STANFORD
COMPUTATIONAL
ANTITRUST

ARTICLE

DOCTRINAL IMPLICATIONS OF COMPUTATIONAL ANTITRUST

Felix B. Chang,[*] Erin McCabe,[**] Zhaowei Ren,[***]
Joshua Beckelhimer,[****] James J. Lee.[*****]

Abstract. Utilizing antitrust decisions extracted from the Caselaw Access Project, we aggregate—or embed—layers of topic modeling into a single set of visualizations. Aggregated models can provide new perspectives on how courts tackle thorny doctrinal questions, such as the measure of market power and the balance between antitrust and regulation. Our central contribution is the improvement of natural language processing to provide greater context for key terms. Our secondary contribution is a new suite of tools to assess the weighty policy arguments that currently dominate antitrust.

* Professor of Law and Co-Director, Corporate Law Center, University of Cincinnati College of Law. Affiliated Fellow, Thurman Arnold Project, Yale School of Management. E-mail: felix.chang@uc.edu.
This project was supported by the Andrew W. Mellon Foundation. We are grateful to Julian Nyarko and Jens Frankenreiter for their insightful comments. Thanks, too, to Rosa Abrantes-Metz, Ned Cavanagh, David Donald, Ezra Edgerton, Harry First, Kelly Fitzpatrick, Peter Grajzl, James Grimmelmann, Jim Hart, Jyh-An Lee, Ephrat Livni, Mike Livermore, Thibault Schrepel, Danny Sokol, and Adam Ziegler.

This essay benefitted from the Online Workshop on the Computational Analysis of Law at Virginia, the Next Generation of Antitrust Scholars Conference at NYU, and the Machine Lawyering Conference at the Chinese University of Hong Kong. We thank Maura Carey, Martin Sicilian, Glen Williams, Ben Halom, Teodora Groza, Aleksandra Wierzbicka, and Alex Sotropa for their careful editing.

** Library Fellow, Digital Scholarship Center, University of Cincinnati.
*** Software Development Engineer, Amazon Web Services (previously, Software Development Engineer, Digital Scholarship Center, University of Cincinnati).
**** Department of English, University of Cincinnati.
***** Associate Professor, Digital Humanities, and Associate Vice Provost for Digital Scholarship, University of Cincinnati.

I. Introduction

This paper introduces modifications to topic modeling devised by the Digital Scholarship Center ("DSC") at the University of Cincinnati. Topic modeling, a form of natural language processing that depicts the probability distribution of terms over a corpus of texts,[1] has been heralded for its ability to process big data.[2] The tool depicts how words cluster around one another, thereby illuminating the textual foundations of legal doctrine. Frequently utilized in computational legal analysis ("CLA"),[3] topic modeling has helped to unearth patterns in judicial opinions,[4] loan agreements,[5] and national constitutions.[6]

For the past several years, DSC has been building a machine learning ("ML") platform that analyzes large datasets through variations on topic modeling.[7] In one variation, we aggregate—or embed—six levels of topic modeling into a single set of visualizations. Using aggregated modeling, the platform reveals linguistic patterns within a corpus of cases extracted from Harvard Law School's Caselaw Access Project ("CAP"), which has recently digitized almost all published decisions in the United States.[8] The ensuing visualizations translate topic modeling into intuitive models for users with little statistical or empirical training.

As a test, we have compiled a large pool of federal antitrust cases to see what our algorithms reveal of two thorny doctrinal questions: the measure of market power and the balance between antitrust and regulation. Our visualizations depict how thousands of market-power and antitrust–regulation cases cluster around different terms—as well as how these clusters have evolved over time. The legal doctrines around market power and the antitrust–regulation balance serve as a back-end check on the precision of aggregated modeling.

Aggregating topic models achieves greater contextualization by helping to generate visualizations that embed topics into neural networks of topic *clusters.* As

---

[1] *See* David M. Blei, Andrew Y. Ng, & Michael I. Jordan, *Latent Dirichlet Allocation*, 3 J. MACH. LEARNING RSCH. 993 (2003); Michael A. Livermore, Allen B. Riddell, & Daniel N. Rockmore, *The Supreme Court and the Judicial Genre*, 59 ARIZ. L. REV. 837, 841–42 (2017).

[2] *See* David S. Law, *Constitutional Archetypes*, 95 TEX. L. REV. 153 (2016); Peter Grajzl & Peter Murrell, *A Machine-Learning History of English Caselaw and Legal Ideas Prior to the Industrial Revolution I: Generating and Interpreting the Estimates*, 17 J. INST. ECON. 1 (2021); Peter Grajzl & Peter Murrell, *A Machine-Learning History of English Caselaw and Legal Ideas Prior to the Industrial Revolution II: Applications*, 17 J. INST. ECON. 201 (forthcoming 2021).

[3] *See* Michael A. Livermore & Daniel N. Rockmore, *Introduction: From Analogue to Digital Legal Scholarship*, *in* LAW AS DATA: COMPUTATION, TEXT, & THE FUTURE OF LEGAL ANALYSIS xvii (Michael A. Livermore & Daniel N. Rockmore eds., 2019).

[4] Livermore et al., *supra* note 1, at 841–42.

[5] Bernhard Ganglmair & Malcolm Wardlaw, Complexity, Standardization, and the Design of Loan Agreements (Apr. 13, 2017) (unpublished manuscript) (available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2952567).

[6] *See* Law, *supra* note 2.

[7] Our team has partnered with scholars from various fields on research questions such as the understanding of race during Shakespeare's era, the causes of placement disruption in foster care, links between race and carbon dioxide emissions in U.S. urban areas, pottery recovered in an early twentieth-century expedition to the city of Troy, and the extent of geographical publication bias in biology publications. For a full list, see *Projects*, UNIV. OF CIN. DIG. SCHOLARSHIP CTR., https://sites.libraries.uc.edu/dsc/research/projects (last accessed Apr. 27, 2021).

[8] *See About*, CASELAW ACCESS PROJECT, https://case.law/about/ (last accessed July 7, 2020).

the next sections reveal, clusters depict the relationship among topics, just as topics show the relationships among terms.

This is an opportune time to fuse CLA and antitrust into *computational* antitrust. Antitrust law has become immensely technical, so judges sometimes default to ideologies on judicial intervention and false positives to steer their decisions.[9] Now, legislative changes are afoot in the U.S.—or, at least, rumblings about legislative changes, driven by ire over big tech.[10] Yet despite bipartisan momentum, an overhaul of antitrust by legislation is uncertain, given political gridlock. Nonetheless, federal and state agencies are barreling ahead: In 2020, the DOJ, FTC, and state attorneys general sued Google and Facebook for a variety of anticompetitive practices, leaving the U.S. District Court for the District of Columbia to rule on the violations and craft remedies, a bedeviling proposition.[11] As in prior generations,[12] federal courts—with their broad injunctive powers—are once again becoming the forum for pushing against entrenched paradigms.

