2020

Law, Artificial Intelligence, and Natural Language Processing: A Funny Thing Happened on the Way to My Search Results

Paul D. Callister
University of Missouri-Kansas City School of Law

Law, Artificial Intelligence, and Natural Language Processing: A Funny Thing Happened on the Way to My Search Results*

Paul D. Callister**

Artificial intelligence (AI), including natural language processing, may challenge the legal profession as much, if not more, than the shift from print to digital resources. We may be inevitably moving toward letting AI become our touchstone for authority or, as Robert Berring has articulated, our "cognitive authority."

** Library Director, Leon E. Bloch Law Library, and Professor of Law, University of Missouri-Kansas City School of Law, Kansas City, Missouri. I wish to give special thanks to Pablo Arrendondo of Casetext for encouraging me to look into words as vectors. I wish to thank Susan Nevelow Mart, Dr. Peter Hook, and Kim Kirschenbaum for the assistance in editing and advice.

"Law must be stable and yet it cannot stand still."—Roscoe Pound[1]

"Like gods, these mathematical models were opaque, their workings invisible to all but the highest priests in their domain: mathematicians and computer scientists."—Cathy O'Neil[2]

"For most of the twentieth century, the legal world had agreed to confer cognitive authority on a small set of resources. By 'cognitive authority' I mean the act by which one confers trust upon a source."—Robert C. Berring[3]

Introduction

¶1 Artificial intelligence (AI),[4] including natural language processing,[5] may challenge this profession, including the larger profession of law practice, as much as, if not more than, the shift from ownership of print resources to licensed digital resources. We may be inevitably moving toward AI becoming our touchstone for

---

1. Roscoe Pound, Interpretations of Legal History 1 (1923).

2. Cathy O'Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy 3 (2016).

3. Robert C. Berring, *Legal Information and the Search for Cognitive Authority*, 88 Calif. L. Rev. 1673, 1676 (2000).

4. "Artificial intelligence is, broadly, a set of computational technologies that aim to sense, learn and reason about their physical or virtual environment and take action based on that." Elliot Jones, Nicolina Kalantery & Ben Glover, DEMOS, Research 4.0 Interim Report 5 (Oct. 2019), https://demos.co.uk/wp-content/uploads/2019/10/Jisc-OCT-2019-2.pdf (last visited Aug. 4, 2020). Some definitions have stressed the human element or equivalence of AI; "an artificial intelligence system [is] a machine behaving in ways thought to be intelligent if a human were so behaving." Casandra M. Laskowski, *AI Defined: Core Concepts Necessary for the Savvy Law Librarian*, *in* Law Librarianship in the Age of AI 3 (Ellyssa Kroski ed., 2020). The same author notes that AI is very much a moving target: "Intelligence is whatever machines haven't done yet." *Id.* at 2 (citing Larry Tesler, *Adage and Coinages*, Curriculum Vitae, http://www.nomodes.com/Larry_Tesler_Consulting/Adages_and_Coinages.html (last visited Aug. 4, 2020)).

5. Natural language processing may be defined as a computer's "way of processing language as actually used rather than set commands." Jones, Kalantery & Glover, *supra* note 4, at 5. "Natural Language Processing (NLP) is the processing and analysis of unstructured language data, essentially enabling computers to understand human language." *Id.* at 8. For example, a lot of scientific information is structured and can be processed easily, but humanities and social science information (including law) is unstructured. *Id.*

authority or, as Robert Berring has articulated, our "cognitive authority,"[6] but we are not there yet. Outside of the field of law, it is easy to see how this shift has already occurred. We can ask Google and Amazon Alexa about all sorts of things and get cogent answers. For example, "Who won the last Chiefs game?" More remarkable than the answers is that we trust their accuracy. These devices and the software that supports them have become part of our cognitive authority. Within law, the relationship between search engines and cognitive authority is more complex.

¶2 To serve the topic, this article will proceed as follows: first, beginning with the introductory quotations, the issues and themes will be introduced along with practical issues surrounding student research habits and problems with inconsistent research results. Next, to understand how effective natural language processing[7] (a subset of AI) is in current legal research, I will go about building a model of a legal information retrieval system that incorporates natural language processing. I have to build my own because we do not know very much about how the proprietary systems of Westlaw, Lexis, Bloomberg, Fastcase, Ravel, and Casetext work.[8] However, there are descriptions in information science literature and on the Internet of how systems with natural language processing actually work or could work. Then, I will compare such systems with the features and search results produced by the major vendors to illustrate the probable use of natural language processing, similar to the models. Next, the use of word prediction or type-ahead techniques in the major research services is also worth studying while considering natural language processing—particularly, how such techniques can be used to bring secondary resources to the forefront of a search. Finally, I will explore how the knowledge gained may help us to better instruct law students and attorneys in the use of the major legal information retrieval systems.

¶3 My conclusion is that the adeptness of natural language processing is uneven among the various vendors and that what we receive in search results from such systems varies widely depending on a host of unknown variables. Natural language processing has introduced new uncertainty to the law. We are a long way from idyllic AI systems that understand, let alone search, legal texts in a stable and consistent

---

6. Robert Berring laid out his idea of cognitive authority in Berring, *supra* note 3, at 1676 ("The cornerstone tools of legal information have been established as unquestioned oracles. They appeared too obvious to examine.").

7. For this article, I am using *natural language processing* as a single example of AI, and will generally refer to it instead of AI. Natural language processing is not the same thing as natural language searching, although the latter tends to incorporate the former. Natural language processing has many other uses than information retrieval incorporating relevancy-based feedback. Its ultimate aim is to process language as humans do and respond in kind. *See* Jones, Kalantery & Glover, *supra* note 4.

8. While I was writing this article, an important article appeared in *AALL Spectrum*, coauthored by several important data scientists from various vendors and librarian Susan Nevelow Mart, that discusses how the major vendors employ search algorithms. Susan Nevelow Mart et al., *Inside the Black Box of Search Algorithms*, AALL Spectrum, Nov./Dec. 2019, at 11. However, the discussion is in general terms about search features, and does not include the level of detail, often mathematical, in this article—incomplete though it may be. The *Spectrum* article mentions TF-IDF (also discussed *infra*, ¶¶ 27–34), automatic query expansion, term proximity, result ranking, term proximity to key words, the importance of context to relevancy, source authoritativeness, the aggregation of user search history, support for search commands, differentiation of search engine by type of document, machine learning algorithms, probabilistic approaches, etc. *See id.* at 11–15. Even with this laundry list of tools, the details of just how current information retrieval features work and are applied are left vague, which is in part due to the short nature of *Spectrum* articles, but may also be a result of the proprietary nature of search algorithms.

way. The use of secondary authority, such as the great legal treatises, is at risk. This may make establishment of cognitive authority more difficult. On the other hand, bestowal of cognitive authority on our new tools may be an act of faith and may even be necessary for us to function as a profession, driving us toward AI systems in the future. It may also be inevitable that only a small cadre of the profession may take advantage of what will increasingly become extremely powerful, but costly, systems for providing legal research assistance.

Issues and Themes

¶4 The objective of this section is to introduce both the theoretical issues surrounding AI and cognitive authority and the practical concerns arising from our current (and future) uses of natural language processing.

Introductory Quotations: Elaborations

¶5 The three quotations that began this article suggest three issues we face as law librarians (and which the larger law profession faces too). The first quotation from legal scholar and educator Roscoe Pound reminds us of the need for tension between stability and adaptability in the law.[9] Yet, how will the law remain stable when, as demonstrated in this article,[10] our tools for accessing the law provide such disparate results? The second quotation goes to the issue of opaqueness in natural language processing and the algorithms that are producing our search results. It was made by a prominent data scientist who wrote a book, *Weapons of Math Destruction*, warning of the devastating effects of unintended consequences of algorithms in everything from lending markets to education. We are unlikely to gain access to how prominent vendors tune their algorithms. They remain opaque, yet authoritative. "Authority is increasingly expressed algorithmically."[11] As in other domains, are there harmful and unforeseeable unintended consequences from natural language processing of legal authority and commentary? There are, primarily because of inconsistency among services and lack of access of most attorneys to all of the search tools providing such differing results.[12]

¶6 The third quote deals with "cognitive authority," which is a conferral of trust on a historically small set of particular legal research resources.[13] This bestowal of authority could be on a primary source, like the *United States Code Annotated*, even though it is not the official version of the *United States Code*. The legal community

---

9. Wikipedia, *Roscoe Pound*, https://en.wikipedia.org/wiki/Roscoe_Pound (last visited Aug. 4, 2020).

10. *See infra* ¶¶ 53–74.

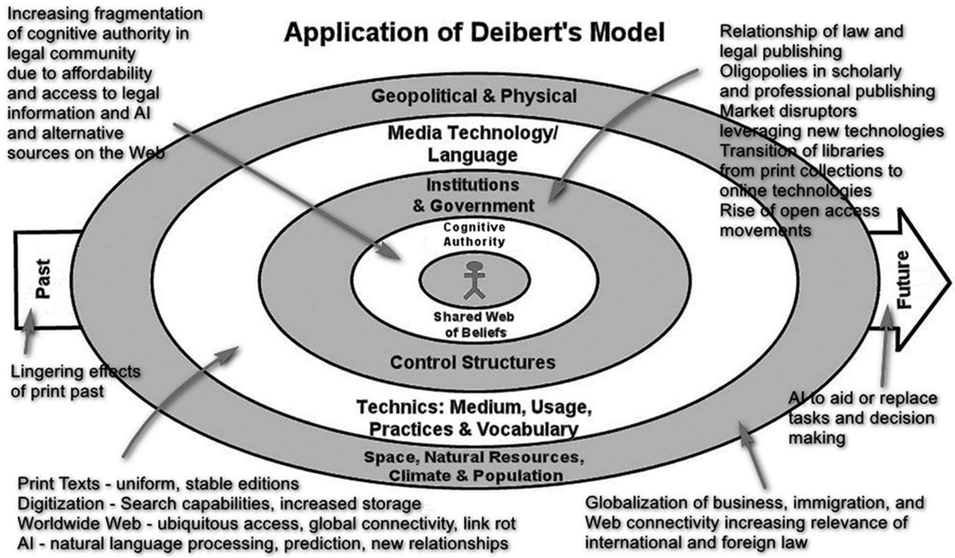
11. Frank Pasquale, The Black Box Society: The Secret Algorithms that Control Money and Information 8 (2015).

12. Although I was able to find surveys on what attorneys use, I was unable to find information about services to which attorneys typically do not have access. *See* Robert Ambrogi, *Survey Finds Virtual Dead Heat in Lawyers' Use of Westlaw, LexisNexis and Fastcase*, LS LawSites, Mar. 13, 2017, https://www.lawsitesblog.com/2017/03/survey-finds-virtual-dead-heat-lawyers-use-westlaw-lexisnexis-fastcase.html (last visited Aug. 4, 2020). Admittedly, my statement about lack of access to the full range of legal research services is based on anecdotal evidence. This would be worth studying in a future article.

13. *See* Berring, *supra* notes 3 and 6 and accompanying text.

Figure 1

Application of Ronald Deibert's Model of Holistic Ecological Media Theory to Law



can entrust cognitive authority to commentary like *Nimmer on Copyright*, *Moore's Federal Practice: Civil*, or BNA Tax Portfolios.

¶7 The concept of *cognitive authority* shares affinity with other core concepts, which in past publications I have referred to as *legal epistemology* or the shared *web of beliefs* common to the legal profession.[14] Ronald Deibert, from whom I adapted the concept of *legal epistemology* for the profession, uses the concept of *social epistemology* for the larger society, which is the "web-of-beliefs into which people are acculturated and through which they perceive the world around them."[15] That *web-of-beliefs* includes what constitutes *cognitive authority*.

¶8 There is a relationship between cognitive authority and legal institutions, technologies, language, and even the geopolitical environment. They affect one another as part of a holistic model of the legal information ecosphere, as illustrated by figure 1, adapted from Deibert's ecological holistic model of media theory.[16]

---

14. *See* Paul Douglas Callister, *Law's Box: Law Jurisprudence and the Information Ecosphere*, 74 UMKC L. Rev. 263, 267 (2005) [hereinafter *Law's Box*]; Paul Douglas Callister, *The Book as Authoritative Sign in Seventeenth-Century England: A Review Through the Lens of Holistic Media Theory*, *in* Law Culture and Visual Studies 49, 51–52 (Anne Wagner & Richard K. Sherwin eds., 2012) [hereinafter *Book as Authoritative Sign*].

15. *See Book as Authoritative Sign*, *supra* note 14, at 51 (citing Ronald J. Deibert, Parchment, Printing, and Hypermedia: Communication in World Order Transformation 94 (1997)).

16. *See* Deibert, *supra* note 15 at 38. For a detailed interpretation of each ring and the arrow connecting past and future based on historical criteria, *see Law's Box*, *supra* note 14, at 267–72. The model is referred to as *holistic* and *ecological* because Deibert wanted to contrast his formulation with prior versions of media theory that were *determined* by changes in technology—most notably from Harold Adam Innis and Marshall McLuhan. *See id.* at 265.

¶9 Note in figure 1 that technology and language are tied together. We can see that quite clearly with natural language processing, which is about the attempts to deal computationally with unstructured language, as is found in legal discourse.[17] Language is a quintessential technology. Other examples of the relationship of language, technology, and institutions include the ancient Egyptians, who because of silent determinants of meaning within hieroglyphic and hieratic writing required a scribal class to interpret rather than just phonetically vocalize law;[18] the Celtic and Icelandic bards who used meter and devices like stating the law in triads, to preserve and communicate the law;[19] the gloss of medieval codices and manuscripts, which sometimes made its way into the law itself;[20] the use of clay "wrappers" or seals to authenticate legal documents in Mesopotamia;[21] the demotic (meaning *common* or *democratic*) alphabet of the classical Greeks used to communicate law openly on stone stele to whomever could read;[22] and the use of pinpoint citation in the era of the printing press to stabilize and create a web of legal authority.[23] Now the issue is potentially the natural language processing of legal texts to enable machine understanding and participation in legal dialogue. It's not just about searching the law; it's about understanding it.[24] Natural language processing may become society's new mediating scribe of the law—able to understand relationships of legal texts invisible to human intelligence without its aid. Such a technology would quite naturally be endowed with cognitive authority by the legal community.

¶10 Returning to figure 1, the arrow on the chart is significant. Its tail suggests that the past bestowals of cognitive authority cannot be ignored. There is not a

---

> New technologies of communication do not *generate* specific social forces and/or ideas, as technological determinists would have it. Rather, they *facilitate* and *constrain* the extant social forces and ideas of a society. The hypothesized process can be likened to the interaction between species and a changing natural environment. New media environments favor certain social forces and ideas by means of a functional bias toward some and not others, much the same as natural environments determine which species prosper by "selecting" for certain physical characteristics. In other words, social forces and ideas survive differentially according to their "fitness" or match with the new media environment—a process that is both open-ended and contingent.

Deibert, *supra* note 15, at 36.

17. *See* Jones, Kalantery & Glover, *supra* note 4. In contrast to the social science (and law), the language of the hard sciences is considered to be structured. *Id.*

18. *See Law's Box*, *supra* note 14, at 297–99.

19. *See id.* at 311–19. The medieval Welsh used triads. *See generally* The Legal Triads of Medieval Wales (Sara Elin Roberts ed., 2007).

20. *See Law's Box*, *supra* note 14, at 308.

21. *See id.* at 285–86.

22. *See id.* at 278. The Greek alphabet, unlike Egyptian hieroglyphics and hieratic, was purely phonetic, without silent determinatives that influenced meaning, requiring a scribe (in the case of Egyptian, but not Greek) to interpret the meaning. *Compare id.* at 278 with *id.* at 296–300. The role of language in liberating the common people from scribal and bureaucratic classes has been noted: "[T]he eminent British legal historian and diplomatist M.T. Clanchy entertained . . . criticisms in his monumental work on the evolution of English legal documents: '[I]t is language itself which forms mentalities, not literacy . . . . Morally and psychologically, depending on the circumstances, literacy may liberate or it may confine.'" *Id.* at 300 (citing M.T. Clanchy, From Memory to Written Record: England 1066–1307, at 9 (2d ed. 1993)).

23. *See Book as Authoritative Sign*, supra note 14, at 66–69.

