Haave™

SentenceReader

Sentence Reader: 17 Times More Power for E-Discovery Search

Every Litigant's Worst Nightmare: Missing the Smoking Gun

Contemporary litigants and their attorneys face a vexing dilemma. On one hand, the ubiquity of Electronically Stored Information (ESI) (predominantly e-mails) has dramatically expanded the volume of documents that must be collected, exchanged, and reviewed in litigation. On the other hand, the escalating costs of legal services can make it prohibitively expensive to pay large teams of attorneys (especially U.S.-based attorneys) to manually review all potentially relevant documents. As a result, the litigator's worst nightmare—being blindsided by a "smoking gun" document that one's own document review team failed to identity—becomes an ever more likely scenario.

In a thorough investigation of this dilemma, the Rand Corporation has correctly identified technological innovation as the most promising solution.[1] Since our mountains of ESI are themselves the product of technological progress, litigators require comparable innovations to help make sense of the massive document sets in their possession. Indeed, if one accepts the basic premises that (1) the volume of ESI will only continue to grow and (2) the hourly rates charged by attorneys can only be constrained so much (typically by using contract or overseas attorneys), then the *only* viable solution becomes a better search engine—one that enables human reviewers to perform their work more efficiently and effectively.

The existing e-discovery products fail because they are simply not designed to tackle the most complex, time-consuming, and expensive task facing litigators: searching for the handful of documents that ultimately matter—the potential litigation exhibits—amidst the haystacks of documents that will never become exhibits. Though tens of thousands if not millions of documents are typically exchanged in each case, the fraction of the exchanged documents that will ultimately become exhibits in depositions or at trial is, by necessity, orders of magnitude smaller. If litigants are lucky and their front-line reviewers are exceptionally well trained and exceptionally diligent, then the tiny fraction of genuinely "hot" documents will all have been accurately identified and successfully passed up the chain of command for careful evaluation and strategic planning.

---

1 Rand Institute for Civil Justice, Where the Money Goes: Understanding Litigant Expenditures for Producing Electronic Discovery. https://www.rand.org/content/dam/rand/pubs/monographs/2012/RAND_MG1208.pdf

Unfortunately, in the real world, the review process rarely works that smoothly, leaving decision makers to constantly worry that crucial pieces of evidence have been overlooked—perhaps resurfacing to ruinous effect in the middle of a deposition or trial. To avoid such costly errors, contemporary litigators need a more effective search tool, one that enables them to quickly identify the documents that ultimately matter.

Sentence Reader, unlike its competitors, has been specifically engineered to accomplish precisely that objective. As detailed below, our experiments prove that Sentence Reader is 17 times better than the industry standard, simultaneously increasing the quality and decreasing the costs of document review. By providing human reviewers with an intuitive and hassle-free, yet genuinely state-of-the-art search engine, Sentence Reader permits lawyers to focus on actually practicing law, instead of drowning in documents and hoping that no unhappy surprises await.

Why Semantic Search is Crucial in E-Discovery

Sentence Reader becomes especially important in large and complex matters, which rarely unfold a predictable or linear fashion. As complex lawsuits progress, the pivotal legal and factual issues frequently shift and evolve—often in unexpected directions.   Indeed, as discovery progresses and as new evidence emerges, new claims and new parties often get added, and formal complaints must be amended.

Establishing the precise nature of the legal claims and parties at issue, moreover, is only the beginning the litigants' wild ride. As the parties exchange documents and as depositions proceed—often in a piecemeal, sporadic manner—litigants must constantly adapt their strategies and agendas to new disclosures and testimony.   In such a turbulent environment, issues that once seemed trivial can suddenly assume unexpected importance, forcing parties to scramble and reassess previously reviewed documents with new questions in mind. In short, litigation requires real-time adaptation to an ever-changing set of facts, questions, problems, and strategic considerations. Keeping tabs on the known facts, the known unknowns, and the unknown unknowns can make for many sleepless nights.

