protection. The Review Board is the entity within the U.S. Copyright Office tasked with making final determinations about the registrability of works.[2] While formal registration is not required to ensure protection, doing so is necessary to litigate instances of suspected copyright infringement. As a result, an average of about 500,000 works are formally registered with the

[1] Zvi S. Rosen, "Examining Copyright" (2021 Workshop of the International Society for History and Theory of Intellectual Property, Bournemouth, 2021), 1–56, https://www.ishtip2021.org/links/link/examining-copyright.

Office every year.[3] Should a submitted work be rejected, its authors (or their attorneys) can appeal such an outcome twice. The second appeal is considered by the Review Board; the outcome is the communicated via letter, itemizing the specific reasons why a work may or may not be deemed registrable. In 2017, an online database of these letters was launched, making available to the public all Review Board decision letters dating from April 2016 to the present.[4] It is these letters that comprise the corpus analyzed in this study.

To provide for a robust understanding of the study's methodology, this paper will begin with an introduction to the use of topic modeling as a research tool. Following this, the study's methods will be detailed and its findings categorically presented and then discussed, largely in light of Rosen's guiding scholarship.

Topic Modeling and Latent Dirichlet Allocation

Like intellectual property, topic modeling is a term that is used to reference a concept with multiple applications. Topic modeling has been defined by David M. Blei as a suite of "statistical models that analyze the words of… original texts to discover the themes that run through them, how those themes are connected to each other, and how they change over time".[5] This study leveraged a probabilistic generative model known as Latent Dirichlet Allocation ("LDA") through the combined use of two software programs – one that performed the statistical

[3] U.S. Copyright Office, "Overview of the Copyright Office," Copyright.gov, accessed January 5, 2022, https://www.copyright.gov/about/.

[4] George Thuronyi, "Copyright Office Launches Online Database of Review Board Decisions | Copyright: Creativity at Work," webpage, June 2, 2017, //blogs.loc.gov/copyright/2017/06/copyright-office-launches-online-database-of-review-board-decisions/; U.S. Copyright Office, "Review Board Opinions," Copyright.gov, accessed January 5, 2022, https://www.copyright.gov/rulings-filings/review-board/index.html?loclr=blogcop.

[5] David M Blei, "Probabilistic Topic Models," *Communications of the ACM* 55, no. 4 (April 2012): 77–78, https://doi.org/10.1145/2133806.2133826.

analysis underlying LDA (a program known as MALLET) and one that provided a graphical user interface to simplify the setting of parameters and preprocessing preferences. Throughout this paper, topic modeling and LDA will be referenced interchangeably. As explained by Benjamin Schmidt, Latent Dirichlet Allocation was initially developed to aid in information retrieval, allowing researchers working with large bodies of unstructured text to "'read' the topic headings first…, only search[ing] out articles in the topics that interest" them.[6]

Scholars unfamiliar with computational methods might think of traditional texts as being structured, given that they have recognizable titles, sections and syntaxes. Reconceptualizing these texts as the unstructured documents that they are in the world of computational methods is helpful in understanding the ways in which topic modeling works. Taking this paper, for instance, one can imagine that each of its component terms were broken from their semantic structures, loosening them from their sentences, paragraphs and punctuation. Each unique term would be conceived of as a "token". Each token would then be tallied such that the document that represents this paper would assume the form of a matrix comprised of unique terms and the number of times each one appears. In the world of LDA, the resulting set of terms and their sum totals is what would be known as a "document". Adding all of the documents together creates what is referred to as a "corpus".[7]

Latent Dirichlet Analysis is known as a "generative probabilistic model," meaning that the algorithm underlying it "contains a model for generating documents by randomly picking

[6] Benjamin Schmidt, "Words Alone: Dismantling Topic Models," *Humanities Journal of Digital Humanities* 2, no. 1 (2012), http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/.

[7] Matthew Burton, "The Joy of Topic Modeling," May 21, 2013, http://mcburton.net/blog/joy-of-tm/.

words from a set of topics based on probability parameters."[8] The topics output through LDA processing can therefore be best understood as a list of words from which each document in the corpus has a certain probability of being generated. Isoaho, Gritsenko and Mäkelä explain generative probabilistic models in the following passage:

> First, topics are modeled as bags of words, with a variable number of each individual word inside each bag…. The documents in a collection are also modeled as such bags, holding all of the words in the document, without regard to the order in which they appear. The… algorithm then tries to re-create these document bags of words through the following process: First, from the set of topics covering the whole document collection, select some number of topics to which a particular document pertains, and the proportion in the document….Then, re-create the document word bag by sampling words at random from each topic bag in the thematic proportions selected previously.[9]

What results is a number of topics, each of which is distributed in some allocation – however small – over each document in the corpus. The topics are therefore "generative" in the sense that there exists a certain probability that terms – or tokens – drawn from a topic's "bag of words" can be used to generate the documents analyzed. An attentive read of the above quote will reveal an important point about topic modeling: while topics are models of relationships between tokens, they are agnostic of the "meaning" of those tokens. Tokens can therefore be – and quite often are – words but they need not be words of a single or specific language. They can be numbers as well.

How topics might best be conceived is the subject of much discussion across related literature, with many suggestions having been offered in hopes that a proper heuristic might help researchers avoid the pitfalls that can arise from thinking too literally about topics being

[8] Karoliina Isoaho, Daria Gritsenko, and Eetu Mäkelä, "Topic Modeling and Text Analysis for Qualitative Policy Research," *Policy Studies Journal* 49, no. 1 (2021): 303, https://doi.org/10.1111/psj.12343.

[9] Isoaho, Gritsenko, and Mäkelä, 303.

indicative of "what" a corpus or its documents are "about." In a 2012 blog post, Ted Underwood addresses the conceptualizing of topics and in his discussion considers them as "discourses," "registers" and "sociolects."[10] He also speaks of them as functioning within specific contexts, which is the term same suggested for consideration by Colin Allen and Jamie Murdock in their forthcoming chapter on the use of LDA in the history and philosophy of science. In defense of this approach, they observe:

> The contexts of writing may include topics in the ordinary sense…. As well as the situation of the writer in the historical moment…. Different writing contexts entail different audiences: letters to friends and family vs. business associates…. Each of these contexts changes the likelihood of the author selecting certain words even when the topic of discussion (in the ordinary sense) is nominally the same.[11]

Throughout this paper topics will be conceived of as contexts, as this understanding provides for a particularly capacious conception of the circumstances under which the outcomes described might arise. These are administrative and even bureaucratic circumstances, in which laws, rules and procedures set forth in a variety of texts are used by a unique cohort of professionals to determine the applicability of those regulations to the products of human creativity, ingenuity and enterprise. These are circumstances in which the regulations have gone unmet and the causal errors explicitly detailed, providing an excellent window into the contours of those regulations.

[10] Theodore Underwood, "What Kinds of 'Topics' Does Topic Modeling Actually Produce?," *The Stone and the Shell* (blog), April 1, 2012, https://tedunderwood.com/2012/04/01/what-kinds-of-topics-does-topic-modeling-actually-produce/.

[11] Colin Allen and Jaimie Murdock, "LDA Topic Modeling: Contexts for the History & Philosophy of Science," Preprint, May 29, 2020, 7, http://philsci-archive.pitt.edu/17261/.