24. "'Legal research' is not merely a search for information; it is primarily a struggle for understanding." Michael J. Lynch, *An Impossible Task but Everybody Has to Do It—Teaching Legal Research in Law Schools*, 89 Law Libr. J. 415, 415 (1997). *See also* Nevelow Mart et al., *supra* note 8, at 15 (describing current application of Westlaw Edge's natural language processing to "understand the meaning of a query").

complete break between successive periods of cognitive authority. Each period from the past influences what is accepted in the future. For instance, Lord Coke, in writing his famous treatise, or *Institute*, or *Commentary on Littleton*,[25] used a format suggestive of the Justinian gloss that came from the manuscript era prior to the printing press (even though Lord Coke's treatise was published on such presses).[26] In more recent times, West and Lexis, including their migrations to electronic versions, have earned what I would argue is cognitive authority for access to primary authority of the law—cases, codes, and regulations—and even secondary authority for the great treatises cited in court. They are accepted almost without reservation despite often not being the official versions, in the case of primary law. However, even online services have to acknowledge the past by preserving page numbers from print sources, and in some cases making available PDF images of the print versions. The point of all these examples and the chart in figure 1 is that technologies such as language affect institutions—including the bar, courts, and legal publishers—which in turn affect (and are affected by) cognitive authority. Natural language processing will have to ground itself in the forms and functions of cognitive authority of the past—perhaps such as giving cognizance to most-cited cases, adhering to jurisdictions, performing citation analysis, building on West's Topic and Key Number System, emphasizing cases annotated in *American Law Reports*, or any number of a hundred factors that make up the current terrain of the legal information environment.

¶11 Looking at the point of the arrow in figure 1, which represents the potential of the future, we live in a remarkable age. The whole purpose of natural language processing is to give machines more "understanding" of (or at least effectiveness with) human speech as it is actually delivered. Efforts to employ natural language processing include application to legal texts. Predictably, there will be an effect upon the legal profession's cognitive authority, while at the same time the legal profession's traditions related to cognitive authority will affect how natural language processing is employed and even accepted. My prediction is that if natural language processing becomes increasingly adept at answering legal questions with precision, reliance upon the great treatises and secondary sources may falter—the relationship between primary law and secondary authorities may be usurped by a new relationship based on natural language processing.[27] However, as this article will hopefully

---

25. Edward Coke, The First Part of the Institutes of the Lawes of England: Or a Commentary upon Littleton (3d ed. corrected, 1633).

26. *See Book as Authoritative Sign*, *supra* note 14, at 66–67 (in particular, note replication of Lord Coke's first *Institutes*, based on Littleton, in the manner of glossed texts, in fig. 3.7).

27. My predictions about the replacement of secondary authority (authored by human experts) by natural language processing are not out of line with forecasts about the relationship of work done by artificial intelligence instead of humans. Just to illustrate, the World Economic Forum estimates that "the average percentage of tasks carried out by machines vs. humans will change from 29% vs. 71% in 2018 to 42% vs. 58% [by 2022]." Mary Lee Kennedy, *What Do Artificial Intelligence (AI) and Ethics of AI Mean in the Context of Research Libraries?*, *in* Association of Research Libraries, *Ethics of Artificial Intelligence*, Research Libr. Issues no. 299, 5 (2019), https://publications.arl.org/rli299/1 (last visited Aug. 4, 2020). Supporting this conclusion: "In a 2018 survey of AI researchers, 50% forecasted that high-level machine intelligence (HLMI) would be achieved within 45 years [and a 10% chance within 9 years]. HLMI is achieved when machines can accomplish every task better and more cheaply than human workers." Jones, Kalantery & Glover, *supra* note 4, at 17 (citing Katja Grace et al., When Will AI Exceed Human Performance? Evidence from AI Experts (2017, revised May 2018), https://arxiv.org/pdf/1705.08807 (last visited Aug. 4, 2020)). In the field of law, Richard

reveal, the current state of technology, while trending away from secondary sources, is still a ways from replacing them. This seems particularly true in that current legal natural language processing has little facility with legal syntax, but depends instead upon general word proximity and co-occurrence. Of course, as shall be shown, the current state of things is more nuanced and complex than simply determining word proximities and co-occurrences.

More Immediate and Practical Concerns

¶12 Turning to practical concerns, I have long maintained that legal researchers should vary search techniques according to the types of problems they face.[28] But I must confront the reality made evident by a recent survey of law students at my school. In discussing how they start their research, almost half the students agreed with this statement: "I just start typing in the search bar on Lexis, Westlaw, or Bloomberg to see what comes up."[29] This held true even though "starting with Google" was also an option on the survey (12.87% of students selected). Starting with legal commentary (something with an index) to get background on the law is far less likely (11.88%) than using a single-search box. Students can filter their results on Lexis, Westlaw, and Bloomberg "post-search" to find secondary sources, but do they bother when primary materials are displayed first and may seem relevant?

¶13 We have entered the age of the one-size (or one search box) fits-all legal inquiry, and dare I suggest that what will follow (especially, in light of vendor advertising) is that one day we will start to behave as if legal search algorithms are intelligent and capable of understanding our queries.[30] This will only get truer with

---

Susskind revised his best-selling book, *Tomorrow's Lawyers: An Introduction to Your Future*, to assert, "as our machines become increasingly capable, they will steadily eat into lawyers' jobs. The best and the brightest human professionals will last the longest . . . ." Richard Susskind, Tomorrow's Lawyers: An Introduction to Your Future 188 (2d ed. 2017). Those jobs may ultimately include the recognized attorney/editors and authors of the major commentaries that are part of the cognitive authority of the law.

28. *See* Paul D. Callister, Field Guide to Legal Research 17–66 (2019); Paul D. Callister, *Time to Blossom: An Inquiry Into Bloom's Taxonomy As a Hierarchy and Means for Teaching Legal Research Skills*, 102 Law Libr. J. 191, 204, tbl. 4, 2010 Law Libr. J. 12, tbl. 4.; Paul D. Callister, *Thinking Like a Research Expert: Schemata for Teaching Complex Problem-Solving Skills*, 28 Legal Reference Servs. Q. 31, 36–38 (2009); Paul Douglas Callister, *Beyond Training: Law Librarianship's Quest for the Pedagogy of Legal Research Education*, 95 Law Libr. J. 7, 37–38, tbl. 4 (2003).

29. The question in the 2019 survey of 101 UMKC law students was:

> Which of the following statements are true about starting your legal research online? Select only one.
>
> ☐ I just start typing in the search bar on Lexis, Westlaw, or Bloomberg to see what comes up. 48.51%
> ☐ I tend to dive into a search of primary law (cases, statutes, etc.) as my first effort. 20.9%
> ☐ I like to start with legal commentary like *Missouri Practice*, *American Law Reports*, or treatises to get background on a topic before beginning my research. 11.88%
> ☐ I have looked up law review and journals online for my topic at the start of my research. 5.94%
> ☐ I start with Google to find information on point. 12.87%

For the survey question, see Question 8 of the Default Report 360 Legal Information Environment Survey (Mar. 7, 2019), https://umkclaw.link/360-survey (last visited Aug. 4, 2020).

30. Westlaw Edge has released WestSearch Plus, which will answer a variety of legal questions with "type-ahead functionality" and search suggestions. Patrick Yatchak, *Thomson Reuters, Answer Legal Questions Faster Than Ever with West Search Plus*, https://legal.thomsonreuters.com/en

each new search engine released by the vendors. *AALL Spectrum* reports, "Recently, Westlaw Edge extended those capabilities through a set of proprietary natural language algorithms that aim to 'understand the meaning of the query' and, when appropriate, provide answer-like results."[31] It is entirely possible that legal information systems will progress from search boxes to research assistants to perhaps (as others have predicted) even providing legal services.[32] The paramount question moves from "will these systems provide stable access to the law?" to "will there even be a stable system of American law in such a world?" Relatedly, will the algorithms or AI driving such systems themselves become the profession's basis for cognitive authority—a concept that asks whether they will be the trusted sources of American law, perhaps in part by subjugating the treatise?

¶14 Historically, in a book environment, legal research (and law) was stable by design.[33] We all had the same digests and indexes from West and citators from Shepard's.[34] But for some time the very fabric of legal research inquiry (at least its initial steps) has been torn by the unpredictability of diverse search algorithms and methods for natural language processing. This concern occurs at a time when,

---

/insights/articles/answer-legal-questions-faster-with-westsearch-plus (last visited Aug. 4, 2020). Thomson Reuters states that with "artificial intelligence, combined with exclusive editorial enhancements," it can deliver answers to thousands of questions. No doubt editors play a role in selecting question types. Thomson Reuters, *WestSearch Plus*, https://legal.thomsonreuters.com/en/products/westlaw/edge/westsearch-plus (last visited Aug. 4, 2020).

Lexis Answers is more circumspect and will answer questions based on a definition, elements, standard of review, burden of proof, a legal doctrine, and statute of limitations. LexisNexis, *You Ask a Question . . . Lexis Answers Understands It*, https://www.lexisnexis.com/pdf/lexis-advance/lexis-answers.pdf (last visited Aug. 4, 2020). Vendor advertising touts the AI applied to its products, even if the application is still limited.

31. Nevelow Mart et al., *supra* note 8, at 15.

32. Jamie Baker has written about AI technologies replacing "lower-level legal professionals." Jamie J. Baker, *2018: A Legal Research Odyssey: Artificial Intelligence as Disrupter,* 110 Law Libr. J. 5, 26, 2018 Law Libr. J. 1, ¶ 72.

> For example a software application could first conduct a fact-gathering intake session to formulate the questions that need to be answered. Then using algorithms that employ natural language understanding, the algorithm would analyze the user inputs to understand the question. The algorithms would generate the appropriate case law, statutes, and regulations to analyze and compile a memo that succinctly describes the current law.

Id. Interestingly, Baker's quote occurs in the context of a discussion of legal ethics, and her claim that ABA Model Rule 5.3, pertaining to supervision of nonlawyer assistants, should apply to assistance from AI technology. *Id.*

33. *See* Robert C. Berring, *Legal Research and the World of Thinkable Thoughts*, 2 J. App. Prac. & Process 305, 305 (2000) ("The world of established sources and sets of law books that had been so stable as to seem inevitable suddenly has vanished. The familiar sets of printed case reporters, citators, and secondary sources that were the core of legal research are being minimized before our eyes.").

34. There are problems with such a system. While not identical, West's and Lexis's "classification system[s] reflect a nineteenth-century worldview." Susan Nevelow Mart, *The Algorithm as Human Artifact: Implications for Legal [Re]Search*, 109 Law Libr. J. 387, 418, 2017 Law Libr. J. 20, ¶ 55.

> The second kind of viewpoint discrimination is one we don't think about that much, and that is the nineteenth-century worldview of the legal system explicitly embedded in Westlaw's Key Numbers and in Lexis Advance's Topics. These classification systems, while not identical, follow a pattern that is familiar to anyone who has taken contracts in law school. It is firmly based in the Langdellian view of the world, where the subject matter is broken down into similar patterns of essentials for formation, interpretation, performance, defenses, and breach. This view is a form of filtering, for better or worse, and the newer legal research databases [Fastcase, Casetext, Ravel, and Google Scholar] may be freer of whatever limitations that worldview imposes.

Id. at 419, ¶ 56. We might wonder why feminist jurisprudence or critical race theory does not appear in such a system.

according to a recent study of 325 decisions in the federal courts of appeals (citing 7552 cases), only 16% of the cases cited in appellate briefs make it into the courts' opinions.[35] The single-search box raises several concerns in addition to the concurrent weakness of case citation in appellate briefs.

¶15 First, Susan Nevelow Mart has demonstrated in a seminal article, *The Algorithm as Human Artifact: Implications for Legal [Re]Search*, that the different online research services (Westlaw, Lexis Advance, Fastcase, Google Scholar, Ravel, and Casetext) produce significantly different results when researching case law with natural language techniques.[36] Roughly 40% of the search results in each studied database were unique.[37] For the top 10 results, when combining unique and relevant results, statistics varied greatly, with new services clustering at lower numbers, and Lexis and Westlaw having higher rates (Casetext 8.2%, Fastcase 13.1%, Google Scholar 14.6%, Lexis Advance 19.7%, Ravel 11.3%, and Westlaw 33.1%).[38] "These algorithmic variations in worldview lead to substantial variations in the unique and relevant results each database provides. The knowledge of this variability expands the opportunities for researchers to find relevant cases that can play 'some cognitive role in the structuring of a legal argument.'"[39] The point is how can the law be stable, a fundamental axiom,[40] when our research tools are providing "substantial variations in the unique and relevant results?"[41] This is a theme that I shall demonstrate also proves true in some of the sample natural language searches done on case law for this article.

¶16 While Nevelow Mart sees the diversity in unique results as an opportunity, few attorneys have access or time to use all of the tools employed in her study (nor can clients afford it). When I asked Nevelow Mart about this issue, her answer is that the solution to having limited access to the different services is to use reiterative searching to discover all of the relevant material.[42] Whether such methods are

---

35. Ken Bennardo & Alexa Z. Chew, *Citation Stickiness*, 20 J. App. Prac. & Process 61, 82, 74–75 (2019). "In our 325-case data set, the parties cited 23,479 cases. Of those, only 16% were later cited by the courts in their opinions—or to use our nomenclature, only 16% of the cases cited in the briefs were sticky." *Id.* at 84. If both parties cited the case, 38% of such opinions were cited by the appellate court. *Id.* The same study found that 49% of cases cited by courts had been cited by at least one party in their brief (only 21% were cited by both parties). *Id.* Samples were taken from each of the Circuit Courts of Appeal. *Id.* at 78. It would be interesting to study whether these statistics hold true in earlier periods, prior to digital search engines.

36. *See* Nevelow Mart, *supra* note 34, at 397, ¶ 16; 409, ¶ 36. Technically, Nevelow Mart conducted "key word" searches, which she distinguishes from natural language searches, the latter of which employ "grammatical structures models." *Id.* at 397, ¶ 16. See *infra* ¶¶ 51–52 for discussion of grammatical techniques. I do not draw such a distinction between key word and natural language searching because I am writing about natural language processing, which includes both key word and "grammatical structures" searches.

37. *See id.* at 413, chart 1.

38. *See id.* at 415, chart 3. The relevancy of top 10 cases also varied with Westlaw at 67%, Lexis Advance at 57%, and the rest clustered around 40%. *See id.* at 414, chart 2.

39. *Id.* at 420, ¶ 57 (citing Stuart A. Sutton, *The Role of Attorney Mental Models of Law in Case Relevance Determinations: An Exploratory Analysis*, 45 J. Am. Soc'y Info. Sci. 186, 187 (1994)). For what Nevelow Mart means by "worldview," see *supra* note 34.

40. *See, e.g.*, Pound, *supra* note 1.

41. *See* Nevelow Mart, *supra* note 34, at 420, ¶ 57. Roughly 40% of search results across the different services are "unique to one database." *Id.* at 390, ¶ 5.

42. Phone conversation, Oct. 24, 2019. Techniques for reiterative searching are discussed *infra* ¶¶ 96–101.

a complete solution to the problem of diverse results among the major and minor search engines is worthy of further study. In particular, we need to test whether reiterative searching on the staples of Lexis and Westlaw and newcomers Fastcase (supplied with some state bar memberships), Casetext, Ravel, Google Scholar, and even Bloomberg will find the same seminal cases that give stability to the law.[43] If not, the profession may be doomed to a lack of confidence in its search results. Although there may be manifold explanations for the phenomenon, given the diversity of search services, there is little wonder courts and attorneys are at odds with respect to case citations in decisions and briefs. Cognitive authority suffers in such an environment.

¶17 The second concern I have is that legal research problems that are primarily *subject* in nature are more readily solved, at least historically, by using different techniques (such as indexes and tables of contents with legal commentary) as opposed to *known-item* problems that best use search algorithms and are fact specific ("I need the case where [insert determinative facts and narrow issues]").[44] I have held to this fundamental belief and, in a general sense, have felt that more often than not, research involves *subject* problems ("I need to understand Wisconsin water law"), and thus requires techniques other than search algorithms to resolve. But do the assumptions inherent in all of this continue to hold true, if they were ever true in the first place? More concretely, is natural language processing, a form of AI, and as used by our major vendors, so good that it no longer matters what problem we face because *something* authoritative and relevant will always appear in the search results for any given inquiry (even prior to filtering sources)?[45] Legal commentary or secondary authority (the classic texts the profession has relied on, say, *Williston on Contracts*, *Nimmer on Copyright*, or *Moore's Federal Practice*) might become invisible with relevant primary authority appearing as defaults in search results. Furthermore, new legal search engines lack connection with much of the commentary and treatises, which are a part of the profession's cognitive authority. Finally, natural language processing may ultimately produce outputs to search queries in forms of abstracts or summaries that replace those commentaries and treatises.