The existing e-Discovery technologies and workflows only make matters worse. Even when the set of documents at issue remains relatively small and associates might, if necessary, be able to click through every single document when a new question presents itself, repeated linear reviews remain too time-consuming and expensive to be feasible. As a result, with every new twist and turn during the course of discovery, the dilemma described above resurfaces: Though clients are understandably reluctant to pay for attorneys to perform multiple linear passes through their entire document set, outside counsel remains justifiably concerned that their associates' previous reviews may not have attended to issues and problems that have only recently assumed center stage. So how are litigants and their clients to balance these competing demands of efficiency, on the one hand, and thoroughness, on the other?

The answer, again, can be found only in a more powerful search tool—a tool that can reduce reliance on linear document reviews and on traditional, linear workflows more generally. Without a search engine that is able to meet their complex and ever-changing informational needs, litigants will always worry that something crucial may have slipped through the cracks: "When we made our initial pass through the documents, we were concerned about questions A, B, and C, but now that D and E have emerged as crucial points of contention, how can we be sure that we have not overlooked anything without re-reading all of our documents again with an eye toward D and E?" The answer is that you cannot be sure unless you have access to the proven power of Sentence Reader, which directs you to the specific sentences in which you may be interested at any given point in time.

To be sure, the challenge of finding the precise sentences that might make or break your case requires a combination of smart lawyering and smart technology: if either half falters, the specter of missing a "smoking gun" document returns. But unless a search tool is able to "see" beneath the mere words contained in an attorney's query and grasp the deeper, underlying concepts and ideas that a query was intended to capture, the sheer complexity of the English language will defeat the attorneys' best efforts. Language affords too many different ways to express the same idea for a keyword search to prove reliable. Without the ability to move beyond keywords to a genuinely semantic, conceptual search, even the most ingenious, resourceful, and determined attorney cannot hope to find what he or she needs without clicking through every single document every single time a new question arises—which can happen many times in the course of a single lawsuit.

Though no search tool can directly read a user's mind, the sheer depth and breadth of Sentence Reader's knowledge does enable it to tease out subtler connections and associations than most users would ever think to query themselves. In that sense, using Sentence Reader is like having a superhuman subject matter expert standing by your side, leveraging its superhuman expertise to either guide you toward your intended target (when it exists) or to reassure that it does not exist (when in fact it does not).

Being able to confirm that a given document does not exist, moreover, is often no less valuable than laying hands on what actually exists. And unless you have confidence that your own search technology does, in fact, represent the current state of the art, you will always be left to wonder whether your opposing counsel's technology is enabling them to find crucial documents that your inferior technology has missed.

Why Sentence Reader is Different: The Importance of Literacy

Exactly what makes Sentence Reader so much better than the competition? Two fundamental and far-reaching innovations:

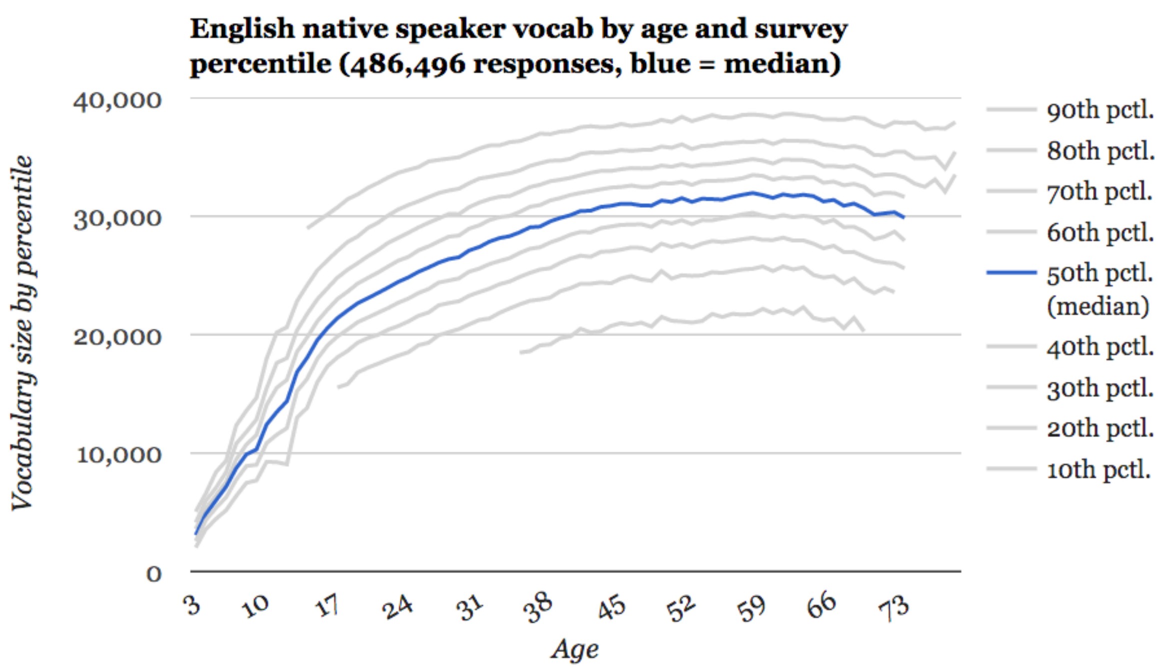
1) Sentence Reader replaces the coarse, document-based searches of the standard e-Discovery algorithms with a much more granular **sentence-level** search results.
2) Unlike other products purporting to offer "conceptual" or "semantic" searches, Sentence Reader is a fully pre-trained and thus genuinely literate sentence **reader**.

Sentence Reader's first innovation is valuable because "smoking gun" e-mails often derive their importance entirely from a single sentence, which may have nothing to do with the document's overall theme or even the sentence's surrounding context.

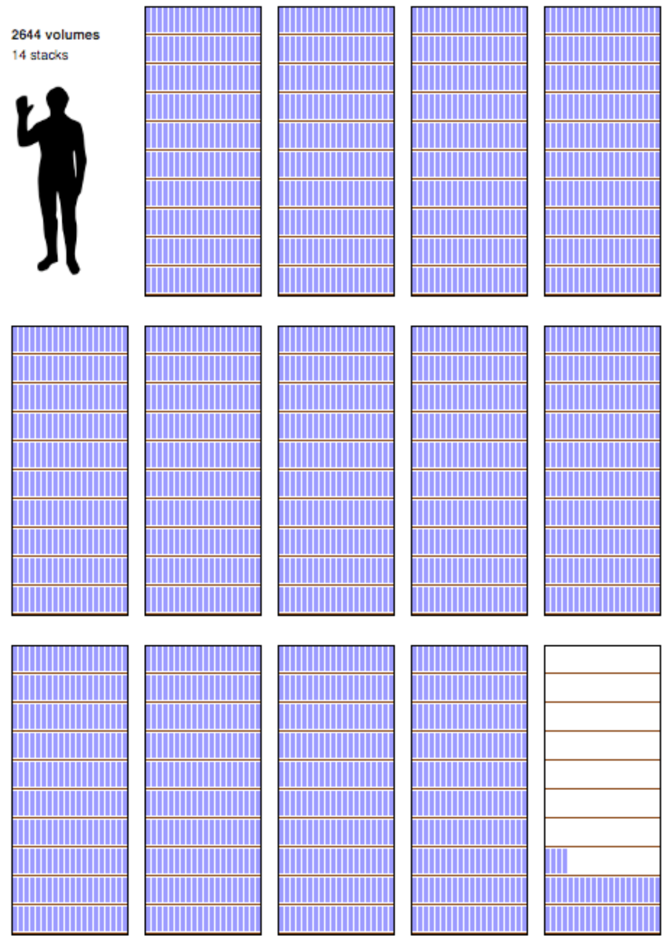
Consider the following: after chatting amiably for several paragraphs about golf, an e-mail's author might casually slip in an insider-trading tip. Because the traditional e-discovery algorithms are all based on Latent Semantic Analysis (LSA, sometimes called Latent Semantic Indexing), they will fail to detect this document. Why? Because LSA only tracks a document's overarching theme, which in this case is golf. Like a lazy lawyer, LSA will dismiss this e-mail as "basically about golf" instead of digging down to see what else might be there beyond golf chatter. As a result, LSA will overlook a "smoking gun" e-mail that has the potential to put someone behind bars.