In this landscape, many scholars have advocated for reforming antitrust doctrine rather than overhauling it. Some call for simplifying antitrust law, for instance by restating its core goals.[13] If antitrust prohibits anticompetitive conduct that increases market power,[14] courts need only engage in two inquiries: whether a defendant is engaging in anticompetitive conduct and whether such conduct is likely to increase market power. Other scholars and advocates have singled out specific tenets, such as predatory pricing, duty to deal, and burdens of proof for reform. This camp would shore up essential facilities obligations, clarify antitrust immunity, relax or reverse plaintiffs' burdens of proof, and recalibrate the deference to direct versus circumstantial proof of anticompetitive effects.[15] The starting point for these proposals, though, is robust empirical analysis on where such doctrines currently stand. We aim to help build that foundation, by designing tools that assess how federal courts approach intractable doctrinal questions.

---

[9] *See, e.g.*, Verizon Commc'ns Inc. v. Law Offs. of Curtis V. Trinko, 540 U.S. 398, 414 (2004) ("Mistaken inferences and the resulting false condemnations are especially costly . . . . The cost of false positives counsels against an undue expansion of § 2 liability.") (internal quotations omitted) (quoting Matsushita Elec. Indus. Co. v. Zenith Radio Corp., 475 U.S. 574, 594 (1986)).

[10] *See* SUBCOMMITTEE ON ANTITRUST, H.R. COM. & ADMIN. L. OF THE COMM. ON THE JUDICIARY, 116TH CONG., INVESTIGATION OF COMPETITION IN DIGITAL MARKETS: MAJORITY STAFF REPORT AND RECOMMENDATIONS 133 (2020).

[11] *See* FTC v. Facebook, Inc., No. 1:20-cv-03590 (D.D.C. filed Dec. 9, 2020); New York v. Facebook, Inc., No. 1:20-cv-03589 (D.D.C. filed Dec. 9, 2020); United States v. Google LLC, No. 1:20-cv-03010 (D.D.C. filed Oct. 20, 2020).

[12] *See, e.g.*, U.S. Dept. of Justice, Modification of Final Judgment § II(A), *reprinted in* United States v. AT&T, 552 F. Supp. 131, 227 (D.D.C. 1982).

[13] In a series of articles, Doug Melamed has championed the simplicity in antitrust. *See* A. Douglas Melamed, *Antitrust Law and Its Critics*, 83 ANTITRUST L.J. 269 (2020); A. Douglas Melamed, *Antitrust Law Is Not That Complicated*, 130 HARV. L. REV. F. 163 (2017).

[14] *See* Michael L. Katz & A. Douglas Melamed, *Competition Law as Common Law:* American Express *and the Evolution of Antitrust*, 168 U. PA. L. REV. 2061, 2071–72 (2020).

[15] *See, e.g.*, LUIGI ZINGALES, GUY ROLNIK, & FILIPPO MARIA LANCIERI, STIGLER COMMITTEE ON DIGITAL PLATFORMS: FINAL REPORT (2019), https://www.chicagobooth.edu/-/media/research/stigler/pdfs/digital-platforms---committee-report---stigler-center.pdf.

II. AGGREGATING TOPIC MODELS

Topic modeling has already gained traction within CLA, so it is not unfamiliar to law scholars. Yet the technique has certain vulnerabilities, as digital humanists and computer scientists have previously pointed out. This Part summarizes the criticisms of topic modeling, as a preview to our improvements to traditional topic modeling algorithms.

A – Criticisms from Digital Humanities and Computer Science

Topic modeling illuminates patterns that cannot be seen by the human eye, at least not with traditional close readings of text. It is a form of distant reading, which considers texts "from afar, using statistics to support large-scale claims."[16] Distant reading can spur interesting collaborations on legal research, particularly in formulating the type of systematic review that can vet the claims of doctrinal work.[17]

Digital humanities ("DH") and computer science have lived with topic modeling far longer than law; there, criticisms of the tool are well-developed. Detractors of the computational approach to reading charge that it is "prone to fallacious overclaims or misinterpretations of statistical results because it often places itself in a position of making claims based purely on word frequencies without regard to position, syntax, context, and semantics."[18] More pointedly, the excitement around topic modeling merely stems from the fact that it seems to work better than other "rearrangement algorithms"; without the proper supervision, the tool resembles a "bad research assistant" that produces inexplicable and misleading results as much as "flickers of deeper truths."[19]

Context is therefore central to the viability of topic modeling. Robust visualizations must be able to show the texts from which the words are drawn—or, with legal texts, the cases that are statistically most likely to feature the terms that constitute a topic. Relatedly, it is possible to focus too much on a few discrete topics and lose the forest for the trees, so topics must be surveyed as a whole rather than in isolation.[20] The opposite is also true: Topic modeling can overwhelm users as much by the grandness of its topics (i.e., too many topics) as by the exquisiteness of its detail (i.e., too many terms within a topic). For this reason, a topic modeling interface must simultaneously be able to break topics down to their constituent words and aggregate them into networks. We respond to these critiques by building visualizations that can do both, as presented in the next Subpart.

---

[16] Michael A. Livermore & Daniel N. Rockmore, *Distant Reading and the Law*, *in* LAW AS DATA: COMPUTATION, TEXT, & THE FUTURE OF LEGAL ANALYSIS, *supra* note 3, at 3, 4. *See also* Lauren F. Klein, *Distant Reading after Moretti*, LAUREN F. KLEIN (Jan. 10, 2018), https://lklein.com/digital-humanities/distant-reading-after-moretti/.

[17] *See* Livermore & Rockmore, *supra* note 16, at 16; William Baude, Adam S. Chilton, & Anup Malani, *Making Doctrinal Work More Rigorous: Lessons from Systematic Reviews*, 84 U. CHI. L. REV. 37 (2017).

[18] Nan Z. Da, *The Computational Case against Computational Literary Studies*, 45 CRITICAL INQUIRY 601, 611 (2019).

[19] Benjamin M. Schmidt, *Words Alone: Dismantling Topic Models in the Humanities*, 2 J. DIGIT. HUMANS., Winter 2012, at 49, 50.

[20] *See* Andrew Goldstone & Ted Underwood, *What Can Topic Models of PMLA Teach Us about the History of Literary Scholarship?*, 2 J. DIGIT. HUMANS., 39 (2012).