Methods

As mentioned, the U.S. Copyright Office established an online database of the Review Board's appeal decision letters in 2017, making available all such letters dated as early as 2016. All 189 rejection letters available at the beginning of November 2021 were downloaded from this database in their original pdf form. They were then converted into plain text (.txt) format using a Python script and truncated such that boiler plate and administrative sections were removed.[12] Because LDA is a probabilistic model, it is important to preprocess data in such a way that removes repetitive text irrelevant to the content being analyzed.[13] In the case of this study, headers, salutations, introductory and concluding sections were trimmed away. So too were the "Administrative Record" sections of each letter, as these simply detail the transactions and communications that take place prior to final review. What remained were, per the interest of this study, the substantive hearts of each letter – the "Discussion" sections, in which relevant legal frameworks were assessed in light of detailed analyses of each work, and all appendices, which often provided significant descriptive detail about submitted works.

In addition to trimming away sections of text that might skew results, common words also need to be removed prior to running a topic modeling session. Those that are common to all texts are removed in the course of MALLET's processing. Others that are unique to a particular corpus must be more manually removed through the use of a "stopwords" list, or a .csv file containing all of the words and other tokens to be removed. In this project, the commonly-used words "work", "works" and "copyright" were removed, as were all digits from 0 to a randomly-

[12] Tyrica Terry Kapral, the Humanities Data Librarian at the University of Pittsburgh's Hillman Library provided invaluable help throughout this stage of the study.

[13] Karoliina Isoaho, Daria Gritsenko, and Eetu Mäkelä, "Topic Modeling and Text Analysis for Qualitative Policy Research," *Policy Studies Journal* 49, no. 1 (2021): 304, https://doi.org/10.1111/psj.12343.

selected 10,767. Numbers were included for removal because those that were present in the sections of text retained for analysis tended to refer to statutes or sections within the *Compendium of U.S. Copyright Practices*, the document that lays out the administrative and procedural rules governing registration submissions and their review.[14] Such numerically-referenced texts were also referenced in the corpus by name, so removing their numerical references succeeded in "cleaning up" resulting lists of topic words without the loss of any text-based meaning.

It was in this preprocessing stage that words in the documents were also stemmed and lemmatized into their roots – a process by which the word "records" becomes "record" and "swam" becomes "swim". What remained was a single .txt file per original letter, comprised of trimmed tokens and with all format encoding removed (see Figure 1). Throughout the preprocessing stage, 18 of the original files suffered processing errors and were removed from the corpus. This resulted in a corpus that was comprised of 171 .txt files, ready to be modeled using MALLET.

Figure 1.

[14] U.S. Copyright Office, "Compendium of U.S. Copyright Office Practices § 101," 3rd Edition (U.S. Copyright Office, 2021), https://www.copyright.gov/comp3/.

As written, MALLET is run via the command line, requiring a certain fluency not uniformly shared by all who might wish to use it. To bridge this gap, a team led by Scott Enderle created the Topic Modeling Tool ("TMT"), a graphical user interface that enables users to leverage MALLET with more familiar tools. Through the TMT, users can identify input and output folders, upload additional stopword lists, dictate the number of topics to be generated, define the number of words output per topic, and modify the "hyperparameters" that determine the balance of statistical weight across tokens, documents and topics.[15] Optimizing parameters is generally done in those contexts where adding more "weight" to a particular topic is desired.[16] Due to the introductory and exploratory nature of this study, all hyperparameters were left to the TMT default values. Three topics were requested to be output, detailed by the 35 words most strongly associated with each.

This request for three topics was run a total of three times, resulting in three sets of output that were then compared against each other to gauge the stability of results. MALLET output is returned in the form of two series of files, one generated as .html and the other as .csv. All validations were performed in Excel using the .csv files. These files included the following:

- docs-in-topics.csv: Ranks in descending order the documents most strongly associated with each topic.
- topics-metadata.csv: Presents the data in topics-in-docs.csv in table form. As Scott Enderle explains, it is generated with the expectation that it might be paired with supplemental metadata to "build a pivot table that groups documents by metadata

[15] Colin Allen and Jaimie Murdock, "LDA Topic Modeling: Contexts for the History & Philosophy of Science," Preprint, May 29, 2020, 4, http://philsci-archive.pitt.edu/17261/.

[16] David Mimno et al., "Topic Modeling," MALLET: Machine Learning for Language Toolkit, accessed January 5, 2022, https://mimno.github.io/Mallet/topics.html.

categories and calculates topic proportions over those… groups."[17] As detailed below, it can also be a helpful resource without leveraging external metadata.

- topics-in-docs.csv – Assigns one document per row, providing in decimal format the proportion of each document that is associated with each topic.
- topic-words.csv – Strings of the top N-number words (or tokens) most strongly associated with each topic. The Topic Modeling Tool's default N value is 20, but this study utilized returns of the top 35 words per topic. The order in which the topics are listed is arbitrary, but the order of words is not: they are listed in descending order from the one most strongly associated with each topic.

Reviewing the topics output across the three sets of returns was the first step in testing the stability of results. To do this, it first had to be determined that the words associated with each topic remained similar across each set of output. Figure 2 demonstrates that, at least per visual review, they did.

A second test allowed for a spot-check of the stability of topic distribution over documents. The topics-metadata file was used towards this end. Each row of this file represents a document while the columns reference the topics and the decimals refer to the proportion of words in each document that are tagged as being in each topic. Across each row they add up to 1.0000, for a complete distribution of each document across topics (see Figure 3).

[17] Scott Enderle, "Quickstart Guide," Topic Modeling Tool Blog, January 6, 2017, https://senderle.github.io/topic-modeling-tool/documentation/2017/01/06/quickstart.html.

Run 1		
Topic 0	Topic 1	Topic 2
article	white	elements
feature	diamond	design
utilitarian	gold	office
sculptural	derivative	register
pictorial	recycle	id
shape	visa	shape
separable	idea	protection
graphic	material	creativity
design	agents	combination
artistic	tcw	authorship
protection	system	creative
office	franchisee	consist
aspects	book	board
separability	18k	arrangement
separate	program	compendium
function	principle	original
athletica	collection	sufficient
test	computer	color
star	earrings	copyrightable
incorporate	tc	find
ct	yellow	feist
separately	necklace	state
usc	office	letter
imagine	fossilize	court
compendium	grey	registration
request	rise	unprotectable
conceptually	slice	qualify
board	brand	standard
exist	section	form
id	bk	cir
conceptual	ivory	common
cir	chain	selection
find	f3d	author
threedimensional	walrus	claim
independently	flag	expression

Run 2		
Topic 0	Topic 1	Topic 2
elements	article	idea
design	feature	visa
office	utilitarian	expression
register	sculptural	agents
id	white	system
shape	graphic	franchisee
protection	separable	book
creativity	pictorial	principle
combination	diamond	program
authorship	artistic	computer
board	separability	describe
creative	shape	section
consist	separate	show
compendium	function	flag
arrangement	gold	ideas
original	office	methods
sufficient	aspects	text
color	athletica	list
copyrightable	test	deposit
find	recycle	material
feist	ct	operation
state	separately	plan
court	incorporate	state
letter	protection	baker
registration	star	brand
unprotectable	usc	merger
cir	imagine	include
qualify	conceptually	method
standard	design	chain
form	exist	data
common	compendium	mark
author	conceptual	doctrine
claim	physically	blank
selection	statute	llc
act	tcw	provide

Run 3		
Topic 0	Topic 1	Topic 2
elements	white	article
design	diamond	feature
office	gold	utilitarian
register	recycle	shape
id	visa	sculptural
shape	agents	graphic
protection	tcw	protection
creativity	franchisee	design
combination	18k	pictorial
authorship	computer	separable
creative	earrings	office
consist	program	artistic
board	tc	separate
arrangement	necklace	aspects
compendium	yellow	separability
original	fossilize	function
sufficient	collection	athletica
color	bk	star
find	ivory	incorporate
copyrightable	chain	test
feist	rise	ct
court	walrus	usc
state	flag	separately
letter	grey	id
registration	office	compendium
unprotectable	slice	elements
cir	moniquepean	imagine
form	brand	protect
qualify	mark	conceptually
common	llc	request
standard	stitch	board
selection	al	exist
expression	cut	conceptual
author	top	cir
cfr	ring	statute

Figure 2. Topics and associated words generated from three "runs" of MALLET.