By contrast, Sentence Reader will never be misled by the fact that bombshell sentence-length admissions are sometimes hidden within perfectly innocuous contexts. Unlike LSA, Sentence Reader scrupulously, obsessively adheres to the instructions given to junior associates being trained to review documents: "Be sure that you read every sentence of every document from start to finish because you never know what might be lurking later in a document. Don't get lazy and stop reading midway through!" Because LSA only detects the topics that predominate within each document, *LSA is explicitly designed to disregard this fundamental principle of proper document review*. Put differently, LSA was never intended to find needles in haystacks. At best, LSA is designed to distinguish between different types of hay. And LSA's fundamental illiteracy makes even that modest goal difficult to achieve.

LSA-based algorithms are illiterate in that they know nothing about the English language, let alone the world at large, before setting eyes on their target corpus. Sentence Reader, by contrast—and this is its second, most fundamental innovation—does its homework first, acquiring a rich understanding not only of the English language, but also the larger world from which words derive their meanings. Sentence Reader does this by ingesting the entirety of Wikipedia.



Whereas Sentence Reader's knowledge base is as deep and as broad as Wikipedia's, LSA models are trained on the same target corpus that they are being asked to interpret. LSA's approach is the equivalent of trying to read *War and Peace* in Russian without bothering to learn Russian first. Such a circular exercise is bound to fail because human languages are far too complex to be absorbed on the basis of any small sample, even one as rich as *War and Peace*. Asking an e-discovery algorithm to learn English based exclusively on a single target corpus dooms it to fail because the model invariably lacks the depth of knowledge and the breadth of linguistic expertise required to accurately situate words and sentences within a larger semantic context. Without such a context, words lose their meaning, and search engines become incapable of distinguishing between what matters and what does not.



To compete with human intelligence, machine learning algorithms must be armed with the same breadth and depth of knowledge as human readers. Better yet, algorithms should be trained on an even larger body of knowledge than most humans absorb in a lifetime—namely, the entirety of Wikipedia, which would comprise some 2,644 volumes if it were printed and bound like a traditional encyclopedia (see fig. 1).

By training on the entirety of Wikipedia, Sentence Reader learns the meanings of over a million different words, names, and concepts. To appreciate the magnitude of that accomplishment, consider that even the most sophisticated human readers know fewer than 40,000 English words (see fig. 2). Our experiments confirm the unparalleled power that Sentence Reader attains by virtue of its grasp not only of English, but of the names, places, dates, and facts set forth in Wikipedia. By doing its homework, Sentence Reader's search engine builds on a linguistic foundation as broad as it is deep.

The Weaknesses of Latent Semantic Analysis: How Illiteracy Cripples Search

To appreciate how Sentence Reader differs from the traditional LSA models that dominate e-Discovery today, it is useful to compare how the two methods compare on a particular search within a large, well-known public e-mail dataset: the Enron e-mail corpus.[2] Imagine that, after reviewing thousands of documents in the Enron dataset, you want to return to one that is proving more significant than you realized upon first encountering it. Unfortunately, all you remember about the e-mail in question is, vaguely, something about seeing a rhinoceros on an African safari. How can conceptual or semantic search be leveraged to relocate this e-mail based on some hazy recollections of its content?