Beyond decontextualization, DH and computer-science scholars have identified other deficiencies of topic modeling. For instance, some have argued that the computer scientists who created topic modeling intended it to perform functions quite different than what DH scholars have made it do.[21] David Blei, one of the pioneers of latent Dirichlet allocation ("LDA"), had envisioned topic modeling as an information retrieval algorithm.[22] Precursors of LDA, including most prominently latent semantic analysis from the 1990s, were designed for information retrieval and indexing as well.[23] However enticing it may be to harness topic modeling and other CLA techniques for prediction of, say, litigation outcomes, these tools might be better restricted to discrete retrieval, indexing, and archival functions in law, at least until legal scholars have a better grasp of their capabilities.[24]

Even if topic modeling is not used to forecast outcomes, it can fail simple robustness and reproducibility checks. Scholars have shown that if a corpus of text is changed slightly (e.g., 1% of the original sample removed), the ensuing topics are entirely different.[25] Similarly, the modeling sampled in prominent DH papers have not always withstood reproduction by others.[26] These methodological concerns question whether topic modeling is stable and verifiable.

B – Aggregated Modeling

We take seriously the criticisms levied at topic modeling from DH and computer science. Accordingly, we prefer to aggregate multiple LDA topic models in one iteration. In this way, we create a "model-of-models" that addresses some of the contextualization, robustness, and reproducibility concerns surrounding the tool. Several improvements to traditional topic models flow from their aggregation. First, our visualizations place topics in both large and small contexts.

[21] Schmidt, *supra* note 19.
[22] *Id.*
[23] *See* Scott Deerwester et al., *Indexing by Latent Semantic Analysis*, 41 J. AM. SOC'Y FOR INFO. SCI. 391 (1990).
[24] To some extent, this inclination is understandable. The predictive possibilities of text analytics draw grant funding and industry–university partnerships. For an interesting account at Georgia State University, see Charlotte S. Alexander, *Using Text Analytics to Predict Litigation Outcomes*, *in* LAW AS DATA: COMPUTATION, TEXT, & THE FUTURE OF LEGAL ANALYSIS, *supra* note 3, at 275.
[25] Da, *supra* note 18, at 628.
[26] *Id.* at 628–29.

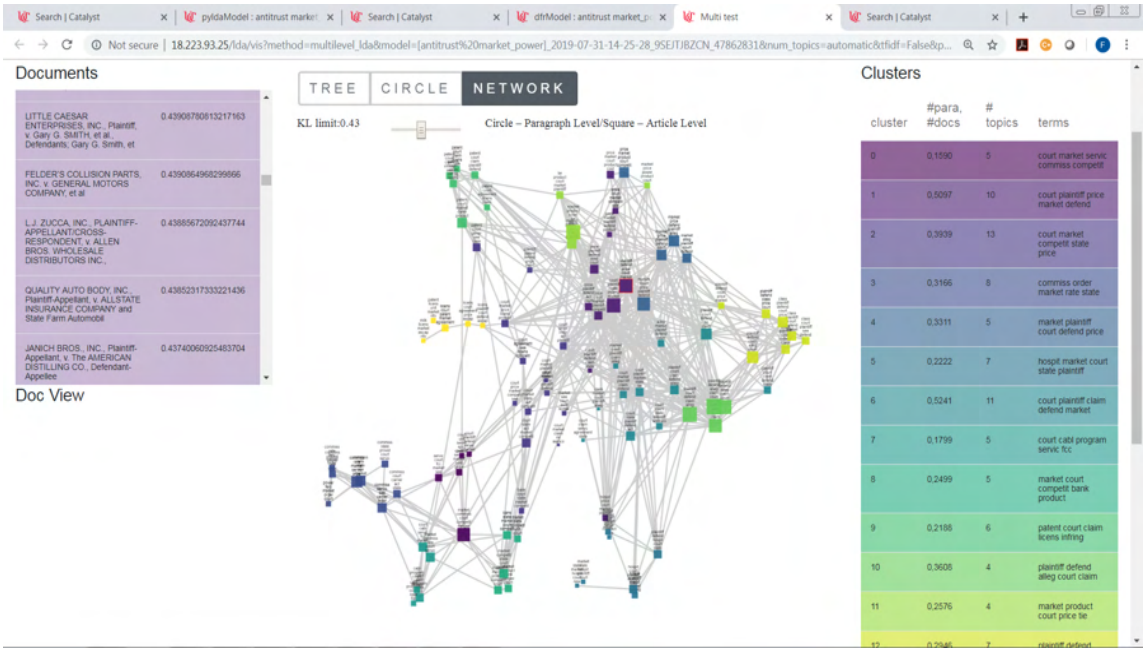

Figure 1: Network View of Market Power Cases in Model of Models

Figure 1 shows a network of antitrust market power cases distributed as topic clusters across space. A topic cluster is an aggregation of multiple topics, where each topic is a collection of terms that are statistically most likely to appear together. The right-hand panel lists each cluster as a distinct shade of color; the clusters are also numbered. In addition, each cluster displays the number of topics that comprise the cluster as well as the top words in the topics. The central graphic depicts the relationship among the clusters. The left-hand panel lists the top "documents," or cases, within a topic as well as the relevant metadata (e.g., case name). It also enables the retrieval of cases.