¶18 I also wonder whether because of AI and natural language processing, we are observing the death throes of the index and the end of human-intermediated access tools to legal information, such as digests and abstract services. My survey of students was not a fine enough instrument to definitively determine this phenomenon, but it does raise important questions. Furthermore, can natural language processing get us into secondary sources as well as indexes have? And where and how will statutory codes be effectively accessed? They are arranged, at least when codified, with a topical structure and have indexes. These are likely to be ignored in

43. There is an economic issue about whether lawyers with access to low-cost or free services such as Fastcase (often comes with bar subscription), Casetext (free plus premium service), Ravel (free for academics plus a premium service), and Google Scholar (free) can through iterative case searching achieve results truly comparable to Westlaw and Lexis. This needs to be studied.

44. *Compare* Callister, Field Guide, *supra* note 28, at 19–24 *with id.* at 24–34. The example I use with *known-item* problems is "You need the California murder case in which the court found that a fetus cannot be a human being, and the defendant was acquitted of murder after beating up his wife or girlfriend resulting in the loss of the fetus." *Id.* at 19, tbl. 3-2. This contrasts with a subject problem where "You are looking for an explanation of low-income housing credits." *Id.* at 24, tbl. 3-3.

45. Usually, we have to filter to get to secondary sources.

the single-search box world. Later, this article will examine the issue and find that both Lexis Advance and Westlaw Edge have taken steps with predictive searching to preserve the role of secondary materials (although circumventing the use of indexes).[46]

¶19 It may give the librarian reader some comfort that not only is the behemoth that is Thomson Reuters (hereinafter Westlaw) sticking to its traditional strength with human "curated" information,[47] but it even outperforms other systems when it employs AI—at least for the examples I provide in this article with respect to case law research.[48] However, the issue is not just case law, where Westlaw has so much curated information or metadata, but the role of secondary sources and how we access them. And, we need to consider access to codes and whether natural language processing works for them. So along with indexes and human-intermediated information, are secondary sources in danger of extinction, at least in terms of use, if not cessation, in the near future? AI may one day replace the "secondary sources" with cognitive authority by writing its own summaries and expounding the law from primary sources. And if it can do it better than human experts, why not? Abstracting and summarizing is already being done with natural language processing.[49] Even a book has been written by such a tool and published by Springer Nature.[50] It is not a great stretch to apply to legal research and writing. Westlaw has 2.4 million state and federal briefs.[51] This is big data that natural language processing can analyze to learn the structure and nuance of brief writing—if not now, soon. Brief analysis is also a feature of Westlaw Edge, Casetext, Ross Intelligence, and several other vendors' products.[52] Actually writing the brief may not be far behind.[53] Indeed, as far back as 2005, Dan Dabney, Senior Director, Thomson Global Services GmbH, described a future in which attorneys would start writing

---

46. *See infra*, ¶¶ 75–85.

47. Thomson Reuters, *Not All Legal AI Is Created Equal: 5 Things to Consider When Evaluating AI Solutions and Why You Need to Pay Attention*, https://legal.thomsonreuters.com/en/insights/white-papers/5-things-to-consider-when-evaluating-legal-ai-solutions (last visited Aug. 4, 2020).

48. *See infra* ¶¶ 53–74. This is also true for the Nevelow Mart study. *See* Nevelow Mart, *supra* note 34, at 414, chart 2 & tbl. 3 (Westlaw led in precision relevance results).

49. Lina Zhou & Dongsong Zhan, *NLPIR: A Theoretical Framework for Applying Natural Language Processing to Information Retrieval*, 54 J. Am. Soc'y Info. Sci. & Tech. 115, 119 (2003).

50. *See* Michael Riley, *Explainable Artificial Intelligence* 28, 40, in Association of Research Libraries, *Ethics of Artificial Intelligence*, Research Library Issues no. 299, 5 (2019), https://publications.arl.org/rli299/1 (last visited Aug. 4, 2020); *see also* John Nay, *Natural Language Processing and Machine Learning for Law and Policy Texts* 1–2 (Apr. 7, 2018), https://dx.doi.org/10.2139/ssrn.3438276 (last visited Aug. 4, 2020) (With reference to natural language processing and legal texts, "[w]e describe methods for automatically summarizing content (sentiment analyses, text summaries, topic models, extracting attributes and relations, document relevance scoring), predicting outcomes, and answering questions.").

51. Thomson Reuters, Using Westlaw to Write a Brief, https://lscontent.westlaw.com/images/content/L-377268_Brief.pdf (last visited Aug. 4, 2020).

52. Robert J. Ambrogi, *AI-Driven Brief Analysis Comes to Westlaw, But Does It Differ from Competitors?*, LawSites (July 12, 2019), https://www.lawsitesblog.com/2019/07/ai-driven-brief-analysis-comes-to-westlaw-but-does-it-differ-from-competitors.html (last visited Aug. 4, 2020).

53. Similar predictions were raised by a *New Yorker* columnist about whether a machine could write for the *New Yorker*. *See* John Seabrook, *The Next Word, Where Will Predictive Text Take Us?*, New Yorker (Oct. 14, 2019), https://www.newyorker.com/magazine/2019/10/14/can-a-machine-learn-to-write-for-the-new-yorker (last visited Aug. 4, 2020). The technology spawning the article was the author's interaction with predictive text using Smart Compose in Google Email. *Id.*

briefs, and the research would be seamlessly supplied to them.[54] Quoting Dabney, "[W]hat is happening here, at least potentially, is that legal research has ceased to be a particularly separate part of the operation. You can just sit down and write a brief and the authorities you need, the law that you are looking for, will find you."[55] Natural language processing, potentially, could support such a future.

Building a Model of an Information Retrieval System with Natural Language Processing

¶20 The query is not the sole focus of natural language processing, although it is part of it. Increasingly, large databases are mined for their inner relationships among documents. This can be done independently of any search, with the purpose of finding relationships of words or documents to each other. Turning words into vectors is an important technique for this process. There are also other tools besides trigonometric vectors, involving probabilities and neural networks. Researchers experiment with all of these to use natural language processing to discover relationships among documents and words in databases. To understand a little more of this activity, we will start with understanding how word vectors (also known as *embeddings*) are employed.

Trigonometric Word Vectors and Cosine Similarity

¶21 The thing to understand is that once words have been turned into vectors (or dimensions), they can be compared, added or subtracted, and then, quite remarkably, the nearest word can be determined.

¶22 As a famous example, vectors for the following words can be processed on a linear basis:[56]

king – man + woman = queen

What happens is the vectors as processed above result in a vector that is nearest the word *queen*. Essentially *king* is to *man* as *queen* is to *woman*. This is analogical computing, and it is done on what is known as a *linear* basis (meaning mathematical functions apply). All sorts of interesting relationships can develop with this or similar processes. In theory, vectors for *purple – red = blue.* Vectors for *France – Paris + Athens = Greece*.[57] All of this is linear processing.

¶23 To understand the determination of vectors, I need to lay out some more concepts and two frequently used natural language processing models.

¶24 Suppose we have the following co-occurrences of terms:

---

54. *See* Paul D. Callister, *Law and Heidegger's Question Concerning Technology: Prolegomenon to Future Law Librarianship*, 99 Law Libr. J. 285, 293–94, 2006 Law Libr. J. 17, ¶ 19.

55. *Id.*, *citing* Dan Dabney, Envisioning the Future: The Publisher's Perspective, Remarks at the Future of Law Libraries Symposium, Florida Coastal School of Law (Mar. 10–11, 2005) (quoted passage transcribed by author from digital recording no longer available over the Internet).

56. Thomas Mikolov et al., *Efficient Estimation of Word Representations in Vector Space* 2 (Sept. 7, 2013), https://arxiv.org/pdf/1301.3781.pdf (last visited Aug. 4, 2020).

57. Thomas Mikolov et al., *Distributed Representations of Words and Phrases and Their Compositionality* 4 (Oct. 16, 2013), https://arxiv.org/pdf/1310.4546.pdf (last visited Aug. 4, 2020).

Table 1

	Rent	Sale	Credit
Warranty	3	4	3
Lease	6	0	2
Contract	0	6	4

We want to know whether *warranty* (including its stem, *warrant*) is closer to *contract* or *lease* (at least as word vectors). To do this, we need the cosines of our terms in relation to each other, or at least:[58]

cos(warranty, contract) and cos(warranty, lease)

The formula for calculating the cosine of two vectors is:

$$\frac{\vec{v} \cdot \vec{w}}{|\vec{v}||\vec{w}|}$$

In the numerator is the "dot product" of the two vectors, which is essentially multiplying two vectors (or each of their members in the set) by each other. In the denominator is the multiplication of each vector's length times the other. I will illustrate later. Let's start with the numerator. It can be calculated:

$$\sum_{i=1}^{N} v_i w_i$$

The denominator (or multiplication of vector lengths) is calculated by squaring each vector set member, adding them together, and then taking the square root. This is done for each set, the results of which are then multiplied by each other. The whole process is summarized:

$$\sqrt{\sum_{i=1}^{N} v_i^2} \sqrt{\sum_{i=1}^{N} w_i^2}$$

Consequently, with the data in the table above, we calculate the numerator for *warranty* and *contract* vectors:

$$(3 \times 0) + (4 \times 6) + (3 \times 4) = 36$$

And the denominator for *warranty* and *contract* vectors:

$$\sqrt{3^2 + 4^2 + 3^2} \times \sqrt{0^2 + 6^2 + 4^2} = 42.04759$$

The cosine requires that we simply divide the numerator and the denominator, which for *warranty* and *contract* equals 0.856173.

58. For the math involved in this and the following computations, see Daniel Jurafsky & James H. Martin, *Vector Semantics*, *in* Speech and Language Processing 11–12 (3d ed., draft Sept. 23, 2018), https://web.stanford.edu/~jurafsky/slp3/6.pdf (last visited Aug. 4, 2020).

Repeating the step for the numerator of the *warranty* and *lease* vectors:

$$(3 \times 6) + (4 \times 0) + (3 \times 2) = 24$$

And the denominator for *warranty* and *lease* vectors:

$$\sqrt{3^2 + 4^2 + 3^2} \times \sqrt{6^2 + 0^2 + 2^2} = 36.87818$$

Again, the cosine requires that we simply divide the numerator and the denominator, which for *warranty* and *lease* equals 0.650791. Thus comparing the two cosines, *warranty* has greater similarity to *contract* than it does to *lease* (at least within our hypothetical database with its small set of data).

¶25 These vectors involve combining vectors for single words, with at least three dimensions (*rent*, *sale*, and *credit*). Extracting cosines can help us cluster words and identify topics in documents or sentences. In large search databases, like Google or *Federal Supplement* on Lexis or Westlaw, the dimensions may be much larger.

¶26 As we scale up for larger databases:

> It turns out, however, that simple frequency isn't the best measure of association between words. One problem is that raw frequency is very skewed and not very discriminative. If we want to know what contexts are shared by [the list of terms], we're not going to get good discrimination from words like the, it, or they, which occur frequently with all sorts of words and aren't informative about any particular word. . . . the [dimension] for the word good is not very discriminative between [Shakespeare's] plays; good is simply a frequent word and has roughly equivalent high frequencies in each of the plays.[59]

The problem addressed above can be addressed by methods I will introduce below. It can also be addressed by filtering *stop words*.[60] The problem with stop words is identifying them ahead of time. In law, there are lots of terms whose frequency across documents in a database may not be helpful in classifying documents—for example, *law*, *legal*, *court*, *judge*, *ruling*. We can either create a list of stop words or decrease the weight of such terms. The latter is preferable because stop words have an absolute rather than graduated effect. The next section will treat this topic by showing a more sophisticated technique.

TF-IDF or Term Frequency and Inverse Document Frequency

¶27 The basic idea behind stop words is to give less weight to terms that occur too frequently in a document, especially if they also appear frequently across all of the documents of a database. Logarithmic scales (based on orders of magnitude) are used to give weights to words and inverse weights if words appear frequently in documents across the database. For the first part of the calculation, a word that appears 10 times in a document would have a weighted term frequency (*tf*) of 2, and one that appeared 100 times would have a weighted *tf* of 3. These *tf*s are then multiplied by the inverse of the term in document frequency (also on an logarithmic scale), known as *idf*.

---

59. *See id.* at 12.

60. *See* Billel Aklouche, Ibrahim Bounhas & Yahya Slimani, *Query Expansion on NLP and Word Embeddings*, Twenty-Seventh Text Retrieval Conference (TREC 2018) Proceedings 2–3, https://trec.nist.gov/pubs/trec27/papers/JARIR-CC.pdf (last visited Aug. 4, 2020).

¶28 Here is the formula for term weight:[61]

$$w_{t,d} = tf_{t,d} \times idf_t$$

So essentially the weight of the term in a given instance of a document is the product of the term frequency in the document (which will be defined below) and the inverse of the term in all documents (which is also defined below). To calculate the term frequency in a document, we use a logarithmic function:

$$\mathrm{tf}_{t,d} = \begin{cases} 1 + \log_{10} \mathrm{count}(t,d) & \text{if } \mathrm{count}(t,d) > 0 \\ 0 & \text{otherwise} \end{cases}$$

So term frequency in a document is determined by first deciding whether the count of the term in the document is greater than zero. If so, the term frequency is represented by adding one to the base 10 logarithm of count of the term in the document. So if the term appears 5 times in the document, 1 + the base 10 logarithm of 5 equals 1.698970004.

¶29 The next step and factor is designed "to give a higher weight to words that only occur in a few documents. Terms that are limited to a few documents are useful for discriminating those documents from the rest of the collection."[62] To calculate the inverse document frequency (*idf*) for a term:

$$\mathrm{idf}_t = \log_{10}\left(\frac{N}{\mathrm{df}_t}\right)$$

N is the number of documents in the database, and df_t is the frequency of documents with the term. Thus the number of documents in the database is divided by the frequency of documents with the term. Then a logarithm base 10 is determined. So if the number of documents in the database is 10, and the number of documents with the term is 5, the result is the logarithm base 10 of 2 (10/5), which is 0.301029996.

¶30 Ultimately, we must determine the product of the previous two functions to gives us the appropriate weight, or in our example 0.511440933.

¶31 We might use the following hypothetical data to illustrate term frequencies across different documents in a database of 1000 documents.

Table 2

Term	Doc 1	Doc 2	Doc 3	Doc 4
law	30	7	17	50
legal	5	9	26	22
tax	105	8	45	3
credit	12	45	5	15
housing	12	21	0	2
allocate	7	19	1	5

61. For math that follows, see Jurafsky & Martin, *supra* note 58, at 13–14.
62. *Id.* at 13.

Using an Excel spreadsheet, the appropriate weight for the terms in Document 1 can be calculated as follows:

Table 3

Term	Term Frequency in Document 1	Document Frequency for Term	1+Log10 for tf	log10(n/dft) or idft	Number of Docs in Database (N)	Product (wtd)
law	30	1000	2.48	–	1,000	—
legal	5	900	1.70	0.05	1,000	0.08
tax	105	90	3.02	1.05	1,000	3.16
credit	12	130	2.08	0.89	1,000	1.84
housing	12	5	2.08	2.30	1,000	4.78
allocate	7	85	1.85	1.07	1,000	1.98

Here we have calculated the appropriate weight of a term in a document (*wtd*) using inverse factors for high frequency of documents with a term across the database (idf_t). Logarithmic functions have facilitated this. The size of our database, or *N*, is 1000 documents.

¶32 We repeat this process for all four documents and get the following weights across the four documents:

Table 4

	Term Weights tf-idf			
Term	**Doc1**	**Doc2**	**Doc3**	**Doc4**
law	–	–	–	–
legal	0.08	0.09	0.11	0.11
tax	3.16	1.99	2.77	1.54
credit	1.84	2.35	1.51	1.93
housing	4.78	5.34	–	2.99
allocate	1.98	2.44	1.07	1.82

Notice that *law* has no weight. This is because it was found in every document in the database (all 1000). As we might expect, *legal* has very low weight because of its frequency (900/1000).

¶33 From these term weights, we can now rank which documents are most similar and dissimilar based on the terms (which is very useful in natural language searching). We can do this by determining the comparative cosines of each of the documents based on their shared terms. For Documents 1 and 2, we make a *dot product* calculation:

$$(.08\times.09) + (3.16\times1.99) + (1.84\times2.35) + (4.78\times5.34) + (1.98\times2.44) = 41.01$$

¶34 We then calculate the square root of the sum of each term frequency squared. We do this for each document. For Document 1,

$$\sqrt{.08^2 + 3.16^2 + 1.84^2 + 4.78^2 + 1.98^2} = 6.34$$

For Document 2,

$$\sqrt{.09^2 + 1.99^2 + 2.35^2 + 5.34^2 + 2.44^2} = 6.63$$

We then perform the following operation:

$$\frac{41.01}{6.34 \times 6.63}$$

The result is a combined cosine for Documents 1 and 2 of 0.98. This means that with respect to the terms we have identified, the documents are almost identical. We can make those comparisons with all of the documents.