1	filename	0 elements design office	1 article feature utilitarian	2 idea visa expression
137	)21-07-19_sculpture-of-locker (1)_processed.txt	0.6477	0.3522	0.0001
138	2016-09-27_b291-bed_processed.txt	0.6448	0.3551	0.0001
139	2021-06-29_manchettes-et-al (1)_processed.txt	0.6373	0.3626	0.0001
140	2021-06-04_led-collection (1)_processed.txt	0.6372	0.3627	0.0001
141	_fiore-sculpture_processed.txt	0.6318	0.3680	0.0002
142	19-09-27_Avalon-B481-09272019_processed.txt	0.6296	0.3703	0.0001
143	2021-07-26_walker-edison (1)_processed.txt	0.6279	0.3720	0.0001
144	2021-05-27_bride-headband (1)_processed.txt	0.6188	0.3794	0.0018
145	azStreetwise(dot)mib-09272019_processed.txt	0.6170	0.0003	0.3827
146	017-05-09_converse-flow-depths_processed.txt	0.6160	0.0019	0.3820
147	2-14_artistic-features-of-fast-xp_processed.txt	0.6143	0.3856	0.0001
148	9-23_kiesel-treated-finger-board_processed.txt	0.6118	0.3880	0.0002
149	2021-06-14_wave-form (1)_processed.txt	0.6107	0.3892	0.0002
150	2019-02-27_glove-overlay_processed.txt	0.6051	0.3947	0.0001
151	2020-06-26_pillow-sculpture_processed.txt	0.5923	0.2876	0.1201

Figure 3. The topics-metadata file.

Conditional formatting was applied in Excel to note in red font the topics to which over half of a given letter were associated, or – stated differently – to note in red the predominant topic represented in each document.[18] The data was then sorted by topic so the that the topic predominant across all of the letters was filtered to the top. This was done for each output in the set of three. Finally, the total number of documents that were primarily associated with a single topic were tallied by topic. Figure 4 displays the results, indicating that the distributions were also relatively stable.

Topic 0	Topic 1	Topic 2
10	7	154
153	6	12
158	7	6

Figure 4, representing the total number of documents primarily associate with each topic.

Having determined that results did appear to be stable, the next step was to randomly select one set of output for analysis. Using an online random number generator, the number output was 2, thereby identifying the set of returns that would be used in this study. The topics and topic words generated from the second run of MALLET are listed in Figure 5. These and the other documents output in the second run form the basis for the analysis that follows.

[18] Green font was similarly used to denote the next class of associations, applied to represent proportions between 0.3000 and 0.4999. This served as a visual cue only.

Topic 0	Topic 1	Topic 2
elements	article	idea
design	feature	visa
office	utilitarian	expression
register	sculptural	agents
id	white	system
shape	graphic	franchisee
protection	separable	book
creativity	pictorial	principle
combination	diamond	program
authorship	artistic	computer
board	separability	describe
creative	shape	section
consist	separate	show
compendium	function	flag
arrangement	gold	ideas
original	office	methods
sufficient	aspects	text
color	athletica	list
copyrightable	test	deposit
find	recycle	material
feist	ct	operation
state	separately	plan
court	incorporate	state
letter	protection	baker
registration	star	brand
unprotectable	usc	merger
cir	imagine	include
qualify	conceptually	method
standard	design	chain
form	exist	data
common	compendium	mark
author	conceptual	doctrine
claim	physically	blank
selection	statute	llc
act	tcw	provide

Figure 5, displaying the top 35 words most strongly associated with each topic analyzed in this study.

In their article on the applicability of topic modeling to policy research, Isoaho, Gritsenko and Mäkelä outline the necessity of an iterative validation process, recommending that researchers working to analyze topic modeling results cycle through three nonlinear steps: validating preprocessing decisions, interpreting output and evaluating "the ability of the topics to

model the phenomenon under investigation."[19] A recursive cycling through these latter two was performed across the MALLET output, allowing the researcher to regularly cross-reference between topic words output and the documents most strongly associated with each topic. To start, topics.metadata.csv was sorted such that the topic most prominently distributed across the corpus was weighted towards the top. Comprising a majority proportion of 158 of the 171 documents analyzed, the topic most prominently distributed across the corpus as output in the second run is Topic 0. The second-most prominently distributed – one which comprises a majority proportion of seven of the documents analyzed – is Topic 1. The least prominent topic across the corpus, comprising a majority of only six documents, is Topic 2.

Beginning with Topic 0 and working downward in distribution through Topics 1 and 2, each topic was analyzed individually, beginning with a close reading of the documents most strongly associated with each topic and noting text corresponding to words in associated topic lists as the analysis progressed. For purposes of scope and expediency, the analysis provided here will be limited to the ten documents most strongly associated with each topic. These will be occasionally referred to as each topic's "top ten".

Please be aware that the U.S. Copyright Office reviews registrations for hundreds of thousands of works each year. These assume all media, shapes, sizes and manners, including those that are adult in nature. One of these more sensitive works fell into the categories analyzed herein. It will be discussed as the matter of fact that it is, using the same language as the rejection letter returned by the U.S. Copyright Office Review Board. Should the reader feel more comfortable forgoing this discussion, it is advised to avoid the analysis of Topic 1.

[19] Isoaho, Gritsenko, and Mäkelä, "Topic Modeling and Text Analysis for Qualitative Policy Research," 306.

Results

Topic 0

Topic Words: elements, design, office, register, id, shape, protection, creativity, combination, authorship, board, creative, consist, compendium, arrangement, original, sufficient, color, copyrightable, find, feist, state, court, letter, registration, unprotectable, cir, qualify, standard, form, common, author, claim, selection, act

As mentioned above, Topic 0 is the most prominent topic distributed across the corpus of documents analyzed. Nearly 94% of the documents are predominantly associated with this topic, the top ten of which are detailed in Figure 7. The numbers in red font in the right-most column of Figure 6 represent the proportion of each document that was associated with the topic. This means that, per the idea of a generative probabilistic model, 99.94% of the CommVault Hexagon letter could be generated from the terms in Topic 0. Due to the fact that so many documents were primarily associated with this topic, many of them shared the same proportion of it. Across the top ten most strongly associated documents there are only three unique proportion values, providing good indication that the works submitted and their respective letters share many features in common.

0 Elements Design Office		
Title	Description	Topic Proportion
CommVault Hexagon	Two images of black parallellograms and triangles situated inside the silhouettes of boxes. The triangle in one is red.	0.9994
Academy Logo	Logo of the Academy of Motion Picture Arts - A black block "A", the center of which takes the shape of the Oscar statue.	0.9994
Equilibrium	Images of geometric shapes "colored in" by fingerprints	0.9994
LA Rocks	Silvery jewelry incribed with collection of words ("brave", "strong", "happy") and sentences.	0.9994
Tire Tread	Small 3D sculptures in the form of tires with differing tread.	0.9993
Wonkey Key	Geometric scroll pattern	0.9993
P-Tech	Blue and white text/graphic logo including the term "P-tech" and concentric circles.	0.9993
Blacka	Block-lettered word "BLACKA", with the "A"s crossed by two dots rather than by bars.	0.9993
Jaipur Link	Chain necklace comprised of O-shaped, hammered gold-tone links.	0.9993
Core Kitchen	A logo reading "Core Kitchen" in standard font with "O" in the shape of a spoon.	0.9992

Figure 6. Topic 0 "top ten" documents.