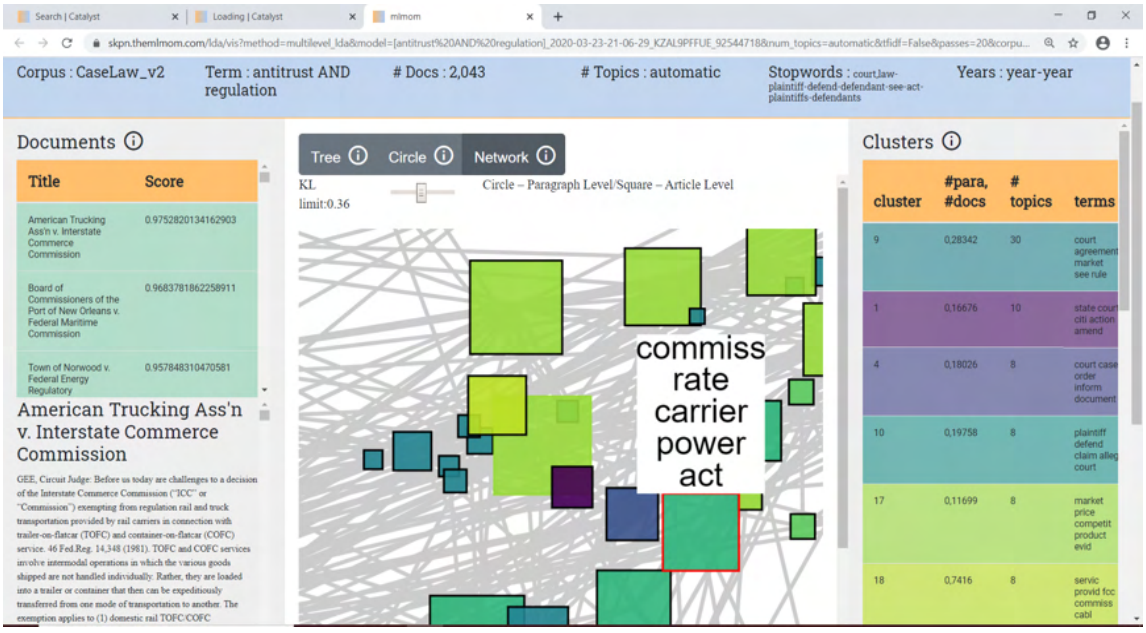

Figure 2: Close-Up View of Antitrust–Regulation Cases with Document Retrieval

Figure 2 demonstrates the case retrieval function on a corpus of antitrust–regulation cases: The highlighted topic cluster in the center encompasses topics with the terms "commiss[ion]," "rate," "carrier," "power," and "act," while the document retrieval feature enables the user to pull up specific cases. Here we have chosen to highlight *American Trucking Associations, Inc. v. I.C.C.*, the top case in the cluster.[27] Note the "top" case means that case that is cross-listed in the most topics.

Our aggregated modeling presents two levels of information: cluster networks show the connections among the topics, while the document retrieval interface shows the specific cases that contribute to each topic. In this way, topics are contextualized at both the macro- and the microscopic levels. The two scales of analysis allow us to see the full complexity of the corpus as a spatial arrangement of how terms are scattered across the cases that comprise the network.

The visualizations employ vector space modeling, with topic clusters strewn across space.

To bolster model stability, a feature that traditional topic modeling sometimes lacks,[28] our algorithms run topic models at least twenty times for each query. As with any empirical project based on copious amounts of data, topic modeling is subject to margins of error, or "wobble."[29] As the next Part details, we run variations of more traditional topic modeling as comparators for each query. In comparison, model aggregation reduces the wobble significantly because the process only picks up the most stable and persistent topics across multiple iterations.

The frequency of iterations also helps to present topics more coherently. Insignificant topic clusters are removed on multiple runs, so the aggregation ensures that visualizations present larger networks that have picked up truly significant term repetitions, rather than statistically aberrant patterns.

III. METHODOLOGY

This Part details how we are using recent technical advances to overcome the hurdles to data extraction and data interpretation. It also reveals how we are checking for model coherence and stability. This Part begins by explaining our data and access procedures, before concluding with our modeling and verification processes.

A – Criticisms

In October 2018, Harvard Law School unveiled its Caselaw Access Project (CAP), which digitized all physically published U.S. case law between 1658 and 2018,

[27] Am. Trucking Ass'n, Inc. v. I.C.C., 656 F.2d 1115 (5th Cir. 1981).
[28] *See, e.g.*, Da, *supra* note 18, at 625.
[29] Margaret E. Roberts et al., *Navigating the Local Modes of big Data: the Case of Topic Models*, *in* COMPUTATIONAL SOCIAL SCIENCE: DISCOVERY AND PREDICTION (R. Michael Alvarez ed. 2016); Antske Fokkens et al., *Offspring from Reproduction Problems: What Replication Failure Teaches Us*, PROCEEDINGS OF THE 51ST ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (2013).

some 40 million pages.[30] One of the projects great advantages is that its APIs feature tools that permit searching through all text in selected cases (as opposed to searches using tags or other metadata). We have created two pools of cases out of the CAP dataset: 36,000 federal cases bearing the word "antitrust" and 305,000 federal cases bearing the word "regulation." We whittled the first pool down to 2,591 cases with the term "market power" (the "Market Power Corpus") and the second pool down to 7,308 with the term "antitrust" (the "Antitrust–Regulation Corpus"). *These corpora* serve as the bases for our Market Power Corpus of 2,591 cases from the "antitrust" pool and our Antitrust–Regulation Corpus of 7,308 cases from the "regulation" pool.

At first glance, these numbers seemed small to us, particularly the count of 36,000 for all federal antitrust cases. The low numbers are partially explained by the fact that the data only goes through 2018 and does not include unpublished decisions.

We verified the case counts in the Market Power Corpus and the Antitrust–Regulation Corpus in several ways. A Westlaw search and subsequent filter for reported federal cases with the terms "antitrust and 'market power'" returned 2,732 cases; for reported federal cases with the terms "antitrust and regulation," this number was 9,775. We also utilized CAP's historical trends interface for verification. CAP has a little over 1.7 million unique federal cases in its corpus, and a search in historical trends reveals that antitrust cases have comprised a low of about 0.1% to a high of almost 4% of all federal cases, with a median roughly short of 2% (or about 34,000 cases).[31] Overall, we have more than a robust sampling for federal antitrust cases.

B – Modeling and Validation

To prepare the corpora, we used the Porter Stemming algorithm to remove suffixes and truncate words down to their roots.[32] The text and metadata of the corpora were then indexed through Elasticsearch, a full-text search and analytics engine.[33]

Utilizing Elasticsearch and the python Gensim package,[34] DSC built a web-based platform that sifts through cases by applying the unsupervised ML algorithm LDA. LDA models are generated on the basis of the distribution of latent topics in a document and the distribution of words in those topics.[35] Each topic is constructed

[30] *About*, CASELAW ACCESS PROJECT, *supra* note 8.

[31] A simple search using CAP's historical trends function reveals that antitrust cases rose to a high of 4% of all federal cases in the 1980s. *See Historical Trends*, CASELAW ACCESS PROJECT, https://case.law/trends/ (search for "us: antitrust"). We also verified CAP's count of federal antitrust cases, which was roughly 32,000.

[32] *See* Martin Porter, *The Porter Stemming Algorithm*, TARTARUS.ORG (2006), https://tartarus.org/martin/PorterStemmer/ (last accessed June 8, 2021).

[33] *Elasticsearch: The Heart of the Free and Open Elastic Stack*, ELASTIC, https://www.elastic.co/products/elasticsearch (last accessed Sept. 14, 2019).

[34] *Gensim 3.8.1*, PYTHON PACKAGE INDEX, https://pypi.org/project/gensim/ (Sept. 26, 2019 data release) (last accessed Oct. 20, 2019).

[35] Blei et al., *supra* note 1.

based on a probability distribution of words.[36] For instance, one topic might feature the term "market" with high probability, whereas its association with another topic will not be as strong. On the platform, users can enter keywords and then retrieve the relevant documents bearing those words (each "document" being an individual decision from our dataset).