Table 5

Summary of Cosines

	Doc 1	**Doc 2**	**Doc 3**	**Doc 4**
Doc 1	1.00	0.98	0.65	0.13
Doc 2	0.98	1.00	0.53	0.16
Doc 3	0.65	0.53	1.00	0.64
Doc 4	0.13	0.16	0.64	1.00

Based on this, we can see that Document 2 is most like Document 1. Document 3 is the next nearest to Document 1, and Document 4 is not very similar at all to Document 1. We haven't exactly created a search engine, but we have a model that would be useful in creating one. Uses of vector comparisons between documents might include Westlaw folder analysis, where based on the documents in a folder Westlaw Edge suggests additional similar documents.[63] Also, it could be used to compare cases for matching with Westlaw Edge's headnotes and Topic and Key Number System, and for Lexis's topic classification system (including uses with "More like this Headnote"). In addition, Westlaw Edge's new *overruling risk* could work by locating documents with similar cosines to a document that has been expressly overruled.[64] I am sure there is more to it than is outlined here, but vector cosines are a powerful tool in helping us think about how our major database services work.

Centroids and Document Similarity

¶35 There is a simple technique for ranking the documents based on the term weights (rather than cosines), and that is to take the average of the sum of term weights for each document. This figure is known as the centroid, and can be used to rank documents in order.

63. *See* Thomson Reuters, *Not All Legal AI Is Created Equal*, *supra* note 47.

64. *See* Thomson Reuters, *KeyCite Overruling Risk: Always Know You're Citing Good Law with Thomson Reuters Westlaw Edge*, https://legal.thomsonreuters.com/en/products/westlaw/edge/keycite-overruling-risk (last visited Aug. 4, 2020).

Table 6

Term	Doc 1	Doc 2	Doc 3	Doc 4
		Term Weights tf-idf		
law	–	–	–	–
legal	0.08	0.09	0.11	0.11
tax	3.16	1.99	2.77	1.54
credit	1.84	2.35	1.51	1.93
housing	4.78	5.34	–	2.99
allocate	1.98	2.44	1.07	1.82
centroid	2.37	2.44	1.09	1.68

Looking at the centroids, we can see that Document 2 is closest to our terms, and Document 1 is a close second. Document 4 is a distant third. The Summary of Cosines is helpful in a search because once a document has been identified as relevant, it is easy to discover what other documents might be relevant.

¶36 Perhaps the most common way to use our packet of tools is to calculate the cosines of words across documents (we have already calculated the cosines of documents across words). Imagine doing this across all 1000 documents in our hypothetical database (or across all the case law in Lexis's *Federal Supplement* database) and using a vocabulary that extends into the tens or hundreds of thousands of words.[65] We can, quite handily, find the synonyms for legal terms by finding the closest cosines using this technique.[66] This can be helpful in expanding the initial query to appropriately related concepts and topics.

Queries as Vectors—An Example

¶37 Not only can documents and words be vectors, but queries can be vectors.[67] Imagine the calculations of the following query for *tax*, *credit*, *housing*, and *allocate*.[68] We can calculate the TF-IDF term weights as follows:

Table 7

Term	Term Frequency in Query 1	Document Frequency for Term	1+Log10 for tf	Log10(n/dft) or idft	Number of Docs in Database (N)	Product (wtd)
law	0	1000	0	–	1,000	–
legal	0	900	0	0.05	1,000	–
tax	1	90	1	1.05	1,000	1.05
credit	1	130	1	0.89	1,000	0.89
housing	1	5	1	2.30	1,000	2.30
allocate	1	85	1	1.07	1,000	1.07

65. *See* Jurafsky & Martin, *supra* note 58, at 15.

66. *Id.* ("[W]e can find the 10 most similar words to any target word *w* by computing the cosines between *w* and each of the *V*-1 other words, sorting, and looking at the top 10.").

67. *See* Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze, Introduction to Information Retrieval 113–14 (2008).

68. In a real search engine, word stems for the terms might be used.

We now have a table of term weights that looks like the following:

Table 8

Term Weights tf-idf					
Term	**Doc 1**	**Doc 2**	**Doc 3**	**Doc 4**	**Query 1**
law	0.00	0.00	0.00	–	0.00
legal	0.08	0.09	0.11	0.11	0.00
tax	3.16	1.99	2.77	1.54	1.05
credit	1.84	2.35	1.51	1.93	0.89
housing	4.78	5.34	–	2.99	2.30
allocate	1.98	2.44	1.07	1.82	1.07

From there we can determine combined cosines for Query 1 and each of the documents using the techniques described above.

Table 9

Cosines for Query and Documents				
	Doc 1	**Doc 2**	**Doc 3**	**Doc 4**
Query1	0.9879	0.9968	0.5594	0.9832

The ranking of documents in a natural language search based on our query is thus: Document 2, Document 1, Document 4, and Document 3, with Documents 2, 1, and 4 being near perfect vector matches.

¶38 Note that the query terms in Table 7 are represented by a 1 or a 0. The zero means we are not including a term from the vocabulary of terms in the database. There would be thousands or tens of thousands of zeros for any given search. Also note that natural language search terms rarely include the term twice, but if a search phrase repeated a term, it could affect the weight of the term in the search, but on a logarithmic scale,[69] and the ultimate ranking of results. This gives the user just a little control in emphasizing certain terms in search results by repeating terms.[70]

¶39 I have presented only the simplest versions of plausible natural language processing with our major online services. Yet, cosine calculations using TF-IDF with database vocabularies of thousands of words (or dimensions if we think in terms of vectors) are voluminous calculations that are costly for computing time.[71] Personally, I marvel to think what might go into our legal search engines.

69. *See* tbl. 7 (the column for *1+Log10 for tf* would be affected on a logarithmic scale).

70. For an additional technique to add weight, see *infra* note 139 and accompanying text.

71. Manning, Raghavan & Schütze, *supra* note 67, at 114.

> Computing the cosine similarities between the query vector and each document vector in the collection, sorting the resulting scores and selecting the top *K* [ranked] documents can be expensive – a single similarity computation can entail a dot product in tens of thousands of dimensions, demanding tens of thousands of arithmetic operations.

Id.

¶40 There is a limitation with this kind of model. It assumes that co-occurrence of terms in a legal document establishes a relationship that is relevant to the search query. Obviously, if we consider case law, many issues (which may be unrelated) may be found in a single case. Think of cases that are unrelated for 95% of the content, but on the precise issue in question they are related. Now, if the co-occurrence were limited by context (i.e., a range of words proximate to a targeted term), rather than co-occurrence scattered throughout the whole document, we might get more relevant search results. Our next model addresses this issue.

word2vec and Skip-gram with Negative Sampling

¶41 Cosine calculations using TF-IDF use thousands and tens of thousands of references (called *dimensions*) of co-occurring words in a document. The resulting vectors or *embeddings* are said to be "sparse," because most of the weights are zero. However, using probabilities instead of cosines employs far fewer dimensions to be used with each word, and are said to be "dense," with many more positive weights. As few as 50 dimensions can be used, but the ranges often extend into the hundreds (rather than the tens of thousands).[72] This simplifies calculations and produces more accurate results.

¶42 Skip-gram with negative sampling (SGNS) is one of two techniques for dense vectors offered by a research package known as word2vec.[73] "Skip-gram predicts surrounding words given the current word."[74] It works by introducing words "into the vector one at a time, and scanning back and forth within a certain range."[75] Context of terms is important to the technique.

¶43 Because of the lengthy complexity of the math (and, candidly, my inability to understand it in other than the most general terms), I will not illustrate it here but, essentially,[76] a target word is identified with a given number of words to the

---

72. *See* Jurafsky & Martin, *supra* note 58, at 18. This is relatively few compared to the embeddings or dimensions described in note 66, *supra*, and accompanying text.

73. "[A]n n-gram is a contiguous sequence of *n* items from a given sample of text or speech." Wikipedia, *n-gram*, https://en.wikipedia.org/wiki/N-gram (last visited Aug. 4, 2020). "Skipgrams are ngrams in which the items are not necessarily contiguous." Skymind, *Glossary* (entry for *skipgrams*), A.I. Wiki, https://skymind.ai/wiki/glossary (last visited Aug. 4, 2020). They are not contiguous because rare words and frequently used words are discarded. Yoav Goldberg & Omer Levy, *word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method* 5 (Feb. 14, 2014), https://arxiv.org/pdf/1402.3722.pdf (last visited Aug. 4, 2020).

74. Mikolov et al., *supra* note 56, at 5, fig.1. In essence, a skip-gram is the opposite of another tool offered by word2vec known as CBOW (Continuous Bag of Words), which predicts the target word based on the context. *Id.* The skip-gram method predicts the context words and "produces more accurate results on large datasets" than CBOW. Skymind, *A Beginner's Guide to Word2Vec and Neural Word Embeddings*, A.I. Wiki, https://skymind.ai/wiki/word2vec (last visited Aug. 4, 2020). The entire text corpus is processed. *Id.*

75. Skymind, *supra* note 74.

76. *See* Jurafsky & Martin, *supra* note 58, at 19–23. Dot products for cosines are turned into probability using a sigmoid function:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Id. at 20, formula 6.25. The function helps determine the probability for the context word for a given target word as well as for random words not being context words ($-x=-t\cdot c$). *See also* Xin Rong, *word2vec Parameter Learning Explained* (June 5, 2016), https://arxiv.org/pdf/1411.2738 (last visited Aug. 4, 2020).

right and left of it, known as context words. Probabilities are calculated for context words. These are considered positive examples of dimensions or embeddings. An even larger set of negative examples is generated at random from the lexicon for the database. These words are known as "noise words." Using probabilistic statistics and statistical regression (and assuming the probabilities for positive and negative examples equals one), weights for embeddings are created for the words in the document. Not only does SGNS yield shorter embeddings, but it does a better job of generalizing and capturing synonyms.[77]

¶44 SGNS works according to the following general rules:

1. Treat the target word and neighboring context word as positive examples;
2. Randomly sample other words in the lexicon to get negative samples;
3. Use logistic regression to train a classifier to distinguish those two cases; and
4. Use the regression weights as embeddings.[78]

Classifiers will be treated below,[79] but basically the end results are weights for positive and negative examples of the probabilities, which collectively equal one. Having produced embeddings for document databases, skip-gram tools can compare the results for similarity of embeddings for search phrases.

Potential Use of word2vec in Information Retrieval

¶45 As with cosine vectors, the potential exists for vectors coming from word2vec to be used with information retrieval. However, such vectors have been primarily used in word clustering or topic modeling.[80] But there are exceptions.

> For instance, if we obtain vector representations of a collection of texts we can apply clustering algorithms directly to these representations to automatically group similar documents together to facilitate searching through a large corpus. Or we can apply supervised learning models that predict an outcome [such as passage of a bill] to the text. The possibilities are almost endless.[81]

¶46 Other data scientists have directly proposed that word2vec vectors or embeddings be used in information retrieval.[82] Unfortunately, we simply do not know whether such tools are used with the major legal databases but, as shall be seen below,[83] we can make some educated guesses.

---

77. *See* Jurafsky & Martin, *supra* note 58, at 18.

78. *Id.* at 19.

79. *See infra* ¶¶ 48–50.

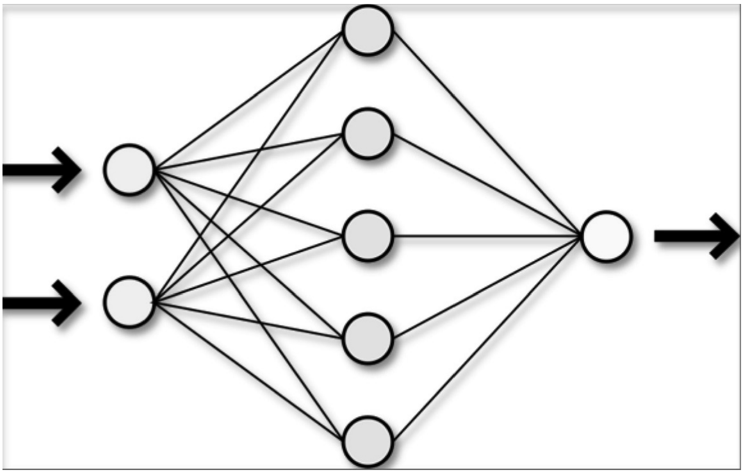
80. *See, e.g.*, Peter Grajzl & Peter Murrell, *Estimating a Culture: Bacon, Coke and Seventeenth-Century England* (Sept. 20, 2019), https://dx.doi.org/10.2139/ssrn.3319386 (last visited Aug. 4, 2020). "[W]e undertake a quantitative, machine learning analysis of the writings of two intellectual giants, Edward Coke (1552–1634) and Francis Bacon (1561–1626)." *Id.* at 1.

81. Nay, *supra* note 50, at 6.

82. *See* Aklouche, Bounhas & Slimani, *supra* note 60, at 3 ("[T]he new terms [for query expansion and from word vectors were] selected based on their similarity to the entire query or their similarity to its individual terms."); Eric Nalisnick et al., *Improving Document Ranking with Dual Word Embeddings* (Apr. 2016), https://www.microsoft.com/en-us/research/wp-content/uploads/2016/04/pp1291-Nalisnick.pdf (last visited Aug. 4, 2020); Zhou & Zhan, *supra* note 49, at 115.

83. *See infra* ¶¶ 53–74.

Figure 2
Neural Network
Source: Mysid Dake, WIKIMEDIA, Licensed for use under Creative Commons 1.0 Generic License.



Regression and Neural Networks

¶47 One of the most advanced techniques for determining embeddings is the use of neural networks (see figure 2). Such computer networks test regressions of variables for terms until the optimum is reached.[84] The number of layers adds to the complexity and computing power of the network. Describing their function other than for use in regression and classification is beyond the scope of this article. The nodes do tend to work by recognizing patterns and using inference, rather than instruction.[85] It is unknown whether the major vendors use neural networks.

Classification

¶48 Besides word vectors, some of the tools of natural language processing come from statistics and include Bayes classification. A basic use is to determine whether text or a document falls within a class. For example, is an email spam? Another use is to determine sentiment—is a movie review positive?[86] In our context of legal research, we can imagine that the major vendors—Lexis, Westlaw, and Bloomberg—could use such techniques to know whether a case has negative treat-

84. For further reading, see Skymind, *A Beginner's Guide to Backpropagation in Neural Networks*, A.I. WIKI, https://skymind.ai/wiki/backpropagation (last visited Aug. 4, 2020); Skymind, *A Beginner's Guide to Neural Networks and Deep Learning*, A.I. WIKI, https://pathmind.com/wiki/neural-network (last visited Aug. 4, 2020) ("Anyone who understands *linear regression* . . . can understand how a neural net works."); and DANIEL JURAFSKY & JAMES H. MARTIN, *Logistic Regression*, *in* SPEECH AND LANGUAGE PROCESSING 1 (3d ed., draft Aug. 25, 2019), https://web.stanford.edu/~jurafsky/slp3/5.pdf (last visited Aug. 4, 2020) ("[A] neural network can be viewed as a series of logistic regression classifiers stacked on top of each other.").

85. *See* WIKIPEDIA, *Machine Learning*, https://en.wikipedia.org/wiki/Machine_learning (last visited Aug. 4, 2020).

86. *See* DANIEL JURAFSKY & JAMES H. MARTIN, *Naive Bayes and Sentiment Classification*, *in* SPEECH AND LANGUAGE PROCESSING 1 (3d ed., draft Aug. 25, 2019), https://web.stanford.edu/~jurafsky/slp3/4.pdf (last visited Aug. 4, 2020).

ment or is distinguished in another case, or with classification in Westlaw Edge's Key Number System, or Lexis Advance's "More like this Headnote." Of course, because these systems are proprietary, we do not know for certain which natural language processing tools are being used.[87] Nonetheless, it is enlightening to study how various natural language processing tools work.