This was indeed found to be the case, as the works referenced include five commercial logos, two graphical creations and three three-dimensional objects.[20] Where text is utilized in these works, it tends to be at least as decorative as it is communicative. In some cases – such as the black, block-lettered "A" serving as the logo for the Academy of Motion Picture Arts – the fact that a letter comprises the majority of the work appears almost incidental to the effect of the work as a whole (see Figure 7).[21]

[20] U.S. Copyright Office Review Board, "COMMVAULT HEXAGON," December 12, 2017, https://www.copyright.gov/rulings-filings/review-board/docs/commvault-hexagon-graphic-artwork.pdf; U.S. Copyright Office Review Board, "Academy Logo," May 7, 2020, https://www.copyright.gov/rulings-filings/review-board/docs/academy-logo.pdf; U.S. Copyright Office Review Board, "Equilibrium et al," August 7, 2020, https://www.copyright.gov/rulings-filings/review-board/docs/equilibrium.pdf; U.S. Copyright Office Review Board, "LA Rocks," June 27, 2017, https://www.copyright.gov/rulings-filings/review-board/docs/la-rocks.pdf; U.S. Copyright Office Review Board, "Tire Tread," December 27, 2019, https://www.copyright.gov/rulings-filings/review-board/docs/tire-tread.pdf; U.S. Copyright Office Review Board, "WONKEY KEY," January 31, 2018, https://www.copyright.gov/rulings-filings/review-board/docs/wonkey_key.pdf; U.S. Copyright Office Review Board, "P-Tech USA Original Logo," September 9, 2019, https://www.copyright.gov/rulings-filings/review-board/docs/p-tech.pdf; U.S. Copyright Office Review Board, "Blacka," December 27, 2019, https://www.copyright.gov/rulings-filings/review-board/docs/blacka.pdf; U.S. Copyright Office Review Board, "JAIPUR LINK Necklace," April 19, 2018, https://www.copyright.gov/rulings-filings/review-board/docs/jaipur-link.pdf; U.S. Copyright Office Review Board, "Core Kitchen," January 19, 2020, https://www.copyright.gov/rulings-filings/review-board/docs/core-kitchen.pdf.

[21] U.S. Copyright Office Review Board, "Academy Logo," May 7, 2020, 1.

Figure 7. Image credit to the Academy of Motion Picture Arts.

While the works submitted are not themselves entirely homogenous, the grounds on which their bids for copyright were rejected are: the works were each found to lack the "creative authorship" needed to secure copyright protection. A variation on this phrase is found in the "Analysis of the Work" sections of each document analyzed in reference to Topic 0, usually in the first sentence. Whether for want of "requisite authorship", "creative authorship" or "requisite originality", each work described by the top ten documents associated with Topic 0 was rejected for the same reason and with the same language deployed at similar locations throughout each letter.[22]

The rejection letters that comprise the corpus used in this study share a relatively standard format. They each contain similar sections – Administrative Review, Discussion (including Legal Framework and Analysis of the Work), and Conclusion – and similar precedent-setting case law is leveraged in illustration of the positions taken. Regular boiler-plate language is used per case referenced such that if the case cited is, for example, *Atari Games Corp. v. Oman,* 888 F.2d 878, it generally is so in reference to the same types of disqualifying factors and utilizing

[22] U.S. Copyright Office Review Board, "Equilibrium et al," August 7, 2020, 5; U.S. Copyright Office Review Board, "LA Rocks," June 27, 2017, 8; U.S. Copyright Office Review Board, "COMMVAULT HEXAGON," December 12, 2017, 4.

the same few paragraphs of text.[23] The sum effect of each of these factors is what led to the similar topic proportions across these documents: because the works were rejected for similar reasons, the letters written to communicate this rejection are also, in some sections, nearly synonymous copies of one another.

One of the most common prior cases referenced within Topic 0's "top ten" documents is *Feist Publications v. Rural Telephone Service Company*, in which the U.S. Supreme Court heard a case regarding alleged copyright infringement of text printed in a white pages telephone directory.[24] The court found that the content of white pages – simple, alphabetical listings of individual names and phone numbers amounting to little more than itemized facts – failed to meet criteria for original or creative authorship, therefore did not qualify as registrable works and thus could not be subject to infringement.[25] This lack of original authorship is also what disqualified each of the works referenced in the top ten documents associated with Topic 0.

However, it is not simply original authorship – or the lack thereof – that sets these documents apart as a cohort. Instead, their similarities owe to the reason why the works were deemed to lack sufficient original authorship in the first place. In the "Analysis of the Work" sections of each of the respective documents, it emerges that the works were all deemed lacking in original authorship due to the relationship between the works themselves and the component elements of which they are comprised. Effectively, it is made clear that while "common elements" such as geometric shapes are not themselves protected by copyright, an arrangement

[23] Atari Games Corp. v. Oman, 888 F.2d 878, No. 88-5296 (D.C. Cir 1989).

[24] U.S. Supreme Court, Feist Publications, Inc. v. Rural Telephone Service Company, Inc., No. 499 U.S. 340 (U.S. Supreme Court March 27, 1991).

[25] Oyez, "Feist Publications, Inc. v. Rural Telephone Service Company, Inc.," Oyez.org, accessed January 4, 2022, https://www.oyez.org/cases/1990/89-1909.

of them that is sufficiently creative and/or original may be. Originality, therefore, is not as important when considered in light of the substantive pieces of the work as is their combination.

Each of the works referenced in the Topic 0 documents analyzed here is in fact composed of simple elements such as geometric shapes arrayed in a manner that is not itself evidently creative (see the CommVault Hexagon logo in Figure 8 for an example). Language referencing this relationship is used across each of the relevant letters, as in this phrase from *Academy Logo*: "[s]ome combinations of common or standard design elements may contain sufficient creativity with respect to how they are juxtaposed or arranged to support a copyright," explains the Review Board, but failing this successful arrangement as they did, the works referenced here did not qualify for protection.[26] As a result, words such as "elements," "shape," "combination," "arrangement," "color," "standard," "common," and "selection" feature repeatedly across the "Analysis of the Work" discussions in each of the documents reviewed, accounting as well for their inclusion in the topic words list of the 35 most strongly associated with the topic.[27]

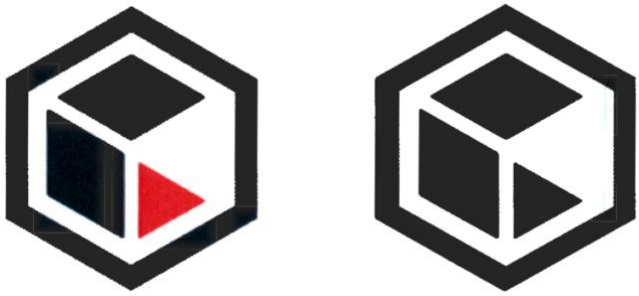

Figure 8. Image credit to CommVault Systems LLC.

[26] U.S. Copyright Office Review Board, "Academy Logo," May 7, 2020, 2.