Researchers typically employ unsupervised ML techniques to generate insights on reams of unstructured datasets. Compared to supervised ML, however, these techniques can be a bit of a black box: difficult to scrutinize and verify. Accordingly, we combine both qualitative and quantitative methods, bringing together technical expertise in ML and subject matter expertise in antitrust, to ensure not only the replicability of the models but also confidence in their results. This has been DSC's approach in its other collaborations.[37]

For transparency and reproducibility, we have made available the underlying code.[38] The models themselves should also be tested for coherence and stability. At the moment, DSC is working toward the capacity to calculate coherence scores on all models generated from its collaborations.[39] This will allow us to gauge whether the alpha and beta hyperparameters (which determine the assumed concentrations of topics per document and words associated with a topic) are properly tuned to maximize the likelihood that top words in each topic are co-occurring in similar contexts.[40]

As for stability, the platform runs six parallel models from different, random seeds on each corpus; the results are then incorporated in a single visualization.[41] The ensuing model-of-models aggregates topics into clusters, a macroscopic output that accentuates the overlaps in terms and topics that recur across the parallel runs while de-emphasizing the topics with rarer terms. This modification to topic modeling enhances the topics with overlapping vocabularies while improving model interpretability.

Previously, LDA has been criticized for relying too heavily on term–topic probability distributions.[42] In response, we incorporated document-level information as well in the construction of our model-of-models. Combining the

---

[36] For a more detailed explanation, *see* Jason Chuang et al., *Interpretation and Trust: Designing Model-Driven Visualizations for Text Analysis*, CHI '12: PROCEEDINGS OF THE SIGCHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS (2012).

[37] For both an example of another collaboration, as well as a more detailed discussion of DSC's methodology, see Margaret V. Powers-Fletcher et al., *Convergence in Viral Outbreak Research: Using Natural Language Processing to Define Network Bridges in the Bench-Bedside-Population Paradigm*, 3.1 HARV. DATA SCI. REV. (2021), https://hdsr.mitpress.mit.edu/pub/xhht1a9s/release/2.

[38] *See* Ezra Edgerton, *Covid Network Bridges Paper Code,* https://github.com/ucdscenter/Covid_Network_Bridges_code (last accessed June 9, 2021). Note that the modeling source code is the same for our project, even though this page is titled under the name of a different collaboration.

[39] For one example where this has been done, *see id.*

[40] *See* Shaheen Syed & Marco Spruit, *Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation*, 2017 IEEE INTERNATIONAL CONFERENCE ON DATA SCIENCE AND ADVANCED ANALYTICS (DSAA) (2017).

[41] Aggregation is an alternative to running only one model, which is more dependent on the initial Bayesian distribution.

[42] *See* David M. Blei, *Probabilistic Topic Models*, 55 COMMUN. ASSOC. FOR COMPUT. MACH. 77 (2012).

term–topic and document–topic matrices allows us to better assess the models' depiction of inter-topic relationships, by illustrating not only the words but also the cases shared by each topic.

Finally, as a back-end check on our models, we are reading the top cases within all topics. Document-level review allows the domain experts among us to gauge how dependable the unsupervised ML results are—and how closely the results hew to antitrust doctrine.

IV. TOPIC MODELING VISUALIZATIONS

Our platform's models provide visualizations of cases grouped by recurring terms, depicting both the relationships among terms and the relationships among groups of cases. We can create three types of visualizations, all built around topic modeling. The first set of visualizations are generated by our aggregated modeling algorithms. These create "multilevel" or model-of-models visualizations that provide a hierarchical view of topics and topic clusters in three different formats—tree, circle, and network (see Figures 3–6).

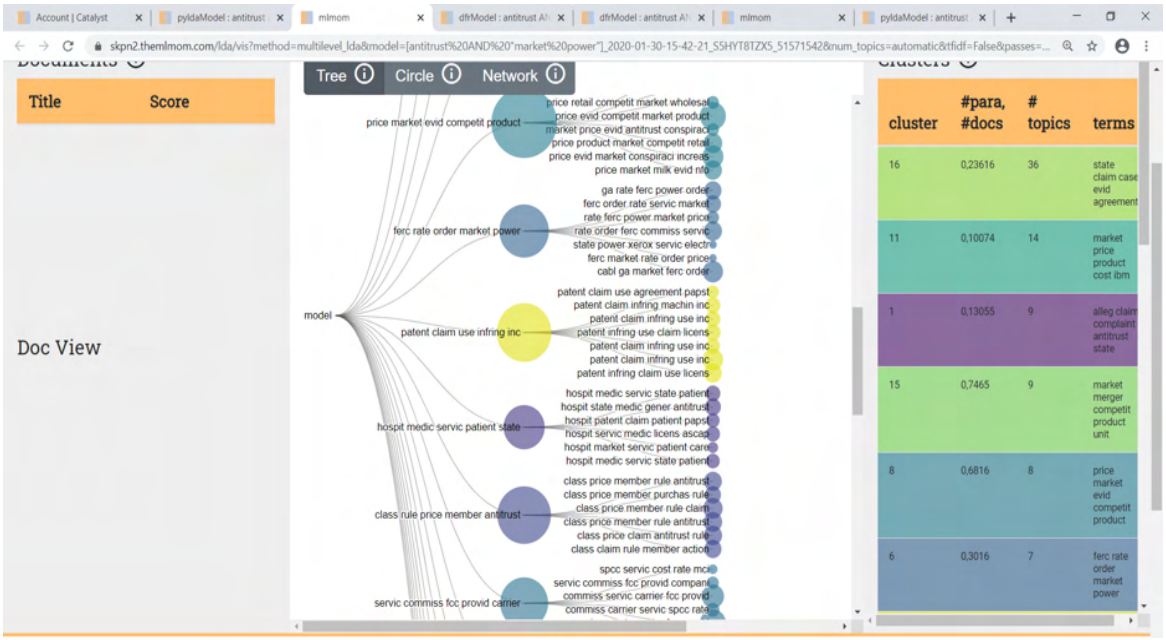

Figure 3: Multilevel Visualization of Market Power Cases in Tree Format

In the tree format of Figure 3, the smaller nodes on the right represent topics (e.g., machine-grouped terms "price," "retail," "competit[ion]," "market," and "wholesal[e]"), while the larger nodes represent clusters of topics (e.g., a cluster with "price," "market," "evid[ence]," "competit[ion]," and "product"). The size of each cluster node or topic node represents the significance of the cluster or topic to the overall corpus. The right-hand bar shows the number of topics within each cluster (thereby functioning as a proxy for the cluster's diversity), and the left-hand bar lists the top cases in each topic.