If one queries the Enron dataset for the term ‘rhinoceros’ using LSA’s version of purportedly semantic search, the search draws a complete blank: LSA has absolutely no idea what a rhinoceros is. The LSA algorithm fails to get any traction because the term ‘rhinoceros’ does not appear anywhere in the Enron corpus, which constitutes LSA’s sole window into the English language and world at large.

So you try a slightly different query: ‘rhino’. This variation seems more promising because the word ‘rhino’ directly appears in the first few hits. Indeed, the dataset includes three identical copies of the short message ‘Rhino here.’, which constitute the first three search results. After that, however, come the following e-mails:

1) “Here ‘tis:”
2) “Here tis.”
3) “Here”
4) “I’m here.”
5) “I case I garbled mine, here ‘tis.”
6) “rhino here, sorry you lost your voice! Just goes to show that sometimes prayers are answered. Just kidding! We missed you at gas fair. It just wasn't the same without you. When are you comming back? Oh damm, I've got to go to a meeting. I'll write later. Rhino, out.”
7) “i am here dog”
8) “Brad & Gerry are here.....”

[2] https://www.kaggle.com/wcukierski/enron-email-dataset

9) "Pizza is here!"
10) "Here's the pics…"
11) "Drink beer! Down here"
12) "are you here?"
13) "are you here"
14) "Here ya go!"

Searching through the top 10,000 hits, the only other 'rhino' references appear in an e-mail containing this advertisement: "Travel between northern and southern California for only $44 each way with roundtrip purchase. Rhino chasers take note!" Still nothing about rhinos on a safari. Yet strangely, those top 10,000 search results do include over 16,000 references to the word 'here.' Exactly what is going on?

This example illustrates LSA's two fundamental limitations. First, because the word 'rhino' appears in the corpus alongside the word 'here' several times, LSA incorrectly concludes that the two terms must have closely related meanings and must bear on the same topic. Because LSA learns what words mean solely on the basis of their co-occurrences within the target corpus, it frequently learns incorrect meanings based on similarly spurious associations. Such misleading associations are inevitable when a training corpus is too small. And every target corpus will be.

Second, because LSA ignores sentences and effectively assumes that documents are homogenous in terms of the topics they contain, LSA is deliberately designed to ignore subsidiary (or contrasting) strains within documents and to focus instead on dominant, overarching themes. If 'rhino' plays no part in what LSA considers a document's dominant topics, then LSA will erroneously dismiss that document as containing little relating to a 'rhino' query.

Taken together, LSA's two weaknesses seriously undermine the algorithm's utility as a search engine. Indeed, the model was never really designed to function as a search engine. Search algorithms require much more fine-grained semantic understandings than the document-based topic modeling for which LSA was designed.

Sentence Reader, on the other hand, was deliberately engineered to function as a search tool. It avoids both of LSA's flaws. First, by pre-training on a very large corpus instead of the target corpus, the model never gets confused about what words mean. Second, by searching sentences instead of documents, it never overlooks sentences merely because they reflect subsidiary topics or themes.

Unlike LSA, which has no idea that the words 'rhino' and 'rhinoceros' are different ways of specifying the same animal, Sentence Reader is not fazed by the fact that the exact keyword 'rhinoceros' does not appear anywhere in the Enron corpus. Sentence Reader top hit for 'rhinoceros' lands us on precisely the passages we remembered reading, without even requiring us to scan through the lengthy e-mail chain within which they appear:

> "We were very privileged to pet baby elephants, rhino and
> zebra…. I was most impressed with the 7 rhino - ok so 6 of them were in the distance and we only knew they were rhino by looking through nans fantastic binoculars, a quick thank you to nan for lending them to me."

Clearly, it pays to use tools that have been specifically engineered for search, instead of repurposing LSA for a task that it was never designed to handle.