[27] U.S. Copyright Office Review Board, "Equilibrium et al," August 7, 2020, 5; U.S. Copyright Office Review Board, "Academy Logo," May 7, 2020, 5; U.S. Copyright Office Review Board, "Blacka," December 27, 2019, 4; U.S. Copyright Office Review Board, "Tire Tread," December 27, 2019, 5; U.S. Copyright Office Review Board, "LA Rocks," June 27, 2017, 9; U.S. Copyright Office Review Board, "COMMVAULT HEXAGON," December 12, 2017, 4–5; U.S. Copyright Office Review Board, "JAIPUR LINK Necklace," April 19, 2018, 5; U.S. Copyright Office Review Board, "P-Tech USA Original Logo," September 9, 2019, 4; U.S. Copyright Office Review Board, "Core Kitchen," January 19, 2020, 3; U.S. Copyright Office Review Board, "WONKEY KEY," January 31, 2018, 4.

Topic 1

Topic 1: article, feature, utilitarian, sculptural, white, graphic, separable, pictorial, diamond, artistic, separability, shape, separate, function, gold, office, aspects, athletica, test, recycle, ct, separately, incorporate, protection, star, usc, imagine, conceptually, design, exist, compendium, conceptual, physically, statute, tcw

Only seven – or about 4% – of the 171 documents comprising this study's corpus were predominately associated with Topic 1, meaning that three of the ten documents most strongly associated with the topic were associated at a rate of less than 50%. This itself means that less than half of each of those documents could be considered to have been generated from this "bag of words". Nevertheless, a clear trend is evident across the ten documents analyzed in reference to this particular topic.

That trend has to do not with the utility of the related works – as might be suggested by the order of the topic words generated – but instead by the applicability of the word "separate" to their essential natures. "Separate," as both an adjective and a verb, played an essential role in the Review Board's considerations and surfaces in four of the returned 35 topic words. While this is true across each of the documents associated with this topic, the particular way in which "separate" is used in the most strongly associated document – the letter related to Solcin Atelier's jewelry collection – is an outlier among the rest. The discussion of Topic 1 therefore begins with the less strongly associated documents.

1 Article Feature Utilitarian		
Title	Description	Topic Proportion
Solcin Atelier	Collection of jewelry	0.9893
L322 Delta Wing	Wheel covers (2) for Jaguar Land Rover	0.8419
Senta	Wheel cover (1) for Jaguar Land Rover	0.8376
Mini Keg Growler	Small keg-shaped container for carrying draft beer	0.8189
Bellocq Tea Caddy	Metal cannister with a twist-off lid	0.7210
Aviator Sculpture	Nine works of housewares, including a chair, a chest and several chandelier-style lamps	0.5681
Ribbon Sculpture	Light fixture comprised of cord, socket and shade constructed from a tangle of translucent red ribbon-like plastic	0.5213
3 Bangs	Three sex toys	0.4895
Ion IQ Headset	Headset with single ear phone and a microphone extended from the earphone to the mouth	0.4738
Zig Zag Chandelier	Cylindrical light fixture in metal with cut-outs within which dangle crystal prisms	0.4680

Figure 9. Topic 1's "top ten" documents.

These letters reference works including wheel covers for the Jaguar Land Rover, a collection of furniture and other housewares, two individual light figures, a headphone/microphone combination headset, two unrelated cannister-shaped containers and a collection of sex toys. As was true of Topic 0 and notwithstanding some variation, these letters are noticeably similar to one another, all containing the same key phrases in effectively the same order. Among these, two stand out as being particularly prominent. The first of these is the phrase "copyright law does not protect useful articles". Relatedly, the second is a variation on "does not contain the requisite separable authorship necessary".[28] The reason for the relation

[28] U.S. Copyright Office Review Board, "L322 Delta Wing et al," August 24, 2016, 3–4, https://www.copyright.gov/rulings-filings/review-board/docs/l322-delta-wing.pdf; U.S. Copyright Office Review Board, "Senta," June 30, 2016, 3–4, https://www.copyright.gov/rulings-filings/review-board/docs/senta.pdf; U.S. Copyright Office Review Board, "Mini-Keg Growler," August 15, 2016, 2–3, https://www.copyright.gov/rulings-filings/review-board/docs/mini-keg-growler.pdf; U.S. Copyright Office Review Board, "Bellocq Tea Caddy," August 24, 2016, 2–3, https://www.copyright.gov/rulings-filings/review-board/docs/bellocq-tea-caddy.pdf; U.S. Copyright Office Review Board, "Ribbon Sculpture Design B," October 13, 2016, 2–4, https://www.copyright.gov/rulings-filings/review-board/docs/ribbon-sculpture-design.pdf; U.S. Copyright Office Review Board, "Aviator Sculpture et al," January 23, 2017, 4–7, https://www.copyright.gov/rulings-filings/review-board/docs/aviator-sculpture.pdf; U.S. Copyright Office Review Board, "3 Bangs for Your Butt et al," November 30, 2016, 3–5, https://www.copyright.gov/rulings-filings/review-board/docs/3-bangs-for.pdf; U.S. Copyright Office, "Review Board Opinions," 3–4; U.S. Copyright Office Review Board, "Zig Zag Chandelier," July 19, 2016, 2–5, https://www.copyright.gov/rulings-filings/review-board/docs/zig-zag-chandelier.pdf.

between these two is that while it is the case that copyright law does not cover purely useful works, it is also the case that the law can be applied to those creative components of a work that can be successfully separated – either physically or conceptually – from that work. This is best explained using the Review Board's own language:

> To satisfy the test for physical separability, a useful article must contain pictorial, graphic, or sculptural features that can be physically separated from the article by ordinary means…. To satisfy the test for conceptual separability, a useful article must contain pictorial, graphic, or sculptural features that can be visualized either on paper or as a free-standing sculpture-as a work of authorship that is separate and independent from the utilitarian aspects of the article and the overall shape of the article.[29]

Without such separability, creative authorship cannot be claimed. As per the discussion regarding Topic 0, without creative authorship, copyright protection cannot apply. It is due in part to this chain of dependencies that the related works were denied protection. The phrase "in part" is added here because three of the works included in this discussion were denied registration for dual reasons. These three works – *3 Bangs*, *Ion IQ Headset* and *Zig Zag Chandelier* – are those that fell lowest in the topic proportion allocation. Topic 1 is therefore not the dominant topic across those documents. Rather, the dominant topic across these documents is Topic 0. Accordingly, the relevant letters contain discussion of creativity as well as discussion of separability. A passage from the "Analysis of the Work" section of *3 Bangs* illustrates well how a discussion of the lack of separability combined with a lack of creative authorship might be suggestive of an association across Topics 1 and 0:

[29] U.S. Copyright Office Review Board, "Mini-Keg Growler," August 15, 2016, 3.

> [N]one of the folds or ridges making up the shape of Dark Vader… are separable, as they constitute the basic shape of the article. However… the oval and triangle imprints on the lower three-quarters of the work are able to be visualized separately and independently from the sex toy without destroying the basic shape of the work…. Of course, for a separable element of a work to be eligible for copyright protection, it must possess more than a "*de minimis* quantum of creativity." *Feist,* 499 U.S. at 363. The imprint on Dark Vader does not meet this low threshold.[30]

Returning to the document that was most strongly associated with Topic 1, it is well to remember that topic modeling in general is agnostic to meaning. This is why it works equally well over different languages and over numerals – the models produced are not models of meaning but instead models of relationships between tokens. The works described in the rejection letter comprising the Solcin Atelier document are pieces of jewelry. The body of the rejection letter as written by the Review Board does not directly address the pieces individually as the unit submitted and therefore considered by the Board was a collection.[31] Regardless of this single unit of submission, the pieces of jewelry are depicted and described individually in the Appendix. This Appendix is 46 pages long (85% of the letter's total 54 pages) and describes approximately 172 pieces of jewelry, resulting in a high level of repetition of words such as "diamonds," "recycled," "white" and "gold".[32] It is because the collection was submitted for registration rather than the individual pieces within it that the bid for registration was rejected. The subsequent discussion comprising much of the body of the letter pertains, then, to the differences in nature between *separate* pieces and a composite whole. The adjective form of

[30] U.S. Copyright Office Review Board, "3 Bangs for Your Butt et al," November 30, 2016, 6; citing U.S. Supreme Court, Feist Publications, Inc. v. Rural Telephone Service Company, Inc., No. 499 U.S. 340 (U.S. Supreme Court March 27, 1991).