¶49 Suppose we are trying to train a case law database natural language processing tool to assess whether a case has negative treatment. We might use a training database of expression that looks like the following:

Table 10[88]

	Negative Case Treatment	
Training	**Category**	**Document Phrases**
	Pos	clearly distinguished by
	Pos	this case was effectively overturned
	Pos	the facts do not set on all fours with
	Neg	closely followed by
	Neg	confirms the holding from
Test	?	case distinguished from

Essentially, phrases in the positive category indicate that subsequent treatment in a case is negative treatment of the original cases. We are going to apply an algorithm known as Naive Bayes Classifier to test the bottom document phrase in the table. We first take the likelihood of negative and positive categories in our dataset of sentence phrases.[89]

$$P(+) = \frac{2}{5} = 0.4$$

$$P(-) = \frac{3}{5} = 0.6$$

We next need the likelihood of the three words in the test phrase being classed as positive or negative.[90]

$$P(\text{"case"}|+) = \frac{count(w,c)+1}{\left(\sum_{w \in V} count(w,c)\right) + |V|} = \frac{1+1}{8+17} = 0.08$$

87. *See supra* note 8.

88. Note that the word *by* is counted only once for the |*v*| or the sum of all negative and positive classifiers excluding duplicates.

89. *See* Jurafsky & Martin, *supra* note 86, at 7.

90. *Id.*

$$P(\text{"case"}|\text{-}) = \frac{count(w,c)+1}{(\sum_{w \in V} count(w,c)) + |V|} = \frac{0+1}{10+17} = 0.037037$$

$$P(\text{"distinguished"}|+) = \frac{count(w,c)+1}{(\sum_{w \in V} count(w,c)) + |V|} = \frac{1+1}{8+17} = 0.08$$

$$P(\text{"distinguished"}|-) = \frac{count(w,c)+1}{(\sum_{w \in V} count(w,c)) + |V|} = \frac{0+1}{10+17} = 0.037037$$

$$P(\text{"from"}|+) = \frac{count(w,c)+1}{(\sum_{w \in V} count(w,c)) + |V|} = \frac{0+1}{8+17} = 0.04$$

$$P(\text{"from"}|-) = \frac{count(w,c)+1}{(\sum_{w \in V} count(w,c)) + |V|} = \frac{1+1}{10+17} = 0.074074$$

Here, *w* is the word, *c* is the class illustration of positive or negative phrases, and $|V|$ is the union of all word types in all classes (i.e., no counting of duplicate words across the class). The two classes are then multiplied by each result for the class (i.e., all positive classes multiplied against each other, and all negative classes against each other). Finally, the results (the probability of our test sentence phrase) are multiplied by the likelihood of the positive and negative categories.[91]

$$P(+)P(S|+) = 0.04 \times (0.08 \times 0.08 \times 0.04) = 1.024 \times 10^{-4}$$
$$P(-)P(S|-) = 0.06 \times (0.037037 \times 0.037037 \times 0.074074) = 6.0966 \times 10^{-5}$$

¶50 It is more probable that our test sentence phrase (*S*) represents a positive instance of negative case history. Thus, we can see the mechanics of how classification of negative case history occurs. It is all based on probability comparisons. This is a far departure from the human expert system established by Shepard's, but coded classifications of when a case has negative treatment is doable as KeyCite and as the above exercise in the probabilities of a classification system have each demonstrated. Furthermore, such systems depend on some human training. In a blog post on *Above the Law*, David Lat confirmed that Thomson Reuters' KeyCite Overruling Risk (which tracks not only the original cases, but related cases whose overruling may affect the original case) relies upon "natural language processing as well as traditional, supervised machine learning [probably referring to human training and testing]."[92] Well, we do not know for certain which natural language processes were

91. *Id.*
92. David Lat, *How Artificial Intelligence Is Transforming Legal Research*, Above the Law 2020,

used; however, the Bayesian exercise we have gone through, or related statistical tools, could be adapted for use. It is important to understand that humans train such systems, but then the systems are turned loose for disintermediated service to the legal profession. How much do we trust opaque natural language processes to approximate a legal mind in identifying negative precedent, or is that even the right question? Perhaps the real issue is for data scientists who can compare results from natural language processing systems against one another.

Grammatical Techniques

¶51 Another use of Bayesian techniques is to combine them with Hidden Markov Models to identify parts of speech and phrases of particular value.[93] One paper on the topic connected sentence structure to "better extraction accuracy,"[94] referring to information extraction of "specified classes or relations from text."[95] With such techniques, sentence structure, rather than co-occurrence of related terms, matters.

¶52 Nevelow Mart has distinguished between key word searching (like when two or three terms are entered into Google) and natural language searching (which is aided by "grammatical structure models").[96] I am not exactly sure what is meant by "grammatical structure models," but statistical techniques employing Hidden Markov Models have been employed to match words with their "parts of speech."[97] For some time, such techniques have also been used for "information extraction" including "summarizing collections of documents, and identifying significant but unknown relations among objects" and for information retrieval systems.[98] Besides parts of speech, other grammatical rules may be in play. One article from 2003 on the subject noted that "grammatical rules contribute to most of the behavior of Natural Language parsers."[99] However, parsing long sentences, even using a complex set of grammatical rules, is avoided due to difficulty.[100] That said, generally, employment of grammatical rules and Hidden Markov Models are beyond the scope of this article, which instead focuses on word vectors (and given the results

https://abovethelaw.com/law2020/how-artificial-intelligence-is-transforming-legal-research (last visited Aug. 4, 2020). The role of natural language processing is also confirmed by Jason Tashea, *Thomson Reuters Announces Westlaw Edge, Increasing Role of AI Analytics*, A.B.A. J. (July 12, 2018), http://www.abajournal.com/news/article/thomson_reuters_announces_westlaw_edge_increasing_role_of_ai_and_analytics (last visited Aug. 4, 2020); *see also KeyCite Overruling Risk*, *supra* note 64.

93. *See* Soumya Ray & Mark Craven, *Representing Sentence Structure in Hidden Markov Models for Information Extraction*, Proceedings of the 17th International Joint Conference on Artificial Intelligence 1 (2001), https://www.biostat.wisc.edu/~craven/papers/ijcai01-hmm.pdf (last visited Aug. 4, 2020); David R.H. Miller, Tim Leek & Richard Schwartz, *A Hidden Markov Model Information Retrieval System*, Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval 214 (Aug. 15–19, 1999), https://dl.acm.org/citation.cfm?id=312680 (last visited Aug. 5, 2020).

94. Ray & Craven, *supra* note 93, at 2.

95. *Id.* at 1.

96. *See* Nevelow Mart, *supra* note 34, at 397, ¶ 16.

97. Luis Serrano, *A Friendly Introduction to Bayes Theorem and Hidden Markov Models*, YouTube (Mar. 27, 2018), https://youtu.be/kqSzLo9fenk (last visited Aug. 5, 2020) (at 31:25).

98. *See* Ray & Craven, *supra* note 93, at 1.

99. Takeshi Matsumoto, David M.W. Powers & Geoff Jarrad, *Application of Search Algorithms to Natural Language Processing*, Proceedings of the Australasian Language Technology Workshop 2003, at 1, https://www.aclweb.org/anthology/U03-1003/ (last visited Aug. 5, 2020).

100. *See id.*

of the next section, which breaks up sentence structure into test searches, this emphasis turns out to be appropriate).

Bag of Words, word2vec, and the Performance of Current Legal Search Engines

¶53 The techniques of TF-IDF vector comparisons and word2vec skip-grams with negative sampling (SGNS) that we have been using in ¶¶41–44 treat all terms as a "bag of words," meaning word order and proximity (at least within the range of context words for SGNS) do not matter. We can ask the question whether this is the case in a review of major and minor case law legal research services. In investigating the issue of "bag of words," I have done some unusual things—run a search on a phrase backwards and then scramble it to see whether similar results would appear. I do not claim this to be the most scientific and disciplined study (it grew organically from a single problem); rather, the activity illustrates, using a fairly difficult and real research question, some of the issues we presently face and the extent to which vendors may incorporate natural language processing. I will explain the research issue more fully as we go through the results.

Westlaw Edge

¶54 Table 11 illustrates a fairly sophisticated search, where we then reverse the order of the search terms and finally scramble them. We then compare the top 10 results (at least for precision)[101] to see whether Westlaw Edge employs a "bag of words" vector approach.[102]

Table 11

Westlaw Edge		
Search Query (federal cases)	**Case**	**Relevant**
qualified retirement plan also meeting 105(c) as a health and disability plan	Beisler v. Comm'r	Y
	Caplin v. U.S.	Y
	Rosen v. U.S.	Y
	Gibson v. U.S.	Y
	Craft v. U.S.	Y
	Berman v. Comm'r	Y
	Abbate v. Spear	N
	Estate of Barnhorst v. Comm'r	Y
	Beisler v. Comm'r (first hearing)	Y
	Hines v. Comm'r	N

101. *See* F.W. Lancaster, *Precision and Recall*, 2 Encyclopedia of Library and Information Science 2346, 2346 (Marcia J. Bates, Mary Niles Maack & Miriam Drake eds., 2d ed. 2003).

102. Except as noted, the date of the following searches was July 9, 2019. I have already discovered that rerunning searches at later dates provides slightly different results. This raises interesting questions about what is changing—perhaps the systems are reacting to training in the form of aggregating user responses. Nevelow Mart documents that Fastcase "aggregates history of more than 100 million searches." Nevelow Mart et al, *supra* note 8, at 13. Probably all of the services take advantage of aggregated user history.

Search Query (backwards)	Case	Relevant
plan disability and health a as 105(c) meeting also plan retirement qualified	Beisler v. Comm'r	Y
	Berman v. Comm'r	Y
	Caplin v. U.S.	Y
	Rosen v. U.S.	Y
	Estate of Barnhorst v. Comm'r	Y
	Gibson v. U.S.	Y
	Craft v. U.S.	Y
	Paul v. U.S.	Y
	Zardo v. Comm'r	Y
	Hines v. Comm'r	N
Search Query (scrambled)	**Case**	**Relevant**
and retirement qualified plan plan also a 105(c) meeting disability as health	Beisler v. Comm'r	Y
	Caplin v. U.S.	Y
	Berman v. Comm'r	Y
	Rosen v. U.S.	Y
	Hines v. Comm'r	N
	Craft v. U.S.	Y
	Paul v. U.S.	Y
	Gibson v. U.S.	Y
	Estate of Barnhorst v. Comm'r	Y
	Kelter v. Comm'r	Y

The issue I am researching lies at the intersection of pension law and two differing tax code schemes for plans that are excluded from income. One excludes retirement plan contributions and earnings; the other excludes health/accident plan contributions, earnings, and distributions based on the nature of the injury. As a former tax and pension attorney, I can state that the goal of some tax attorneys is to combine the two different kinds of plans (or at least claim that a retirement plan also operated as a health/accident plan). It is a tough issue to research and requires vigorous inquiry into case law.

¶55 *Beisler* and *Caplin* were two of the cases I relied upon heavily in practice and in publication on this topic,[103] although there were many more cases that addressed the issue. Note the results are very similar regardless of how the query is entered—even if backwards or scrambled. In fact, the backwards and scrambled queries have only one irrelevant result each (marginally better than the straightforward search). This stresses how little word order and proximity (within the context range for SGNS) may matter to the search engine. This suggests use of word vectors (whether determined as cosines of co-occurrence or as probabilities based on context). Also note, however, the lists, while very close, are not exactly identical. An

103. *See* Paul D. Callister, *Dual Purpose Retirement/Disability Plans: Can a State Ignore Federal Precedent?*, 8 J. Multistate Tax'n 119, 121, nn.8, 11 (1998); Paul D. Callister, *Dual-Purpose Retirement/Section 105(c) Disability Plans: Chasing the End of the Rainbow or Sound Planning*, Cal. Tax Law., Summer 1998, at 3, 4–5.

educated guess is that while Westlaw Edge is using "bag of words" vector comparison algorithms discussed in the previous section, there may be some small consideration of word proximity and order in its algorithms, but we really don't know.[104]

¶56 The research director at Thomson Reuters freely admits that tinkering with the algorithm may produce different results in terms of recall and precision—some algorithms favoring the former and some the latter:

> As Curtis explains, "In technological terms, it's the trade-off between precision—finding the right things—versus recall—finding all the things. If we tune an algorithm to score high in precision by delivering the 'best' results, there's a trade-off in terms of some relevant results getting missed—in other words, lower recall. On the other hand, if we tune an algorithm to score high in recall by not missing relevant results, the best results will be somewhere in there, but not at the top, or maybe even buried—in other words, lower precision."[105]

¶57 This balancing act affects two very different measures of search success—precision measures how many results from an inquiry are relevant, but recall measures how many relevant documents were found compared to relevant documents in the database as a whole.[106] Furthermore, what Curtis is acknowledging is that you cannot have it both ways—both high precision and recall scores.

¶58 Additionally, the range of the context window[107] for a targeted term in a skip-gram is subject to adjustments for words that are too frequent or too few in the database. Thus, the window size is dynamic.[108] Indeed, the size of the context window (known as the *k* variable) is set by the algorithm's designer.[109] How big a context window is can have a huge impact on results:

> Shorter context windows tend to lead to representations that are a bit more syntactic, since the information is coming from immediately nearby words. When vectors are computed from short context windows, the most similar words to a target word *w* tend to be semantically similar words with the same parts of speech. When vectors are computed from long context windows, the highest cosine words to a target word *w* tend to be words that are topically related but not similar.[110]

There are many variables that have to be selected for a search algorithm—it is a balancing act—and they may a make tremendous difference in terms of results. Thus, we should expect a significant variation in results among legal research services.

Lexis Advance

¶59 The results in Lexis Advance suggest that it is probably not using a "bag of words" vector comparison approach and that word order and, perhaps, proximity matter. The initial results are less than stellar for this complex search.[111] In fact, only

---

104. *See* Nevelow Mart et al., *supra* note 8.
105. Lat, *supra* note 92.
106. *See* Lancaster, *supra* note 101, at 2346.
107. *See supra* note 74 and accompanying text.
108. *See* Goldberg and Levy, *supra* note 73, at 4–5.
109. *See id.* at 4.
110. Jurafsky & Martin, *supra* note 58, at 23.
111. The pivotal work on the differences in search results (for key word searches) among the major legal research services is Nevelow Mart, *supra* note 34. For an important study in differences in headnotes between Westlaw and Lexis, see Peter A. Hook & Kurt R. Mattson, *Surprising Differences: An Empirical Analysis of LexisNexis and West Headnotes in the Written Opinions of the 2009 Supreme Court Term*, 109 Law Libr. J. 557, 2017 Law Libr. J. 27 [Ed. note: while this article was in press, Lexis released Lexis+, which might behave differently than described herein].

eight cases were identified by the first search, of which four were relevant (and none of these were *Beisler* and *Caplin,* two cases that Westlaw ranked most prominently, and which I heavily relied upon in practice and in publication).[112] Furthermore, reversing the terms in the search query results in totally different results with no relevant cases.[113] Based on the initial search in July 2019, the order of search terms seems to matter in a Lexis Advance search, suggesting that the "bag of words" word vector approach may not be in use. However, a subsequent search, in November 2019, produced very different results:[114] more relevant results and key cases were recorded on the initial search; reversing the search terms still produced no results; and finally, scrambling the search terms yielded a few relevant cases, although very different search results from the straightforward search. The conclusion is that word order still matters in Lexis Advance searching,[115] and there is less, if any, reliance on a "bag of words" vector approach. There also appears to be little stability in search results. I have found that even running a search a few days later gives new results (perhaps because the system is reacting to my behavior as a user).[116]

¶60 That reversing the order of the query causes a return of entirely irrelevant results is convincing evidence that "bag of words" and word vectors are not being used, but proximity and word order are important.

Table 12

Lexis Advance

Search Query (federal cases)	Case	Relevant
qualified retirement plan also meeting 105(c) as a health and disability plan	ABA Ret. Funds v. U.S.	N
	Antioch Co. Litig Trust v. Morgan	N
	Kifafi v. Hilton Hotels Ret. Plan	N
	Fr. v. Comm'r	Y
	Wright v. Comm'r	Y
	Chernik v. Comm'r	N
	Berman v. Comm'r	Y
	Minnequa Univ. Club v. Comm'r	N

112. *See supra* note 103.

113. At the time of the original search, because of the lack of relevant results, there seemed to be no point to scrambling the search terms as I did with Westlaw Edge.

114. *See* appendix A. The initial search was done on July 9, 2019. Since the initial search, more relevant results appear in the straightforward search query. *See* appendix A (searched Nov. 13, 2019). Indeed, every time I run the same "scrambled" search, the results change and the precision often improves. *Compare* appendix A and appendix B (searched Nov. 18, 2019).

115. Indeed, techniques like use of Bayesian methods with Hidden Markov Model require ordered inputs or observations to predict outputs or "emissions." Such techniques predict a probable path of emission such as the structure and order of a sentence. *See* discussion of Hidden Markov Models, *supra* ¶¶ 51–52. Of course, it is pure speculation (without any evidence) whether natural language processing in Lexis Advance might employ such methods.

116. *Compare* appendix A and appendix B.