Hunting for Easter Eggs: The Test Dataset

To systematically test the power of Sentence Reader against the industry's standard semantic search algorithm, LSA, we utilized a large, hand-labeled public dataset created by researchers at the Center for Intelligent Information Retrieval at the University of Massachusetts, Amherst: the Web Answer Passages (WebAP) dataset.[3] The WebAP dataset has been used for a variety of different web-based information retrieval tests.[4]

For sentence-level retrieval tasks, the relevant dataset consists of 80 different search queries spanning 80 different topics.[5]

---

[3] https://ciir.cs.umass.edu/downloads/WebAP/

[4] Daniel Cohen, W. Bruce Croft. 2018. "A Hybrid Embedding Approach to Noisy Answer Passage Retrieval". In *Proceedings of the 40th European Conference on Information Retrieval (ECIR 2018)*. Liu Yang, Qingyao Ai, Damiano Spina, Ruey-Cheng Chen, Liang Pang, W. Bruce Croft, Jiafeng Guo and Falk Scholer. 2016. "Beyond Factoid QA: Effective Methods for Non-factoid Answer Sentence Retrieval". To appear in *Proceedings of the 38th European Conference on Information Retrieval (ECIR 2016)*. Evi Yulianti, Ruey-Cheng Chen, Falk Scholer, and Mark Sanderson. 2016. "Using Semantic and Context Features for Answer Summary Extraction". *In Proceedings of the 21st Australasian Document Computing Symposium (ADCS '16).* Ruey-Cheng Chen, Damiano Spina, W. Bruce Croft, Mark Sanderson and Falk Scholer. 2015. "Harnessing semantics for answer sentence retrieval". In *Proceedings of the Eighth Workshop on Exploiting Semantic Annotations in Information Retrieval, ESAIR '15 (CIKM'15 workshop)*.

[5] In their original form, prior to any preprocessing (stemming, lemmatizing, removing stopwords), each of the queries took the form either of a question, such as "What is artificial intelligence?" or a command such as, "Describe the history of the U.S. oil industry."

The creators of the WebAP dataset used a web search engine to collect a set of potentially responsive documents for each of the 80 queries.[6] They then asked a human annotator to rate the sentences in those documents in terms of their relevance to the original query.[7] All told, nearly a million sentences were assigned one of five possible labels: PERFECT, EXCELLENT, GOOD, FAIR, or NONE. Note that because the documents themselves were initially collected on the basis of their relevance to a particular query, a NONE label does not mean that the sentence has no relevance whatsoever to the topic of its corresponding query. Rather, the labels reflect the human annotator's estimate of how directly a given sentence responds to the likely intent of the original query. Of the million sentences in the entire WebAP, fewer than 1% were assigned a rating better than NONE:

	Entire WebAP Corpus / 80	= Average per query
PERFECT	4,508 sentences	56 sentences
EXCELLENT	4,283 sentences	54 sentences
GOOD	783 sentences	10 sentences
FAIR	149 sentences	2 sentences
NONE	938,977 sentences	11,737 sentences

Because this dataset was created (through a long chain of modifications and refinements recounted in the papers cited above) for the purpose of web passage retrieval, for purposes of our e-mail-based application, we combined sentences drawn of the WebAP dataset with the Enron dataset. To increase the difficulty of locating the sentences deemed relevant by the human annotator, moreover, we separated all of the WebAP sentences from each other and randomly injected them into the Enron corpus. In effect, we created an exceptionally difficult Easter Egg Hunt, asking the search algorithms to identify a tiny set of approximately 122 Easter Eggs per query: the sentences labeled PERFECT, EXCELLENT, GOOD, or FAIR for a given query.

---

[6] The human annotators actually labeled documents from 82 queries, but since two of those queries yielded no sentences considered PERFECT, EXCELLENT, GOOD, or FAIR, we and other researchers using the WebAP dataset have focused on the remaining 80 queries.

[7] To mitigate the tedium of labeling nearly a million sentences, annotators were allowed to assign labels to passages instead of individual sentences, with every sentence within given passage (typically only 2-3 sentences) receiving the passage's label.