[31] U.S. Copyright Office Review Board, "Solcin & Atelier 2015 Collection," August 3, 2017, https://www.copyright.gov/rulings-filings/review-board/docs/solcin-atelier.pdf.

[32] U.S. Copyright Office Review Board, 8–54.

"separate" is used multiple times throughout the letter, conflating in the "bag of words" with the verb form of "separate" that is used so regularly in other documents' discussion of the separability of useful and creative elements. It is suspected that the combination of the heavily weighted use of jewelry-related terms and the unusual and repeated usage of "separate" as an adjective served to place *Solcin Atelier* at the top of the list of those associated with Topic 1.

Topic 2

Topic Words: idea, visa, expression, agents, system, franchisee, book, principle, program, computer, describe, section, show, flag, ideas, methods, text, list, deposit, material, operation, plan, state, baker, brand, merger, include, method, chain, data, mark, doctrine, blank, llc, provide

Topic 2 emerges as the noisiest and most conglomerate of the three. It is also the thinnest, with only 25 of the 171 documents comprising the corpus associated with it in proportions over 10%. Despite its limited reach and diverse assortment, Topic 2 remains a grouping indicative of a perceptible context. Its context is one in which there was something fundamentally wrong with the works submitted: either they were lacking in copyrightable substance or they were submitted in unacceptable form.

Topic 2: Idea, Visa, Expression		
Title	Description	2 idea visa expression
2015 EHIM Essentials List	Alphabetical listings of pharmaceutical schedules by state	**0.7786**
Sunset Lumenscript et al	Data files	**0.7695**
Range et al	Printed carpet designs	**0.6943**
2017 Restaurant Directory et al	Directories of restaurant-related businesses	**0.6659**
Visa Flag Symbol Work et al	Branding instructions for use of Visa logos	**0.6089**
Misc. CA Statutes	Revisions to compilations of CA state insurance regulations	**0.5391**
Dubai Frame	Building in the shape of the frame of a rectangle	0.4841
NazStreetwise.mib	Computer program	0.3827
Converse Flow Depths	jpeg images depicting formulas and illustrations of water flow patterns	0.3820
Canetti Cursivus Guide for Cursive Writing and Calligraphy	A workbook intended for use by those learning calligraphy and cursive	0.3032

Figure 10

Towards the latter, two documents falling into Topic 2's "top ten" describe works that were submitted for registration in formats that effectively ensured their rejection. These were *Visa Flag Symbol Work et al* and *Range*. The *Range* submission was made seeking protection over printed carpet designs, the original submission for which provided photographs of the carpets installed on floors as well as scans of patterns used by the tufting machines in the carpets' manufacturing process. In their response, the Review Board references the *Compendium* multiple times, citing its guidelines in making clear that the proper unit of deposit would have instead been a swatch of the printed carpet.[33]

Whereas format was the issue with *Range*, it was the scope of materials submitted pertaining to *Visa Flag Symbol et al* that negated its bid for copyright. In that case, the works submitted were instruction sheets outlining proper branding guidelines for the use of Visa logos – sheets that contained two-dimensional depictions of the Visa flags as well as lengthy written text including instructions for their use and creation. The letter accompanying the original

[33] U.S. Copyright Office Review Board, "Range et al," April 13, 2017, 3, https://www.copyright.gov/rulings-filings/review-board/docs/range.pdf; U.S. Copyright Office, "Compendium of U.S. Copyright Office Practices § 101."

submission requested copyright protection over just the Visa Logos, which the Review Board determined to be lacking in sufficiently creative elements to be protectible. Had the request included consideration of the written instructions, however, protection may have been granted.[34] The body of the *Visa Flag Symbol* letter therefore illuminates the inclusion of "compendium" and "deposit" in Topic 2's top 35 terms. The Appendix of that letter helps to account for the words "visa," "flag," "show," and "brand", as the Appendix includes copies of the original submissions in which each of those words is used with great regularity (some over 100 times).

Returning to the cases of works lacking in copyrightable subject material, an evident majority of these relate to works that ran afoul of what is known as the "idea/expression" dichotomy. As discussed previously, copyright protections cover "original work[s] of authorship fixed in any tangible medium of expression."[35] Authorship, however, must be original, creative and the result of human (as opposed to machine, animal or otherwise natural) activity – it will not qualify as such if it is a simple, unembellished expression of fact. Likewise, to be a work, it must be fixed in a tangible medium, effectively drawing a line between unprotectible ideas and protectible expressions.

Four of the documents in the "top ten" of Topic 2 pertain to works that did not meet the standards set forth by the idea/expression dichotomy. These include *Converse Flow Depths*, *nazStreetwise.mib*, the *Canetti Cursivus Guide* and *Misc. CA Statutes*. Each letter regarding these works includes discussion of a case entitled *Baker v. Selden* (101 U.S. 99), in which it was

[34] U.S. Copyright Office Review Board, "Visa Flag Symbol Work et al," February 7, 2017, 5, https://www.copyright.gov/rulings-filings/review-board/docs/visa-flag-symbol-work.pdf.

[35] 17. U.S. Code 17 (1990), §§ 102

determined that the "system" outlined in a particular bookkeeping book – one that consisted, as printed and sold, merely of blank lines – could not be protected by copyright because what comprised its substance effectively amounted to an idea or system rather than a work of authorship.[36] As a result, discussion of the idea/expression dichotomy across the relevant rejection letters regularly employs the words "idea," "expression," "baker," "book" and "system", helping to account for the emergence of those terms among the top 35 of Topic 2.[37]

Closely related to the idea/expression dichotomy is what is known as "the merger doctrine". In copyright law, the merger doctrine provides a subtle but vital distinction between 1) the dichotomy standing between ideas and their expressions and 2) the conflation of an idea with its expression. This conflation – or merging – occurs when there is effectively only one way or a small number of ways to express an idea.[38] In such cases, that expression is not copyrightable. It was observed that every document in Topic 2's "top ten" that references the idea/expression dichotomy also references the merger doctrine, accounting for the inclusion of an additional three terms – "merger," "doctrine" and "principle" – in the top 35 terms associated with Topic 2.

[36] Justia, Baker v. Selden, No. 101 U.S. 99 (U.S. Supreme Court 1879).

[37] U.S. Copyright Office Review Board, "Converse Flow Depths," May 9, 2017, 3, https://www.copyright.gov/rulings-filings/review-board/docs/converse-flow-depths.pdf; U.S. Copyright Office Review Board, "NazStreetwise.Mib," September 27, 2019, 5, https://www.copyright.gov/rulings-filings/review-board/docs/nazStreetwise(dot)mib-09272019.pdf; U.S. Copyright Office Review Board, "Canetti Cursivus Guide for Cursive Writing and Calligraphy," September 27, 2016, 2, https://www.copyright.gov/rulings-filings/review-board/docs/canetti-cursivus-guide.pdf; U.S. Copyright Office Review Board, "Misc CA Statutes," December 12, 2017, 3, https://www.copyright.gov/rulings-filings/review-board/docs/misc-ca-statues.pdf.