Search Query (backwards)	Case	Relevant
plan disability and health a as 105(c) meeting also plan retirement qualified	U.S. v. Hyppolite	N
	In re Galvin	N
	U.S. v. Hutchins	N
	U.S. v. Hale	N
	U.S. v. Christopher	N
	Pittman v. U.S.	N
	Wilson v. Fed. Mine Safety & Health Review Comm'n	N
	In re Disney ERISA Litig.	N
	SEC v. Revelation Capital Mgmt.	N
	In re CHC Grp.	N

¶61 To get Lexis Advance to improve its relevancy ratings, I had to revert to a terms and connectors search (outside the realm of natural language processing). The search terms were:

("profit-sharing plan" OR "401(k) plan" OR "pension plan") AND 105(c)

What I have done is replace "qualified retirement plan" (something Westlaw Edge was able to parse, probably by comparing vectors of terms to find terms that "qualified retirement plan" included) with specific types of *retirement plans*. I also had to delete references to "health and disability plans."

¶62 Interestingly, in terms of precision,[117] the results of Lexis Advance's *terms and connectors* search were on par with Westlaw Edge's natural language results (9 out of 10 results are relevant, although *Beisler* was missing). However, Lexis Advance's results through terms and connectors are outside the inquiry of this article, which examines the scope of natural language processing and AI. What is really noteworthy, in comparing Lexis Advance and Westlaw Edge, is the ability of West to identify "profit-sharing plan" and other types of retirement plans with "qualified retirement

Table 13

Lexis		
Search Query (terms and connectors)	**Case**	**Relevant**
("profit-sharing plan" OR "401(k) plan" OR "pension plan") AND 105(c)	Berman v. Comm'r	Y
	Gibson v. U.S.	Y
	Gordon v. Comm'r	Y
	Hall v. Comm'r	Y
	Caplin v. U.S.	Y
	Zardo v. Comm'r	Y
	Enloe v. Comm'r	Y
	Rosen v. U.S.	Y
	Cardinale v. S. Homes of Polk Cty.	N
	Christensen v. U.S.	Y

117. Results in the first 10 hits that were relevant.

plan"—the more abstract term for all such plans in general. This may be the result of vector or embedding comparisons among terms.

Bloomberg Law

¶63 Bloomberg Law also has some rather notable results.

Table 14

Bloomberg Law		
Search Query (federal cases)	**Case**	**Relevant**
qualified retirement plan also meeting 105(c) as a health and disability plan	Berman v. Comm'r	Y
	Gordon v. Comm'r	Y
	Wood v. U.S.	Y
	Green v. Comm'r	N
	Castellano v. City of N.Y.	N
	Armstrong v. Comm'r	Y
	Atkins v. Bert Bell/Pete Rozelle NFL Ret. Plan	N
	Watts v. U.S.	Y
	Hannington v. Sun Life & Health Ins. Co.	N
	Lovely-Beyea v. Me. State Ret. Sys.	N
Search Query (backwards)	**Case**	**Relevant**
plan disability and health a as 105(c) meeting also plan retirement qualified	Grose v. Grose	N
	Wood v. U.S.	Y
	Gordon v. Comm'r	Y
	Berman v. Comm'r	Y
	Stewart v. U.S.	N
	Penn Allegh Coal Co. v. Holland	N
	Saunders v. Teamsters Local	N
	Castellano v. City of N.Y.	N
	Marrs v. Motorola Inc.	N
	Atkins v. Bert Bell/Pete Rozelle NFL Player Ret. Plan	N
Search Query (scrambled)	**Case**	**Relevant**
and retirement qualified plan plan also a 105(c) meeting disability as health	Gordon v. Comm'r	Y
	Green v. Comm'r	N
	Wood v. U.S.	Y
	Berman v. Comm'r	Y
	Estate of Hall v. Comm'r	Y
	Barnhorst v. Comm'r	Y
	Berman v. Comm'r	Y
	Christensen v. U.S.	Y
	Maller v. Comm'r	N
	Caplin v. U.S.	Y

For Bloomberg, the best-performing group is when we scramble the search query (8 of 10 results are relevant, including the seminal case, *Caplin*). In comparison, the first search with the query in a straightforward order yields only 5 of 10 as relevant. Reversing the search algorithm gets only 3 of 10 relevant. What is going on here? Only the data scientists at Bloomberg know for sure, but the fact that we can scramble our search terms into a phrase with no syntactical cohesion and still get good (even better) results suggest there must be a role for word vectors and "bag of words."[118] However, there must also be something else affecting results.

Fastcase

¶64 Fastcase's search engine cannot parse *105(c)*, and so the search term was shortened to *105*, which certainly results in more "noise." It also could not produce any relevant hits on the initial, straightforward query.[119]

Table 15

Fastcase

Search Query (federal appellate cases)	Case	Relevant
qualified retirement plan also meeting 105 as a health and disability plan	Ashcroft v. Iqbal	N
	Bell Atl. Corp. v. Twombly	N
	Erikson v. Pardus	N
	Johnson v. U.S.	N
	Lucia v. SEC	N
	Harrington v. Richter	N
	Sessions v. Dimaya	N
	Pearson v. Callahan	N
	Pearson v. Callahan[120]	N
	Molina v. Astrue	N

¶65 We may want to stop here since we didn't get any relevant results, but it might be interesting to determine whether we get the same irrelevant results by reversing the search terms or scrambling. Notably, the reversing of the search terms or the scrambling of them (omitting starting with *and*) produced identical results as the initial query, except the order of results is slightly different for the last three results.

118. The context width, or *k* value, may be quite large since, as suggested above, large context ranges result in topical relatedness rather than semantic similarity. *See supra* note 110 and accompanying text.

119. The test for Fastcase was run on Nov. 7, 2019, a different date than the other databases. This is because during the initial test on July 9, 2019, I did not notice that natural language searching is not the default if entering terms into the "Quick Caselaw Search" bar.

120. Pearson v. Callahan appears twice. Both cites are to 555 U.S. 223 (2009).

Table 16

Fastcase		
Search Query (backwards)	**Case**	**Relevant**
plan disability and health a as 105 meeting also plan retirement qualified	Ashcroft v. Iqbal	N
	Bell Atl. Corp. v. Twombly	N
	Erikson v. Pardus	N
	Lucia v. SEC	N
	Johnson v. U.S.	N
	Harrington v. Richter	N
	Sessions v. Dimaya	N
	Molina v. Astrue	N
	Pearson v. Callahan	N
	Pearson v. Callahan	N
Search Query (scrambled)	**Case**	**Relevant**
retirement qualified plan plan also a 105 meeting disability as health	Ashcroft v. Iqbal	N
	Bell Atl. Corp. v. Twombly	N
	Erikson v. Pardus	N
	Lucia v. SEC	N
	Johnson v. U.S.	N
	Harrington v. Richter	N
	Sessions v. Dimaya	N
	Pearson v. Callahan	N
	Pearson v. Callahan	N
	Molina v. Astrue	N

¶66 While the reader may question the efficacy of disclosing search results with no relevant hits, the point is that the nearly identical results indicate that Fastcase may be using a "bag of words" vector approach to its natural language processing. Word order and proximity do not matter. However, none of Fastcase's results were tax cases,[121] which makes Westlaw Edge's results even more remarkable. Westlaw's algorithm detected, even though there was no reference to tax law or title 26 of the U.S. Code, that the relevant cases were all tax and pension cases through a natural language search. Although not producing as many relevant hits, Bloomberg Law picked up some tax cases. Lexis also located some tax cases, although only on a straightforward natural language query and a terms and connectors query (outside the scope of natural language processing).

121. "According to Fastcase's promotional material, 'natural language searches are much less precise' than Boolean searches, 'but they are a good place to start if you are new to legal research, or if you are delving into a new area of the law.'" Nevelow Mart, *supra* note 34, at 402, ¶ 24.

Casetext

¶67 Casetext produces some interesting results in its straightforward query.[122]

Table 17

Casetext		
Search Query (federal cases)	**Case**	**Relevant**
qualified retirement plan also meeting 105(c) as a health and disability plan	Berman v. Comm'r	Y
	Caplin v. U.S.	Y
	Estate of Barnhorst v. Comm'r	Y
	In re Eagle Food Ctrs.	N
	In re Diet Drugs	N
	Beisler v. Comm'r	Y
	Crouch v. Brase Elec. Contracting Corp.	N
	Rashiel Salem Enters. v. Bunton	N
	In re Phillips	N
	Chevron Corp. v. Barrett	N
Search Query (backwards)	**Case**	**Relevant**
plan disability and health a as 105(c) meeting also plan retirement qualified	Berman v. Comm'r	Y
	Caplin v. U.S.	Y
	Estate of Barnhorst v. Comm'r	Y
	In re Eagle Food Ctrs.	N
	In re Diet Drugs	N
	Gordon v. Comm'r	Y
	Gibson v. U.S.	Y
	Zardo v. Comm'r	Y
	Beisler v. Comm'r	Y
	Watts v. U.S.	Y
Search Query (scrambled)	**Case**	**Relevant**
and retirement qualified plan plan also a 105(c) meeting disability as health	Berman v. Comm'r	Y
	Caplin v. U.S.	Y
	Estate of Barnhorst v. Comm'r	Y
	In re Eagle Food Ctrs.	N
	In re Diet Drugs	N
	Moore v. Raytheon Corp.	N
	In re Ullico Inc. Litig.	N
	Albertson's Inc. v. Comm'r	N
	In re Cook	N
	Findling v. U.S.	N

122. Search done on Nov. 7, 2019. The initial straightforward search on July 9, 2019, produced all irrelevant results, which I did not think to record as a table, and so they are not included. It appears the search algorithm improved between these dates. Search engines are not static, and so rerunning searches can be profitable, though unsettling if you desire consistency.

¶68 However, reversing the terms of the query actually improves the results (with some of the key cases still listed first). Finally, scrambling the search query also produces new results (that are irrelevant), while preserving some key cases at the top of the results list. It is hard to know what is going on in Casetext. The results vary based on search term but, regardless of order, the same three relevant cases appeared in the top three positions in each search, and indeed the top five cases (two of which are irrelevant) were the same in all three searches. This suggests a "bag of words" vector approach with some other elements in the search algorithm. Casetext did seem to get nearer to the issues in its results because they were all about retirement plans, but not all the cases were on point for the specific issue of combining a retirement plan with an accident/health plan.

Ravel

¶69 Ravel did not produce any relevant results in a straightforward query, but when reversed the results were identical.[123] No results appeared in the scrambled version of the query, which had a problem with starting a query with the term *and*. Once that was removed, results were similar but not identical to the prior two queries, although without relevant results.

Table 18

Ravel

Search Query (federal cases)	Case	Relevant
qualified retirement plan also meeting 105(c) as a health and disability plan	Taylor v. Phoenixville Sch. Dist.	N
	Terry v. Bayer Corp.	N
	Regents of the Univ. of Cal. v. Bakke	N
	Gaworski v. ITT Commercial Fin. Corp.	N
	Schwalm v. Guardian Life Ins. of Am.	N
	Smith v. City of Jackson	N
	Bogan v. Holland	N
	Goldstein v. Johnson & Johnson	N
	Funk v. Cigna Grp. Ins.	N
	Tobin v. Liberty Mut. Ins.	N
Search Query (backwards)	**Case**	**Relevant**
plan disability and health a as 105(c) meeting also plan retirement qualified	Taylor v. Phoenixville Sch. Dist.	N
	Terry v. Bayer Corp.	N
	Regents of the Univ. of Cal. v. Bakke	N
	Gaworski v. ITT Commercial Fin. Corp.	N
	Schwalm v. Guardian Life Ins. of Am.	N
	Smith v. City of Jackson	N
	Bogan v. Holland	N
	Goldstein v. Johnson & Johnson	N
	Funk v. Cigna Grp. Ins.	N
	Tobin v. Liberty Mut. Ins.	N

123. Searches performed on Nov. 19, 2019.

Search Query (scrambled)	Case	Relevant
and retirement qualified plan plan also a 105(c) meeting disability as health	Taylor v. Phoenixville Sch. Dist.	N
	Terry v. Bayer Corp.	N
	Bogan v. Holland	N
	Schwalm v. Guardian Life Ins. of Am.	N
	Goldstein v. Johnson & Johnson	N
	Gaworski v. ITT Commercial Fin. Corp.	N
	Regents of the Univ. of Cal. v. Bakke	N
	Smith v. Jackson	N
	Funk v. Cigna Grp. Ins.	N
	Hunt v. Hawthorne Assocs.	N

Because of the similarity in results, Ravel appears to be using a "bag of words" vector process.

Google Scholar

¶70 Google Scholar produced no relevant results with any of the searches or any consistent pattern of results suggesting "bag of words."[124] Citation counts are an important feature of Google Scholar results.[125] So is term frequency.

Table 19

Google Scholar		
Search Query (federal cases)	**Case**	**Relevant**
qualified retirement plan also meeting 105(c) as a health and disability plan	Pilot Life Ins. Co. v. Dedeaux	N
	Firestone Tire & Rubber Co. v. Bruch	N
	Mass. Mut. Life Ins. Co. v. Russell	N
	Metro. Life Ins. Co. v. Mass.	N
	McDonnell Douglas Corp. v. Green	N
	Alexander v. Choate	N
	Shaw v. Delta Airlines	N
	Allis-Chalmers Corp. v. Lueck	N
	Varity Corp. v. Howe	N
	L.A. Dep't of Water & Power v. Manhart	N

124. Search done on Nov. 7, 2019.
125. Nevelow Mart, *supra* note 34, at 405, ¶ 28.

Search Query (backwards)	Case	Relevant
plan disability and health a as 105(c) meeting also plan retirement qualified	Alexander v. Choate	N
	Firestone Tire & Rubber Co. v. Bruch	N
	Metro. Life Ins. Co. v. Glenn	N
	Hendrick Hudson Dist. Bd. of Ed. v. Rowley	N
	San Antonio Indep. Sch. Dist. v. Rodriguez	N
	Metro. Life Ins. Co. v. Mass.	N
	McDonnell Douglas Corp. v. Green	N
	Pilot Life Ins. Co. v. Dedeaux	N
	School Bd. of Nassau Cty. v. Arline	N
	Goldberg v. Kelly	N
Search Query (scrambled)	**Case**	**Relevant**
and retirement qualified plan plan also a 105(c) meeting disability as health	Varity Corp. v. Howe	N
	Pilot Life Ins. Co. v. Dedeaux	N
	Metro Life Ins. Co. v. Mass.	N
	Firestone Tire & Rubber Co. v. Bruch	N
	Alexander v. Choate	N
	Metro. Life Ins. Co. v. Glenn	N
	McDonnell Douglas Corp. v. Green	N
	Goldberg v. Kelly	N
	St. Mary's Honor Ctr. v. Hicks	N
	Mathews v. Eldridge	N

¶71 Google appears not to have a "bag of words" approach. It differs markedly from other legal search tools, a fact that should be stressed with students.

Caution and Summary

¶72 The reader should be cautioned about generally discounting the merits of some of these search engines based on the foregoing. Other techniques or different queries may produce more relevant results than in this illustration, which is not a proper study. Obviously, many more test searches would need to be run to determine definitively what is going on with each major vendor, but the examples above reveal stark differences. It appears that Westlaw Edge is the most prone to use the techniques based on "bag of word" vectors we have described herein, followed by Casetext. For Lexis Advance, word order seems to matter, suggesting that it uses other algorithms than have been discussed in this section, and indeed a *terms and connectors* search ultimately produced comparable results to Westlaw. Bloomberg also produced a few similar results among the three searches. Ironically, the best search for Bloomberg was when we scrambled the terms in the query. Fastcase produced similar results between searches, perhaps suggesting "bag of words," but because it could not parse *105(c)*, none of those similar results were relevant. Ravel produced no relevant results, but identical results for forwards and backwards queries, and similar results for a scrambled study. For Google Scholar, no conclusions can be drawn because of irrelevant results, at least for the straightforward test query. Reversing and scrambling the results also did not yield similar results that would have suggested a "bag of words" vector approach. It is important to stress that Google Scholar appears to operate differently from other legal search engines. Google and Lexis Advance differ most notably from the other search engines that seem to rely on a "bag of words" approach.

¶73 Besides the issue of natural language processing, there is the issue of my frustration at rerunning searches on the same service on different dates (even sometimes a few hours later) only to get different results. I have documented one instance of that with Lexis Advance in appendixes A and B (especially when compared with Table 12). Similar discrepancies are noted with Casetext.[126] There were also things I didn't document, such as Lexis Ravel graphs radically changing between iterations of running the same search.[127] This is another affirmation about the instability of our search systems.[128] How can the legal profession bestow cognitive authority when not only are there such wide variations between systems, but the results of any search on any given day can change? It may be that we simply don't have another option. Finally, I must note that for all their sophistication, existing search technologies do not answer the question of whether retirement and disability plans can be successfully combined. The answer to that actually requires human reading and analysis of the relevant cases—so much for AI solving our research problems without us.