Circle view presents the same information, but in a format that more clearly conveys the topics where each word appears. Clicking on a specific word pulls up how it is shared across topic clusters. For example, Figure 4 (below) shows the recurrence of the term "market" within all topics. In contrast, network view constructs a spatial representation where each topic comprises a vector in space (see Figure 1 above). It is adapted from the neural network architecture Word2Vec, where each word represents a vector.[43]

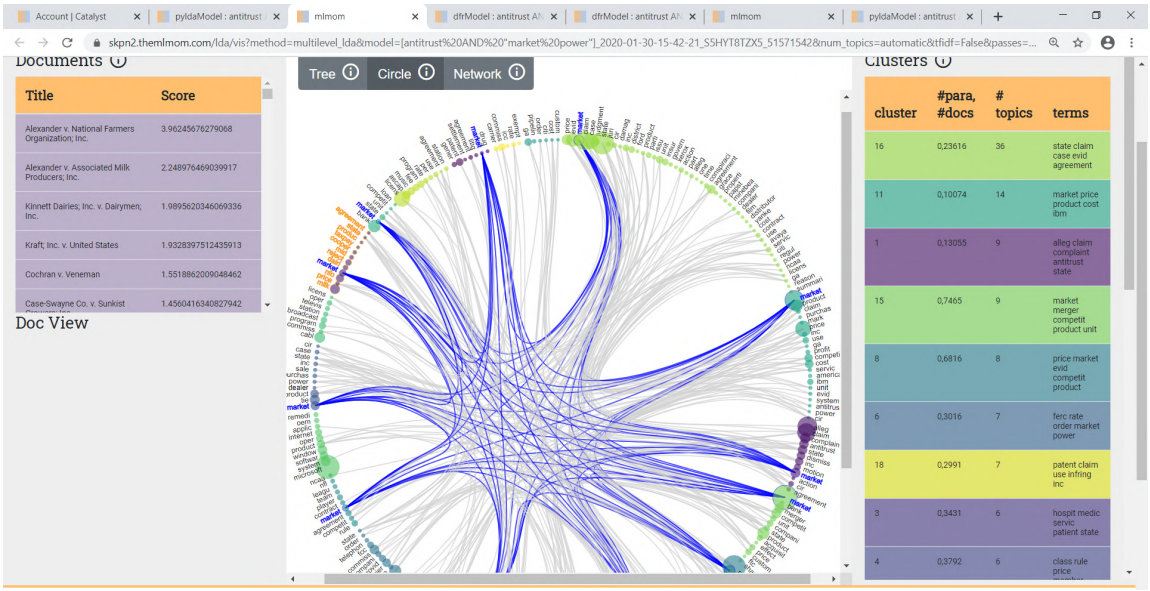

Figure 4: Multilevel Visualization Showing the Recurrence of the Term "Market"

The second set of visualizations, "topic browser," are generated from the DFR framework of Andrew Goldstone, a DH scholar.[44] Topic browser visualizations organize cases into topics, enabling detailed analyses of where (i.e., in what topics) certain terms recur (see Figures 5 and 6).

[43] Thomas Mikolov, *Distributed Representations of Words and Phrases and their Compositionality*, 26 ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS (C.J.C. Burges et al. eds., 2013); Elliott Ash & Daniel L. Chen, *Case Vectors: Spatial Representations of the Law Using Document Embeddings*, *in* LAW AS DATA: COMPUTATION, TEXT, & THE FUTURE OF LEGAL ANALYSIS, *supra* note 3, at 315–7.

[44] *See* Andrew Goldstone, *DfR-Browser: Take a MALLET to Disciplinary History*, https://agoldst.github.io/dfr-browser/ (last accessed Feb. 27, 2020).