[38] U.S. Copyright Office Review Board, "Converse Flow Depths," May 9, 2017, 3; U.S. Copyright Office Review Board, "NazStreetwise.Mib," September 27, 2019, 5; U.S. Copyright Office Review Board, "Canetti Cursivus Guide for Cursive Writing and Calligraphy," September 27, 2016, 2; U.S. Copyright Office Review Board, "Misc CA Statutes," December 12, 2017, 3.

Discussion

Through the application of topic modeling to a collection of rejection letters written by the U.S. Copyright Office Review Board and received by that many unrequited applicants, this study has suggested three contexts in which bids for copyright protection are rejected. In the first of these contexts – that which is represented by Topic 0 – the selection, combination and arrangement of elements in submitted works are insufficiently original. In the second – illustrated by Topic 1 – the works submitted fail the separability test that endeavors to find independent creative elements within otherwise utilitarian objects. In the third – suggested by Topic 2 – either substantive or procedural errors are made.

This study addresses copyright law and its application. However, if understood from the vantage afforded by a broader understanding of intellectual property law, it becomes possible to perceive some broader trends at work. Understood in the broadest possible terms, the works addressed by copyright law are works of creative authorship. Those addressed by patent law are utilitarian and those addressed by trademark are, in essence, commercialized. Specifically, trademarks protect "any word, name, symbol, or design used in commerce to identify and distinguish the goods of one manufacturer or seller from those of another."[39] While applicable to packaging, odors and a variety of other broadly understood "media," trademarks are often recognized as logos.

Logos such as P-Tech's (featured below in Figure 11), CommVault Hexagon Systems' (above, Figure 8) and the Academy of Motion Picture Arts (above, Figure 7) each meet these high-level criteria, as likely do the works referenced in three referenced in Topic 0's most

[39] U.S. Code 15 (1999), §§ 1127

strongly associated documents – *Blacka, Tire Tread* and *Core Kitchen*.[40] While the combination of their particular component parts might have lacked sufficient creative authorship to qualify for copyright protection, they may well have qualified as trademarks, thereby affording the protection that their creators sought. Likewise, if the objects described in the documents comprising Topic 1 were to be considered as works of intellectual property, it could be argued that they would be better suited for patent consideration than for copyright. Such bids might still be rejected, either due to "prior art" (evidence that the useful object has, in essence, already been invented) or due to a lack of originality, but the objects' utilitarianism arguably makes for a better IP fit with patent rather than copyright protection.

Figure 11. Image credit China Motors and Components, Inc.

Interestingly, Zvi S. Rosen, the scholar whose work was cited earlier as the inspiration behind this study, observes similar trends in some of his more recent work on the history of U.S. copyright registration. In a preprint posted to SSRN in May 2022, he meticulously details trends in copyright registration from 1781 to 2020, demonstrating among other things that visual works tend to fare poorly. Reflecting specifically on the period between 1959 and 1977, he surmises

[40] U.S. Copyright Office Review Board, "P-Tech USA Original Logo," September 9, 2019, 1; U.S. Copyright Office Review Board, "COMMVAULT HEXAGON," December 12, 2017, 1; U.S. Copyright Office Review Board, "Academy Logo," May 7, 2020, 1; U.S. Copyright Office Review Board, "Blacka," December 27, 2019, 1; U.S. Copyright Office Review Board, "Tire Tread," December 27, 2019, 2; U.S. Copyright Office Review Board, "Core Kitchen," January 19, 2020, 1.

that among submissions of visual works, it may be the case that "… in a great number of cases technical drawings… have been submitted in an attempt to get a pseudo-patent for a design of invention via copyright law". He continues, reflecting as well on artwork: "protection might be attempted for what is more properly a trademark, or simply lacking sufficient creativity." Rosen qualifies his observations as "supposition", for want of "more granular data".[41] This study benefits from – indeed, it was designed to leverage – extensive documentary data as well as the highly granular analysis afforded by topic modeling, but the data used emerged over fifty years after those referenced by Rosen in this specific passage. Nonetheless, the fact that these trends appear to have persisted over such a length of time suggests that Rosen's suppositions might indeed be accurate.

Less directly, this study's findings also appear to corroborate another of Rosen's positions. Throughout his work, Rosen proposes that copyright examination might be streamlined by the adoption of review procedures differentiated by type of work submitted.[42] Indeed, one of his most consistent observations is that the fate of submitted works often ultimately comes down to their form – to the type of work they are.[43] Some enjoy high levels of relatively seamless registration, while others – such as blank books and forms – have fared poorly over the years.[44] Others yet face unpredictable outcomes. If nothing else, Topic 2 illustrates this miscellany, simultaneously shedding light on a potential copyright illiteracy extending across what is sure to be a motley cohort of prospective registrants. Considered in light

[41] Zvi S. Rosen, "Examining Copyright," SSRN Scholarly Paper (Rochester, NY, May 4, 2022), 85, https://doi.org/10.2139/ssrn.4099976.
[42] Rosen, 91–94.
[43] Rosen, 3.
[44] Rosen, 33.

of the relative consistency demonstrated across Topics 0 and 1, this finding helps to substantiate Rosen's call for varied copyright review practices.

At this point it should be acknowledged that results are presented here with an awareness of some weaknesses in the study design and execution. To start, at 171, the number of documents comprising the corpus analyzed is quite small. Due to the probability model upon which it is premised, Latent Dirichlet Allocation has been shown to perform most successfully when run over datasets that number in the thousands of documents.[45] Further, many experts in the use of topic modeling have urged future users to employ the method only when it and its results are understood very well.[46] Without a solid understanding of the statistics that underlie it and without proper validation of both preprocessing decisions and results output, confirmation bias can undermine a researcher's work. This study was performed by a single researcher using LDA for the first time. While it was completed with the iterative guidance and feedback of a cohort of experts at the University of Pittsburgh, it remains likely that the researcher's still-maturing familiarity with topic modeling has affected its execution. Towards this end, it is requested that findings be considered preliminary. It is also recommended that exploration is continued into the use of topic modeling to study applied copyright law.

Conclusion

By using topic modeling to analyze select sections of letters from the U.S. Copyright Office Review Board recounting reasons for copyright rejection, this study suggests three broad

[45] Isoaho, Gritsenko, and Mäkelä, "Topic Modeling and Text Analysis for Qualitative Policy Research," 306.

[46] Isoaho, Gritsenko, and Mäkelä, "Topic Modeling and Text Analysis for Qualitative Policy Research"; Trevor J. Owens, "Discovery and Justification Are Different: Notes on Science-Ing the Humanities," *Trevor Owens* (blog), November 19, 2012, http://www.trevorowens.org/2012/11/discovery-and-justification-are-different-notes-on-sciencing-the-humanities/; Schmidt, "Words Alone: Dismantling Topic Models."

reasons why copyright was regularly denied between 2016 and 2021. These include a lack of original authorship evinced by the arrangement of component elements within a given work; insufficient independent creative aspects of predominantly utilitarian objects; and fundamental errors made, whether administrative or substantive. Considered together, the results of this analysis add to a small but consequential body of scholarship examining the contours of U.S. copyright registration and rejection. Specifically, they illuminate prior analyses of copyright examination and its implications. Considered more broadly, the results shared here contribute to a preliminary case for the use of LDA topic modeling in the study of applied copyright law. Deployed by well-trained researchers over a dataset of sufficient size, LDA topic modeling could prove to be a powerful tool in helping scholars investigate the ways in which copyright and intellectual property law have been leveraged over time, contributing to a broader understanding of how "creations of the mind" have been valued, devalued and commodified over time.