¶74 The question of whether iterative searches on a single service or subset of services will replicate the results of searches on all the search services, or at least find a stable basis of precedent, is one that needs to be and can be studied.[129]

Word Prediction Techniques and Secondary Sources

¶75 Skip-grams and Bayesian algorithms represent only some of the many tools that data scientists use. One variant of the application of such tools is used to predict word searches—which is not only a feature of Lexis, Westlaw, and Bloomberg, but is familiar on Google searches and typing on text messaging.[130] However, Lexis and Westlaw have taken word prediction to a new level, by predicting not only search terms or phrases but also applicable questions, documents, and general sources. In figure 3 is a response from searching in Lexis's search bar to the single term *copyright*.

¶76 What is interesting is how Lexis ties in secondary sources as well as search terms and *Suggested Questions*. Lexis's strength is in rich secondary materials. However, the results are somewhat troubling about why the particular volume and sections of *Nimmer on Copyright* are selected with so little input from the user, other

---

126. *See supra* note 122.

127. *See infra* figure 9 for example of Lexis Ravel graphic.

128. The Nevelow Mart study confirms this by showing that approximately 40% of results in each service are unique. *See supra* notes 34 through 37 and accompanying text.

129. *See supra* note 43.

130. A typical tool used would be *Interactive Query Refinement*, a variant of *Automatic Query Expansion*. *See* Claudio Carpineto & Giovanni Romano, *A Survey of Automatic Query Expansion in Information Retrieval*, 44 ACM Computing Surveys, Jan. 2012, at 1, 8. *Google Suggest* is an example of such a feature. *Id.*

Figure 3

Lexis Search Bar with Predictive Response to Search Term

(Reprinted from LexisNexis with permission. Copyright 2020 LexisNexis. All rights reserved.)

copyright

Sources	Nimmer on **Copyright**
	Nimmer on **Copyright** - Index
	ALI CLE Course of Study - Trademarks, **Copyrights** & Unfair Competition
	A Practical Guide to **Copyright** Law in the Digital Age (MCLE)
	Associate's Guide to the Practice of **Copyright** Law
Documents	**copyright** act
	4 Nimmer on **Copyright** § 13.05
	4 Nimmer on **Copyright** § 13.03
	4 Nimmer on **Copyright** § 13.01
	2 Nimmer on **Copyright** § 7.16
Legal Phrases	**copyright**
	copyright infringement
	copyright law
	copyright protection
	copyright owner
Suggested Questions	What are the elements of **copyright** infringement?
	What is the burden of proof for **copyright** infringement?
	What is the statute of limitations for **copyright** infringement?
	What is the definition of **copyright** owner?
	What is the definition of common-law **copyright**?

Figure 4

Lexis Search Bar with Predictive Response to "copyright fair use transformative"

(Reprinted from LexisNexis with permission. Copyright 2020 LexisNexis. All rights reserved.)

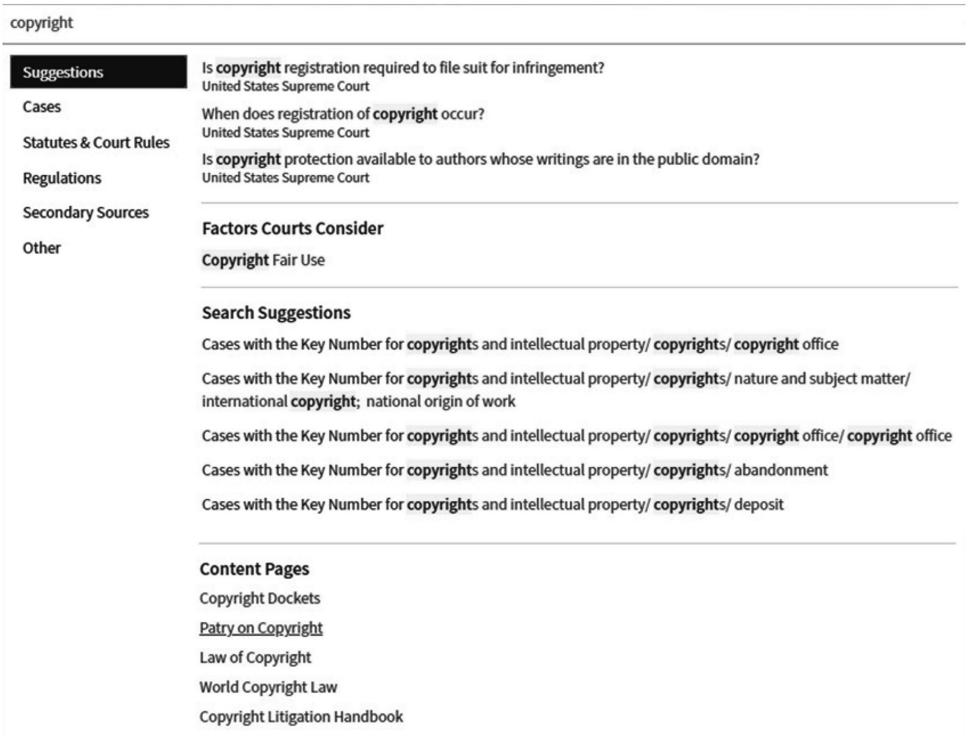
copyright fair use transformative

Legal Phrases	fair use and transformative
	copyright fair use

Figure 5

Westlaw Search Bar with Predictive Response to Search Term

(Reprinted with permission of Thomson Reuters)



than the single term, *copyright*. There is simply going to be a lot of noise or irrelevant material in our results.

¶77 When I add more terms, "copyright fair use transformative," the results are narrower, but they now are limited to *Legal Phrases* (see figure 4).

¶78 If we select "fair use and transformative," it leads us to a lot of law review articles. By filtering, for *Secondary Material*, we can get back to *Nimmer on Copyright,* in this instance to volume 4 and § 13.05, which has some quite persuasive material on the matter. Therein lies the problem. The process fails to emphasize that some material is more "cognitively" authoritative (or persuasive as recognized by the profession) and relevant than others. The user has to discern through filtering successfully where that material might lie.

¶79 Lexis should be applauded. It has taken steps, using predictive techniques, to try to keep secondary material as part of the search inquiry, even in the age of the single-search box and natural language processing, but the tool needs refinement with more complex search inquiries. Nonetheless, perhaps there is hope for secondary material after all. Of course, this is in Lexis's interest. Secondary materials is where it really shines. The only problem is that indexes (and human intermediation) get circumvented in the process. Appealing to an index will always be last, if thought of at all.

Figure 6

Westlaw Search Bar with Secondary Sources Resulting from Predictive Response

(Reprinted with permission of Thomson Reuters)

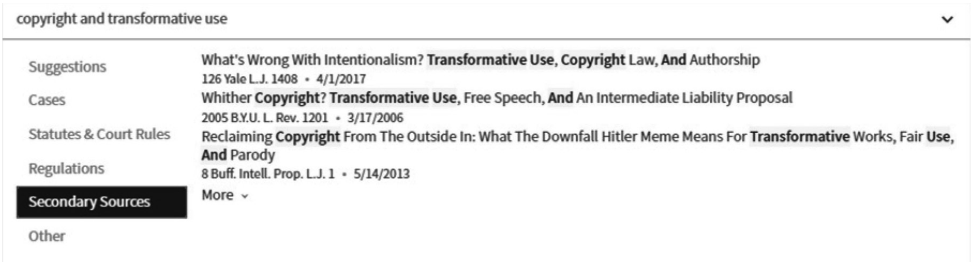
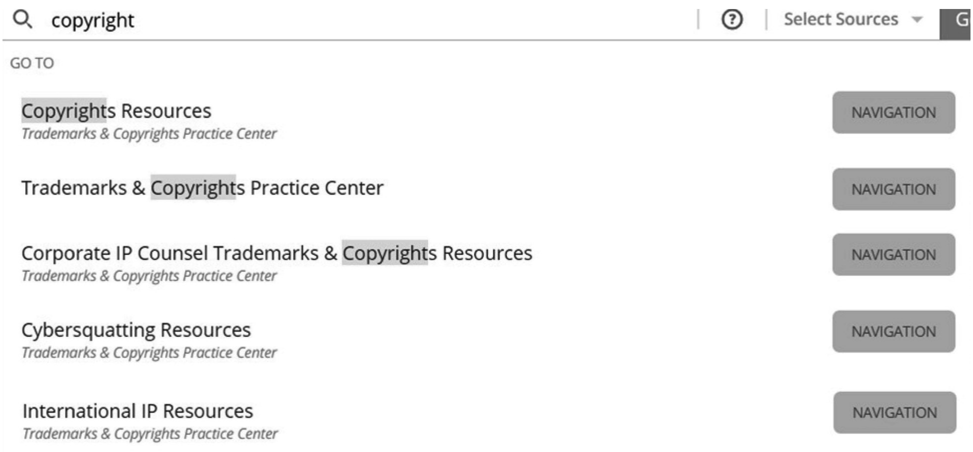


Figure 7

Bloomberg Search Bar with Predictive Searching

(Reproduced with permission from Bloomberg Law. Copyright 2020 by The Bureau of National Affairs, Inc. (800-372-1033) http://www.bna.com.)



¶80 Advancement of these technologies is critical, especially if indexes are going unused as a method of access and are being replaced by the single-search box. As I mention in the introduction, many, if not most, questions are *subject* in nature and need different tools, or at least methods, than in a *known-item* search (which favor online search queries).

¶81 Not to be outdone, Westlaw Edge also uses predictive searching to suggest a variety of material and searches (see figure 5).

¶82 Furthermore, adding words to Westlaw Edge's search bar actually yields better results for searches, in this instance "copyright and transformative use" (see figure 6).

¶83 The algorithms for predictive searching may work on the same principles as the prediction of context words with word2vec's Skip-gram with Negative Sampling.[131] Probabilities are calculated using Bayesian techniques (which provides for

131. *See supra* ¶¶ 41–44.

training by humans), and the leading contenders are set forth.[132] The use of secondary sources (in this case, law review articles) is a step toward preserving the role of secondary materials. What would really be great is the appearance of *American Law Reports* or a treatise.

¶84 Rather than suggesting particular resources, Bloomberg will suggest *Practice Centers* based on predictive search terms (see figure 7).

¶85 More specific terms, like "copyright transformative use," do not produce suggestions. What is important to note is that the whole process bypasses indexes and tables of contents—human-intermediated access tools. It is a shift in how cognitive authority is accessed.

Teaching Search Inquiry in the Age of Natural Language Processing

¶86 Teaching natural language processing may be overlooked, but it is necessary to make students aware of the diversity of results between services, some of the shortcomings of the services, and things they can do to improve results.

Differences in Results

¶87 Because of the manifold variables in natural language processing, it is imperative that students be aware of the differences that various services are likely to provide. Fortunately, Nevelow Mart has provided a ground-breaking study that ought to be read by students to facilitate their understanding of the different and unique results that the various services provide.[133] Given the current popularity of word2vec (and chances of its use in legal databases), the importance of the range of context words (or the *k*) variable in capturing results ought to be stressed as affecting outcomes. Shorter ranges yield terms that are similar to the target term, and larger ranges are more inclined to pick up the broader topic.[134] Unfortunately, we do not know what search engines are utilizing. It is possible that longer search inquiries may take better advantage of search algorithms with larger context ranges (as in my examples of the dual-purpose health-disability qualified retirement plans), but this is still speculation (and the opposite occurs with word prediction in the previous section). In my example of case law research, Westlaw Edge performed particularly well in this situation. Furthermore, it seemed to be able to parse a more general term, "qualified retirement plans," for the many other particular retirement plans that would fall within this umbrella. This is probably because of the use of word vectors or embeddings, which were in evidence by the reversing and then scrambling of the search terms. It even identified the problem as a tax problem without the word *tax* or complete citations to tax code sections appearing in the search query. For Lexis Advance, we had to revert to terms and connectors (and the use of terms within the umbrella of "qualified retirement plans") to get adequate performance. Rerunning the search months later also improved results, and indeed running searches again even after a few days provided new results.[135] This does not produce confidence in stability of results, but perhaps this is a method of searching

---

132. *See supra* ¶¶ 48–50.

133. *See* Nevelow Mart, *supra* note 34.

134. *See* Jurafsky & Martin, *supra* note 110, and accompanying text.

135. *Compare* Lexis "scrambled" searches in appendixes A and B. *Compare also* original Lexis search in Table 12 with appendixes A and B.

that should be taught. "If you don't like what you get, rerun the exact same search again."

¶88 Certain natural language processing activities, such as classification, are useful in determining negative and positive citations to a work, and may use Naive Bayes classification. What is important to understand is that there are human trainers that help *teach* the algorithms correct classification. Thus, the human element is not totally divorced from natural language processing. The reputation of services like Thomson Reuters Westlaw and LexisNexis for employing human editors is thus a real benefit. Do services like Bloomberg, Fastcase, Casetext, Ravel, and even Google have legal experts helping train their systems? There is an alternative to in-house legal experts, and that is to track users' (especially attorneys') responses to make search results more precise. Without humans, the different search tools may be limited in what they can accomplish. The important thing is to recognize the differences.

Lengthening Search Inquiries

¶89 It is uncertain how long and detailed a query should be. In some circumstances, for natural language processing, having a rich and lengthy search query can greatly improve performance (in terms of precision or recall).[136]

> When a user query contains multiple topic-specific keywords that accurately describe his information need, the system is likely to return good matches; however, given that user queries are usually short and that the natural language is inherently ambiguous, this simple retrieval model is in general prone to errors and omissions.[137]

For Google the average search query is 2.3 words.[138] Maybe this is why search tools try to lengthen searches with predictive terms and automatic query expansion. Students should be challenged to lengthen their searches. It may also be helpful to repeat key terms to give them more weight.[139] As an alternative to repeated terms, Westlaw has introduced the "^" as a way to emphasize terms.[140]

Recognizing Biases of Search Algorithms

¶90 Nevelow Mart lists the following necessary attributes of search algorithms that may bias results and contribute to algorithmic opacity:

- prioritization ("emphasiz[ing] . . . certain things at the expense of others"; like relevance ranking);
- classification (putting an "entity [in a] constituent . . . class"; data training may import human biases);
- association ("marks relationships between entities"); and
- filtering, which "includes or excludes information according to various rules or criteria."[141]

---

136. *See supra* ¶¶ 53–74. Tables 11–19 illustrate a fairly detailed search phrase.

137. Carpineto & Romano, *supra* note 130, at 1.

138. *Id.*

139. *See supra* note 69.

140. Barco Law Library, *Search Term Emphasis on Westlaw*, Barco 3.0: Law Library Reference, Oct. 8, 2019, http://barcorefblog.blogspot.com/2019/10/search-term-emphasis-on-westlaw.html (last visited Aug. 5, 2020). Simply place the ^ at the tail end of the word.

141. Nevelow Mart, *supra* note 34, at 394–95, ¶ 12 (citing Nicholas Diakopoulos, *Algorithmic Accountability: Journalistic Investigation of Computational Power Structures*, 3 Dig. Journalism 398, 399 (2015)).

¶91 Nevelow Mart calls for disclosing how these factors are used for "algorithmic accountability."[142] Helping students see these categories of potential bias is something worthy of classroom discussion.

Recognizing Out-of-Context Matches

¶92 There are a variety of problems with natural language search inquiries that students would do well to recognize:

- Polysemy (or the multiple meanings of words, such as "organization");
- Relationships based on order ("man bites dog," rather than "dog bites man");
- Out-of-phrase terms (failure to treat a phrase as a single unit, such as "North Atlantic Treaty Organization");
- Secondary topic key words ("Roth IRA" versus "IRA");
- Noncategorical terms ("fair use" is an instance of an exception under both copyright law and trademark law).[143]

¶93 To this list, I would add my own frustration in doing online search queries of the codes of all 50 states on Lexis Advance and Westlaw Edge to determine the states that had repealed or otherwise done away with the "rule of perpetuities." All my results affirmed the rule in state after state. I conclude that besides being difficult to do natural language searching in codes, it is difficult to prove a negative (or a repealed statute). Only after locating a secondary source on the topic was I able to start to find states that had done away with the rule.