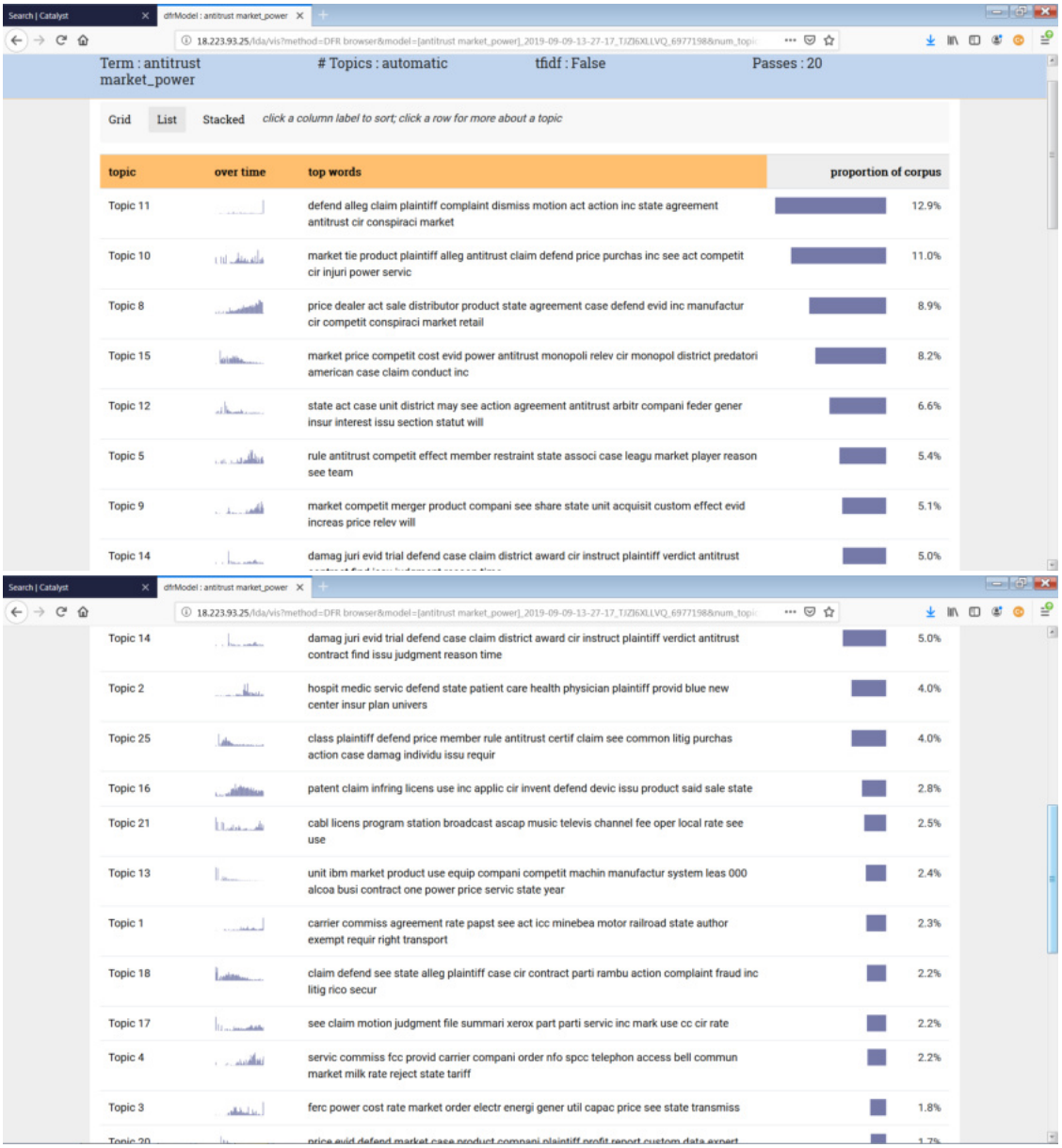

Figure 5: Topic Browser Visualization of Market Power Cases in List Format

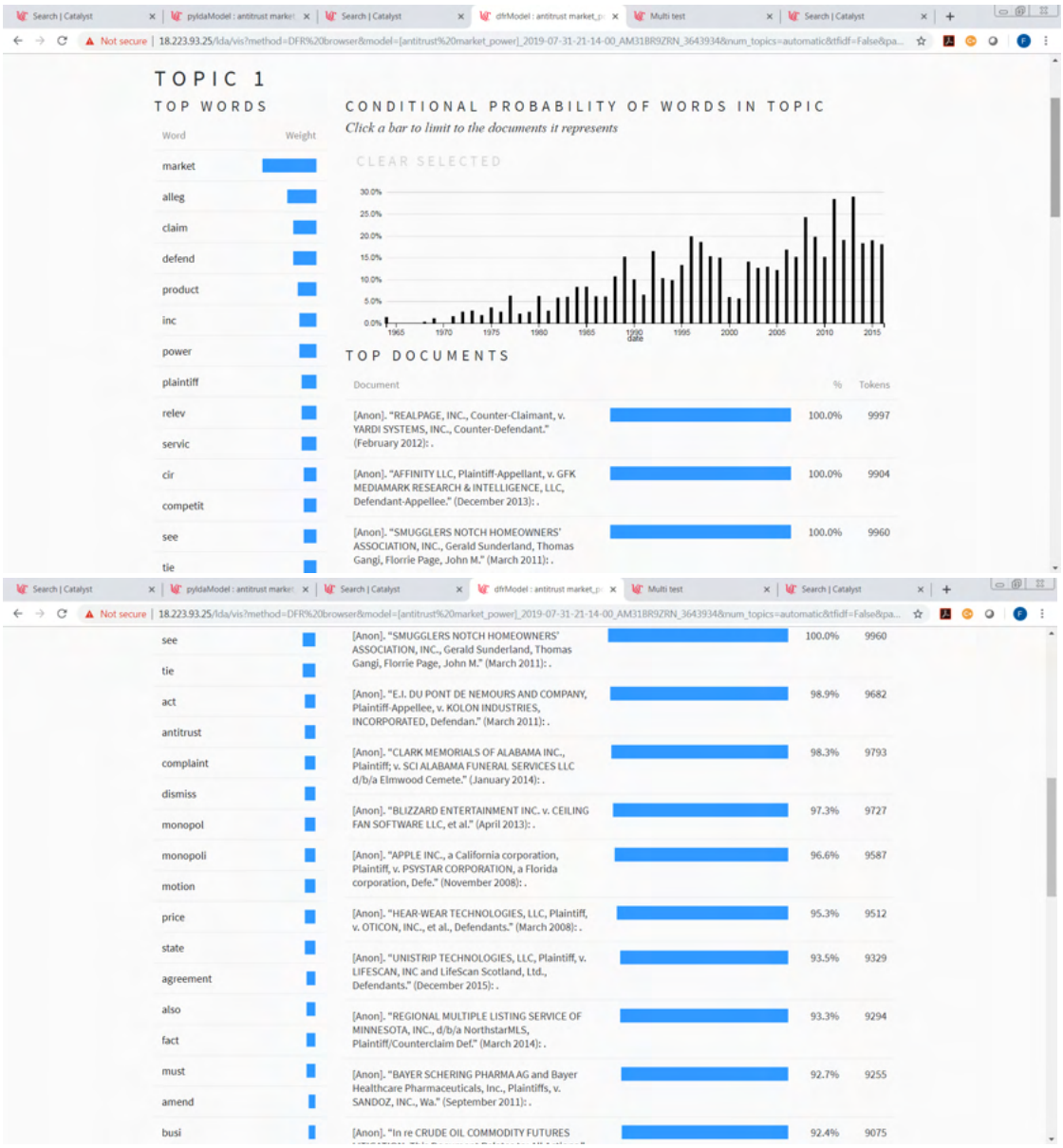

Figure 6: Breakdown of Terms and Cases within a Topic in Topic Browser View

From the overview in Figure 5, the user can browse a specific topic by clicking on it, which brings up the topic's top terms and cases as shown in Figure 6.

Both the overview and single-topic view display histograms on the time periods when certain topics were more prevalent. Clicking on each term pulls up the topics where the term appears.

The third set of visualizations, python-based LDA visualizations ("pyLDAvis"), is built from the framework of the programmer Ben Mabey.[45] PyLDAvis depicts the distance between topics, in a format that most closely resembles the Word2Vec

[45] *See* Ben Mabey, *Welcome to PyLDAvis's Documentation!*, https://pyldavis.readthedocs.io/en/latest/index.html (last accessed Feb. 27, 2020).

architecture (see Figure 7). Word2Vec is a two-layer neural network devised by Google that assigns each term onto a vector in space. The totality of such a graph represents the entire corpus and can have hundreds of vectors, each corresponding to a term, thereby illustrating the proximity and distance among terms.[46] In the pyLDAvis adaptation, the size of each topic bubble represents the weight of that topic. When a topic is highlighted, the platform pulls up the top probable words contained in that topic.[47]