Bibliography

Allen, Colin, and Jaimie Murdock. “LDA Topic Modeling: Contexts for the History & Philosophy of Science.” Preprint, May 29, 2020. http://philsci-archive.pitt.edu/17261/.

Atari Games Corp. v. Oman, 888 F.2d 878, No. 88-5296 (D.C. Cir 1989).

Blei, David M. “Probabilistic Topic Models.” *Communications of the ACM* 55, no. 4 (April 2012): 77–84. https://doi.org/10.1145/2133806.2133826.

Burton, Matthew. “The Joy of Topic Modeling,” May 21, 2013. http://mcburton.net/blog/joy-of-tm/.

Enderle, Scott. “Quickstart Guide.” Topic Modeling Tool Blog, January 6, 2017. https://senderle.github.io/topic-modeling-tool/documentation/2017/01/06/quickstart.html.

Isoaho, Karoliina, Daria Gritsenko, and Eetu Mäkelä. “Topic Modeling and Text Analysis for Qualitative Policy Research.” *Policy Studies Journal* 49, no. 1 (2021): 300–324. https://doi.org/10.1111/psj.12343.

Justia. Baker v. Selden, No. 101 U.S. 99 (U.S. Supreme Court 1879).

Mimno, David, Charles Sutton, Gaurav Chandalia, and Al Hough. “Topic Modeling.” MALLET: Machine Learning for Language Toolkit. Accessed January 5, 2022. https://mimno.github.io/Mallet/topics.html.

Owens, Trevor J. “Discovery and Justification Are Different: Notes on Science-Ing the Humanities.” *Trevor Owens* (blog), November 19, 2012. http://www.trevorowens.org/2012/11/discovery-and-justification-are-different-notes-on-sciencing-the-humanities/.

Oyez. “Feist Publications, Inc. v. Rural Telephone Service Company, Inc.” Oyez.org. Accessed January 4, 2022. https://www.oyez.org/cases/1990/89-1909.

Rosen, Zvi S. “Examining Copyright.” SSRN Scholarly Paper. Rochester, NY, May 4, 2022. https://doi.org/10.2139/ssrn.4099976.

———. “Examining Copyright,” 1–56. Bournemouth, 2021. https://www.ishtip2021.org/links/link/examining-copyright.

Schmidt, Benjamin. “Words Alone: Dismantling Topic Models.” *Humanities Journal of Digital Humanities* 2, no. 1 (2012). http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/.

Thuronyi, George. “Copyright Office Launches Online Database of Review Board Decisions | Copyright: Creativity at Work.” Webpage, June 2, 2017.

//blogs.loc.gov/copyright/2017/06/copyright-office-launches-online-database-of-review-board-decisions/.

Underwood, Theodore. "What Kinds of 'Topics' Does Topic Modeling Actually Produce?" *The Stone and the Shell* (blog), April 1, 2012. https://tedunderwood.com/2012/04/01/what-kinds-of-topics-does-topic-modeling-actually-produce/.

U.S. Copyright Office. "Compendium of U.S. Copyright Office Practices § 101." 3rd Edition. U.S. Copyright Office, 2021. https://www.copyright.gov/comp3/.

———. "Overview of the Copyright Office." Copyright.gov. Accessed January 5, 2022. https://www.copyright.gov/about/.

———. "Review Board Opinions." Copyright.gov. Accessed January 5, 2022. https://www.copyright.gov/rulings-filings/review-board/index.html?loclr=blogcop.

U.S. Copyright Office Review Board. "3 Bangs for Your Butt et al," November 30, 2016. https://www.copyright.gov/rulings-filings/review-board/docs/3-bangs-for.pdf.

———. "Academy Logo," May 7, 2020. https://www.copyright.gov/rulings-filings/review-board/docs/academy-logo.pdf.

———. "Aviator Sculpture et al," January 23, 2017. https://www.copyright.gov/rulings-filings/review-board/docs/aviator-sculpture.pdf.

———. "Bellocq Tea Caddy," August 24, 2016. https://www.copyright.gov/rulings-filings/review-board/docs/bellocq-tea-caddy.pdf.

———. "Blacka," December 27, 2019. https://www.copyright.gov/rulings-filings/review-board/docs/blacka.pdf.

———. "Canetti Cursivus Guide for Cursive Writing and Calligraphy," September 27, 2016. https://www.copyright.gov/rulings-filings/review-board/docs/canetti-cursivus-guide.pdf.

———. "COMMVAULT HEXAGON," December 12, 2017. https://www.copyright.gov/rulings-filings/review-board/docs/commvault-hexagon-graphic-artwork.pdf.

———. "Converse Flow Depths," May 9, 2017. https://www.copyright.gov/rulings-filings/review-board/docs/converse-flow-depths.pdf.

———. "Core Kitchen," January 19, 2020. https://www.copyright.gov/rulings-filings/review-board/docs/core-kitchen.pdf.

———. "Equilibrium et Al," August 7, 2020. https://www.copyright.gov/rulings-filings/review-board/docs/equilibrium.pdf.

———. “JAIPUR LINK Necklace,” April 19, 2018. https://www.copyright.gov/rulings-filings/review-board/docs/jaipur-link.pdf.

———. “L322 Delta Wing et al,” August 24, 2016. https://www.copyright.gov/rulings-filings/review-board/docs/l322-delta-wing.pdf.

———. “LA Rocks,” June 27, 2017. https://www.copyright.gov/rulings-filings/review-board/docs/la-rocks.pdf.

———. “Mini-Keg Growler,” August 15, 2016. https://www.copyright.gov/rulings-filings/review-board/docs/mini-keg-growler.pdf.

———. “Misc CA Statutes,” December 12, 2017. https://www.copyright.gov/rulings-filings/review-board/docs/misc-ca-statues.pdf.

———. “NazStreetwise.Mib,” September 27, 2019. https://www.copyright.gov/rulings-filings/review-board/docs/nazStreetwise(dot)mib-09272019.pdf.

———. “P-Tech USA Original Logo,” September 9, 2019. https://www.copyright.gov/rulings-filings/review-board/docs/p-tech.pdf.

———. “Range et al,” April 13, 2017. https://www.copyright.gov/rulings-filings/review-board/docs/range.pdf.

———. “Ribbon Sculpture Design B,” October 13, 2016. https://www.copyright.gov/rulings-filings/review-board/docs/ribbon-sculpture-design.pdf.

———. “Senta,” June 30, 2016. https://www.copyright.gov/rulings-filings/review-board/docs/senta.pdf.

———. “Solcin & Atelier 2015 Collection,” August 3, 2017. https://www.copyright.gov/rulings-filings/review-board/docs/solcin-atelier.pdf.

———. “Tire Tread,” December 27, 2019. https://www.copyright.gov/rulings-filings/review-board/docs/tire-tread.pdf.

———. “Visa Flag Symbol Work et al,” February 7, 2017. https://www.copyright.gov/rulings-filings/review-board/docs/visa-flag-symbol-work.pdf.

———. “WONKEY KEY,” January 31, 2018. https://www.copyright.gov/rulings-filings/review-board/docs/wonkey_key.pdf.

———. “Zig Zag Chandelier,” July 19, 2016. https://www.copyright.gov/rulings-filings/review-board/docs/zig-zag-chandelier.pdf.

U.S. Supreme Court. Feist Publications, Inc. v. Rural Telephone Service Company, Inc., No. 499 U.S. 340 (U.S. Supreme Court March 27, 1991).