Secondary Sources and Codes

¶94 The problem is that many searches should start with secondary sources (or codes),[144] but the interfaces of the major vendors place them in a secondary position. Lexis Advance and Westlaw Edge (and to some extent Bloomberg Law) are to be applauded by suggesting the user types in the search box sources that could be used. But these services are still primitive: the more you type in the search box, the less you get. Compared to a good index in a secondary source, the insufficiency is noteworthy.

¶95 Most of the time, to get to secondary sources, users have to search and then select filters that bring secondary sources to the forefront. This totally bypasses any use of human-intermediated indexes. Students need to be instructed both to get into the habit of "drilling down" to secondary sources and, in many instances, to search out the index for initial queries. And then there is the question of what to do about statutory codes. My own experience in natural language searching on codes has limited utility. Indexes and tables of contents are still key and need to be taught at every opportunity with students.

Reiterative Searching

¶96 Often, it is not the initial search query that produces the best results, but what the researcher does with those results that ultimately finds the best cases or authority. This is illustrated in figure 8.

---

142. *See* Nevelow Mart, *supra* note 34, at 395, ¶ 14.
143. Carpineto & Romano, *supra* note 130, at 5.
144. *See* CALLISTER, *supra* note 28, at 24–34.

Figure 8

The Legal Research Cycle

Source: Paul D. Callister, Field Guide to Legal Research 123 (West 2019).

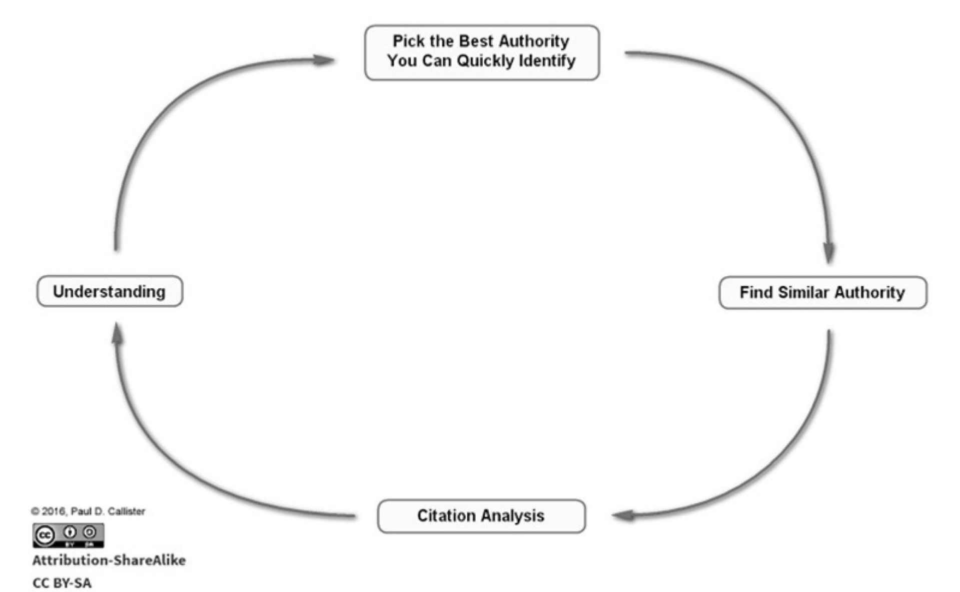


Figure 9

Lexis Advance Ravel View of Search for “Fair Use and Transformative”

(Reprinted from LexisNexis with permission. Copyright 2020 LexisNexis. All rights reserved.)

¶97 Teaching students to find similar authority is thus vital when initial search queries produce such varying results. There are a variety of techniques for finding similar authority. Westlaw Edge or Lexis Advance headnotes can lead to similar points of law, if the particular point in question has been assigned to a unique headnote. As with search queries, the differences between headnotes is significant between the two vendors, and so using them is more art than science.[145] Users may also search headnotes and particular terms at the same time, yielding the greatest opportunity for stumbling on similar cases. Citation analysis can also lead to more relevant cases. Both the cases cited in a case (from the table of authorities), and the cases that cite a relevant case can lead to stronger authority. Finally, understanding key cases is important. Consequently, tracking commentary (like an *American Law Reports* annotation or an encyclopedia, such as *California Jurisprudence*) that has cited a case may uncover better authority—the ones practitioners rely on.

¶98 One technique for reiterative searching is to follow up a natural language search with a terms and connectors search, just to compare. A particularly useful tool is having identified a key term or phrase to search within results using an "atleastN" command. But the use of commands takes us out of natural language processing and this article's quest to explore how responsive such systems are to legal parlance.

Search Result Visualizations

¶99 Data visualizations, like those provided by Lexis and Fastcase, often can be used to winnow key authority to seminal cases on an issue. Figure 9 shows my search of federal case law for cases with "fair use" and "transformative." While the first case that appears is *Fox News Network v. TVEyes*, I can readily discern by rolling over the large circle in the Supreme Court segment of the *Ravel View* chart that *Campbell v. Acuff-Rose Music* is seminal to citations about fair use and transformative use. Thus, while my natural language search may not have produced the most relevant or seminal cases first, the Lexis Advance *Ravel View* visualization function quickly got me the information I needed.[146]

¶100 Likewise, with Fastcase, a terms and connectors search for "copyright and 'fair use' and transformative"[147] yields *Fox News Network v. TVEyes* as the first hit, but when I switch to interactive timeline, with the vertical axis set to court level, I can readily see that *Campbell v. Acuff-Rose Music* is the key Supreme Court precedent, heavily cited, on the issue (see figure 10). Interestingly, Westlaw Edge (lacking the visualization tools) in a natural language search of "fair use transformative" produced *Campbell v. Acuff-Rose Music* as its third entry (neither Fastcase nor Lexis produced the case in top results). The point is the means to get to the seminal case, preferably a Supreme Court case, varies from service to service, and this needs to be taught to students.

¶101 Thus, visualizations can be a powerful tool in discerning the most relevant authority on an issue after conducting a primary search.

Conclusion

¶102 By undertaking this article, I assumed a large risk—chiefly because I don't know how the major legal search engines employ natural language processing and

145. *See* Hook & Mattson, *supra* note 111 (study on difference of headnotes between Westlaw and Lexis).

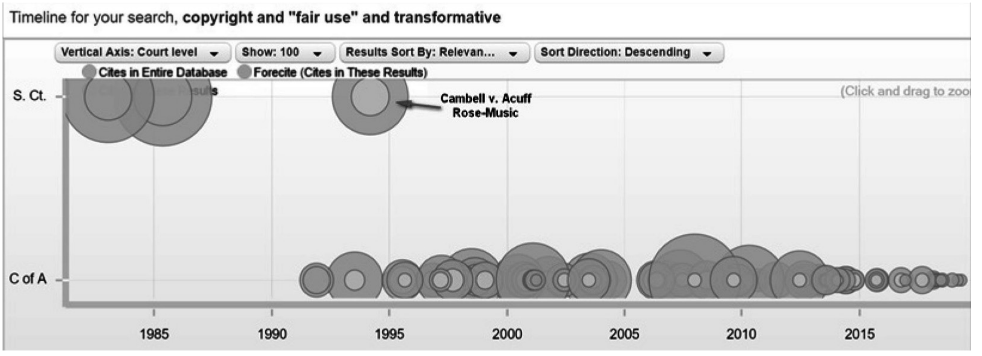
146. Rerunning a search at a later date or time may produce different results and a different graphic.

147. A natural language search of the same terms yielded too many non-copyright cases.

Figure 10

Fastcase Timeline for Search of "Copyright and 'Fair Use' and Transformative"

(Reprinted with permission of Fastcase)



AI.[148] I have made some educated guesses based on the literature about natural language processing in hopes of gleaning some understanding of what may be going on with searches across the various vendors. I have even tried to find evidence for tools, such as "bag of words," by reversing and scrambling search inquiries to see what results were produced.[149] Besides search tools (and comparison of documents for similarity), I have introduced Bayesian classifiers that may be used for a variety of tasks, such as detecting when cases are distinguished or not followed (or classification of headnotes by topic or Key Numbers).[150] Are any of these tools used? I have no way of knowing, other than statements by the major services that they use AI.[151]

¶103 Vendor promotional material proclaiming, *You Ask a Question . . . Lexis Answers Understands It*,[152] is marketing research services that are intelligent. However, as we understand more about natural language processing, some of the "magic" goes away, and we see algorithms at work. If the major vendors are using neural networks[153] to find through regression optimum weights for vectors or embeddings, this may be one step closer to AI (and is a part of *machine learning*).[154] But this really is the application of the brute force of computing power. The problem of such applications of "brute force" is that they command vast computing resources and expense, perhaps making themselves available to only the most prosperous sector of the legal profession.[155]

148. *See supra* note 8.
149. *See supra* ¶¶ 53–74.
150. *See supra* ¶¶ 48–50.
151. *See supra* notes 8, 30, 47.
152. *See* LexisNexis, *supra* note 30.
153. See *supra* ¶ 47.
154. Jason Brownlee, *Logistic Regression for Machine Learning*, MACHINE LEARNING MASTERY (Apr. 1, 2016), https://machinelearningmastery.com/logistic-regression-for-machine-learning/ (last visited Aug. 5, 2020) ("Logistic regression is another technique borrowed by machine learning from the field of statistics."); WIKIPEDIA, *supra* note 85 ("Machine leaning (ML) is the scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead.").
155. *See* JONES, KALANTERY & GLOVER, *supra* note 4, at 18 ("Unless the cost of compute

¶104 The larger concerns recited in the beginning of this article were stability of the law, opacity, and deemphasizing or even replacing secondary authority (the great treatises), thereby affecting cognitive authority.[156] Stability of law is a concern because of inconsistent results and the corresponding fact that very little of what attorneys are citing in briefs appears to make it into court decisions.[157] Inconsistent results among services is compounded by the fact that most attorneys may not have access to all the services, nor can clients pay for such prodigious use of research time. However, assuming access to the right research services, there are techniques to take initial results and find seminal cases through headnotes or visualization.[158] But the first round of searching can produce widely different results. Following up with additional steps to the first round of results can improve performance.

¶105 Is opacity among search engines a threat? We have turned over the stability and structure of law to data scientists. However, as the author of *Weapons of Math Destruction* points out, recent history is rife with examples of unintended and harmful consequences from algorithms.[159] The lack of stable responses to search queries is a threat, but there may be others, such as biases yet unseen. Nevelow Mart has suggested that the case classification systems used by West and Lexis are *possibly* biased for being based on the 19th century Langdellian conceptions of the law.[160] Nevelow Mart even speculates about the possible bias (or at least the differences in results) "from the very different list of secondary sources in Westlaw and Lexis Advance [that] are baked into their respective search results . . . ."[161] Secondary sources may well be part of the algorithmic equation when searching case law in the two dominant services.

¶106 Finally, there is the question about cognitive authority and how it might change. To answer the question, I have introduced Deibert's holistic ecological model of media theory that eschews technical determinism.[162] Within its rings, language reigns supreme as a technology through time. That humans are driven to have machines process and even understand language, including in such technical fields as law, is inherent in our nature. The introduction of natural language processing is already affecting law's cognitive authority, but its historical parameters (like 19th century classification systems) will also affect natural language processing and its use. At least, that is the prediction of Deibert's model. "Langdellian" conceptions may survive the migration to new cognitive authority.

¶107 In the introduction, I was also concerned about the demotion of secondary sources because a single-search box tends to default to case law, and finding secondary material requires filtering. This is somewhat alleviated by the tools of Lexis Advance and Westlaw Edge that suggest secondary materials as part of the predictive searching,[163] but these services are still primitive, and when more text is entered

---

drastically decreases, experiments will grow too large to be affordable by anyone but the US or Chinese governments."). *Compute* is the correlation between the amount of computing power used to train AI and the power necessary for the resulting AI model. *Id.*

156. *See supra* ¶¶ 75–85.
157. *See* Bennardo & Chew, *supra* note 35.
158. *See supra* ¶¶ 96–101.
159. *See* O'Neil, *supra* note 2.
160. *See supra* note 34.
161. Nevelow Mart, *supra* note 34, at 419, ¶ 55.
162. *See supra* notes 14 through 16 and accompanying text.
163. *See supra* ¶¶ 75–85.

into the query, secondary material seems to disappear, leaving the user to rely on post-search filtering. Finally, the indexes and tables of contents are lost in the single-search box world. That connection to human-intermediated information is disappearing.[164]

¶108 There are a number of reasons why we cannot yet rely upon legal search engines like we do with Google or Amazon Alexa, at least with respect to natural language processing. The legal search services are turning to the single-search bar, but their natural language algorithms appear to be based on a "bag of words," with scarce evidence of syntax playing any significant role.[165] For the most part, by treating words as vectors, we are relying on proximity—whether terms in the same document or within a "context window"—to accord with meaning. I believe legal discourse to be too subtle to be boiled down to proximity of words, regardless of the method of doing so. Perhaps more important, the role of secondary authority (the great treatises) is often subjugated in search processes, especially with respect to access through indexes and tables of contents.[166] Indeed, the use of the latter is dying. Many up and coming legal research services do not even have secondary materials. Add to that the difficulty searching legal codes. Often, commentary and codes should be the starting point of legal research rather than an afterthought.

¶109 In conclusion, the best we librarians can do in the face of uncertainty is to teach our users about the limitations of these systems, disillusioning them of computer intelligence doing the work for them—at least for now. If anything, AI is a tool and, one day perhaps—assuming a humanistic techno-central vision—a partner. Perhaps the day will soon come when law firms will list IBM's Watson as a partner, and we will be able to ask questions of legal search engines like we do with Google and Alexa. But that day is still a ways off. Whatever may happen, the profession's shared cognitive authority is shifting to the algorithm—it is in it that we will entrust our future.

---

164. *See id*. Hopefully West's reliance on the human-intermediated Key Number System will counter this trend. *See supra* notes 47–48 and accompanying text.

165. There is evidence that sometimes word order does matter in some of the searches conducted, *see supra* ¶¶ 53–74, because not every search engine produced identical results regardless of search order.

166. *See supra* ¶¶ 75–85.

Appendix A

Lexis Advance		
Search Query (federal cases)	**Case**	**Relevant**
qualified retirement plan also meeting 105(c) as a health and disability plan[167]	Gibson v. U.S.	Y
	Gordon v. Comm'r	Y
	Caplin v. U.S.	Y
	Wright v. Comm'r	Y
	Paul v. U.S.	Y
	Berman v. Comm'r	Y
	Fr. v. Comm'r	Y
	Chernik v. Comm'r	N
	ABA Ret. Funds v. U.S.	N
	Enloe v. Comm'r	Y
Search Query (backwards)	**Case**	**Relevant**
plan disability and health a as 105(c) meeting also plan retirement qualified	No cases	
Search Query (scrambled)	**Case**	**Relevant**
and retirement qualified plan plan also a 105(c) meeting disability as health	In re Disney ERISA Litig.	N
	In re Alpha Natural Res.	N
	U.S. v. Martinez	N
	ABA Ret. Funds v. U.S.	N
	Antioch Co. v. Morgan	N
	Zardo v. Comm'r	Y
	Gentile v. Comm'r	Y
	Kifafi v. Hilton Hotels Ret. Plan	N
	Thomas v. Comm'r	Y
	Thomas v. Comm'r[168]	Y

167. Search run Nov. 13, 2019.

168. The results are both to the United States Tax Court Opinion, T.C. Summary Opinion 2007-110 (Jun. 28, 2007).

Appendix B

Lexis Advance

Search Query (federal cases)	Case	Relevant
qualified retirement plan also meeting 105(c) as a health and disability plan[169]	Gibson v. U.S.	Y
	Gordon v. Comm’r	Y
	Caplin v. U.S.	Y
	Wright v. Comm’r	Y
	Paul v. U.S.	Y
	Berman v. Comm’r	Y
	Fr. v. Comm’r	Y
	Chernik v. Comm’r	N
	ABA Ret. Funds v. U.S.	N
	Enloe v. Comm’r	Y
Search Query (backwards)	**Case**	**Relevant**
plan disability and health a as 105(c) meeting also plan retirement qualified	No cases	
Search Query (scrambled)	**Case**	**Relevant**
and retirement qualified plan plan also a 105(c) meeting disability as health	Gordon v. Comm’r	Y
	Caplin v. U.S.	Y
	Berman v. Comm’r (6th Cir.)	Y
	Gibson v. U.S.	Y
	Berman v. Comm’r (Tax Ct.)	Y
	Hall v. Comm’r	Y
	West v. Comm’r	Y
	Kelter v. Comm’r	Y
	Dorroh v. Comm’r	Y
	Armstrong v. Comm’r	Y

169. Search run Nov. 18, 2019. Even identical searches run the same day may have different results on Lexis Advance.