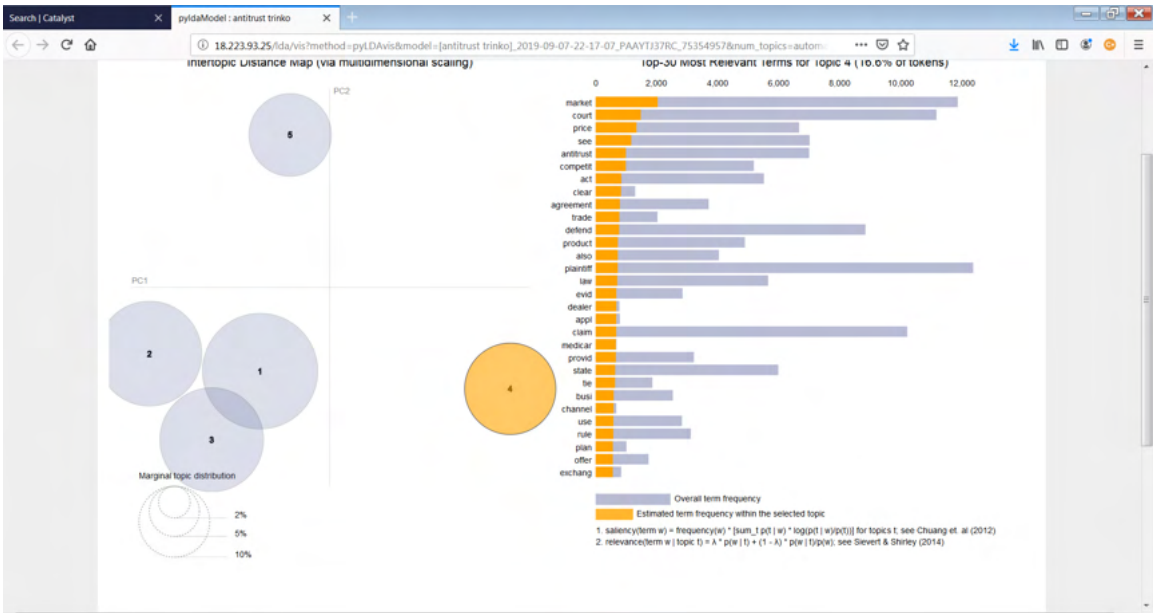

Figure 7: pyLDAvis View of Antitrust Cases Containing "Trinko"

Figure 7 shows how our algorithms have sorted antitrust cases with the word "Trinko" into four topics.[48] In the screen shot, topic 4 is highlighted, bringing up its top terms. With pyLDAvis and the other visualizations, the platform user can set the number of topics manually. Here, the model is comprised of five topic bubbles.

In totality, these three sets of visualizations—multilevel, topic browser, and pyLDAvis—allow for easier interpretation of machine analysis. They add a translational step between topic models and the user by generating visual depictions that are intuitive and easy to grasp without necessarily requiring statistical training. The features we have embedded into the platform also allow users to situate terms and topics in multiple contexts, such as historical (through histograms), constituent cases, constituent terms, meta-topics, and distances among topics and meta-topics.

---

[46] *See* Mikolov, *supra* note 43. For an illustration of Word2Vec, *see* Jay Alammar, *The Illustrated Word2Vec* (Mar. 29, 2019), https://jalammar.github.io/illustrated-word2vec/.

[47] For a mathematical expression of probability, one of the key concepts in this statistical analysis, *see* Carson Sievert & Kenneth E. Shirley, *LDA vis: A Method for Visualizing and Interpreting Topics*, 2014 PROCEEDINGS OF THE WORKSHOP ON INTERACTIVE LANGUAGE LEARNING, VISUALIZATION, AND INTERFACES 63, 66 (Jason Chuang et al. eds. 2014). The probability of any term within a topic is its *relevance* within that topic. Relevance can be expressed as $r(w,k) \mid = \lambda \log(\phi_{kw}) + (1 - \lambda) \log (\phi_{kw} / p_w)$, where λ is the weight of the probability of term w under topic k relative to its lift.

[48] *See* Verizon Commc'ns Inc. v. Law Offs. of Curtis V. Trinko, 540 U.S. 398, (2004). *Trinko* reset the balance between antitrust and regulation while also gutting the essential facilities doctrine.

Two additional points are notable. First, generic words such as "court," "see," "claim," and "plaintiff" are prevalent in the initial results. Although their presence renders the topics more generic, their appearance corroborates the fact that our ML is identifying the top terms. Judicial decisions are replete with these wordsthat algorithms are not trained to filter out.[49] We can refine the results by excluding generic words from the visualizations.[50] Second, these three types of visualizations are different than Word2Vec, which has been the visualization of choice on many legal research projects so far. From a methodological perspective, our project therefore pushes ML in legal scholarship beyond word-level analysis, by building topic and even meta-topic models.

V. CONCLUSION

Topic modeling algorithms can be modified to address the criticisms of its detractors by providing greater context at the micro- and macroscopic levels. We have found that aggregating topic modeling over many iterations helps to eliminate aberrant results while providing contextualization.

Our visualizations show how antitrust caselaw fall into different categories. In the Market Power corpus, clusters span telecommunications, mergers, patent, technology, sports, tying, health care, class action, and civil litigation topics and terms, among others.[51] Meanwhile, the Antitrust–Regulation corpus splinters into Interstate Commerce Commission ("ICC"), immunity, insurance, health care, telecommunications, labor, energy, banking, securities, tax, class action, and civil litigation topics and terms, among others. These results implicate additional trends, such as the decline of ICC cases and a rise of civil litigation topics over time. For researchers delving into a particular doctrine for the first time, the pairing of topic modeling with traditional research tools is particularly exciting because of its ability to show connections across doctrines.

At the same time, however, topic modeling calls into question the search algorithms of commercial databases such as Westlaw and Lexis. Frequently, the top documents for topics and topic clusters are not the top cases that are returned on a Westlaw or Lexis search.[52] We hope these disparities push the operators of these databases to be more transparent in how they define relevance.

There is still much to be done with our platform and visualizations. Looking ahead, we plan to enhance our model validation capabilities by calculating coherence scores and extracting random samplings of documents within our corpora to see whether the topics reflect the same clustering as in the corpora overall. We undertake these validations with the hope of illuminating the black box of our unsupervised ML tools.

---

[49] Our platform has the capacity to exclude these generic terms in the construction of visualizations.

[50] Excluded words are tagged as "stop words." At this point, the platform can only filter out up to nine stop words.

[51] The results are detailed in Felix B. Chang et al., *Modeling the Caselaw Access Project: Lessons for Market Power and the Antitrust–Regulation Balance*, 22 NEV. L. J. __ (forthcoming 2022).

[52] For instance, in the Antitrust–Regulation corpus, in clusters where "immunity" is a top term, the top results are neither *Gordon v. New York Stock Exchange, Inc.*, 422 U.S. 659 (1975), nor *Parker v. Brown*, 317 U.S. 341 (1943). By contrast, these two cases, which cover the conflicts between antitrust and regulation as well as antitrust immunity for state action, are among the top results in Westlaw (using "antitrust and regulation" and a filter for "immunity").