The resulting dataset, while exceptionally complex insofar as it consists of an artificial merger of sentences covering a multitude of unrelated topics, remains very realistic both in terms of its overall size and in terms of the proportion of documents that one ultimately seeks to identify. For each of the 80 queries, we are asking the search engines to find approximately 122 individual "needles" that have been randomly hidden within an immense haystack, consisting of 517,400 documents and 6.6 million sentences. Again, the ratio of needles to hay remains very realistic in terms of the proportion of reviewed documents that might become formal exhibits, or evidence in litigation.

Given that "smoking gun" e-mails often derive their importance entirely from a single sentence, moreover, and not from the document's overall theme or topic, e-discovery search tools must not be misled by the fact that bombshell disclosures are sometimes hidden in perfectly innocuous contexts. By randomly interspersing the target WebAP sentences amidst the Enron e-mails (typical of e-mail communications within a large business), our dataset explicitly tests a search engine's ability to obey the instructions given to associates performing document review: be sure that you read each and every sentence of each and every document. Because the WebAP Easter Eggs have been carefully hidden within the more familiar Enron e-mails, the resulting dataset creates an exceptionally difficult search challenge.

Results: Sentence Reader is 17 Times More Efficient than LSA

Given that the goal of a search engine is to rank a set of documents in terms of their relevance to a given query, the labels from the WebAP dataset can be used to gauge the relative effectiveness of the LSA and Sentence Readers search engines. To be sure, the notion of relevance is a notoriously slippery one, leaving a great deal of room for disagreement depending on a user's ultimate objective and on the subjective views of any particular relevance assessor. Put simply, if one were to present the same 80 WebAP queries and documents to a different set of human annotators, one would expect the new set of human annotators to reach significantly different conclusions as to which sentences constituted PERFECT, EXCELLENT, GOOD, or FAIR results for a given query. Indeed, the more closely one inspects the WebAP labels, the more one appreciates the difficulties associated with creating an objective, gold standard for search relevancy.

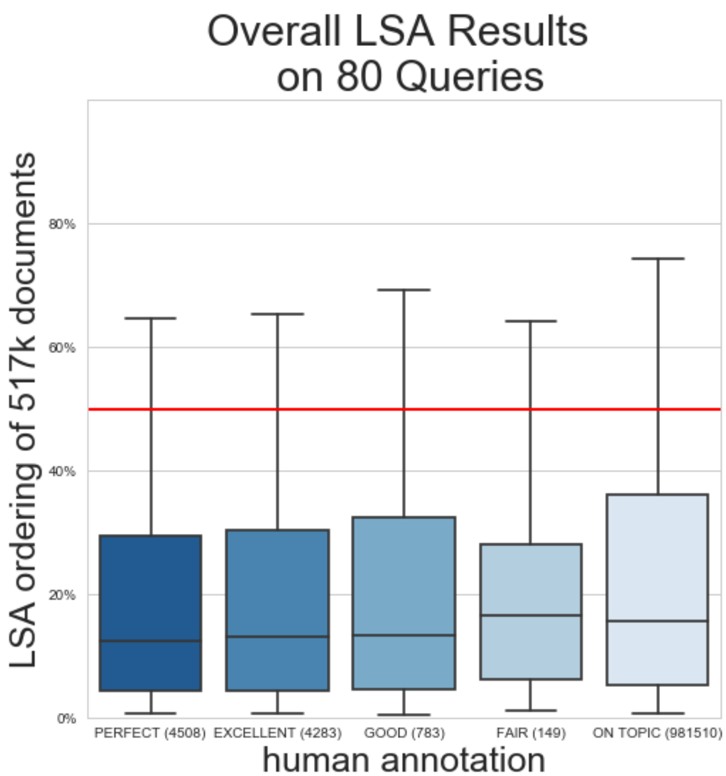
Nevertheless, despite the large dose of subjectivity that invariably infuses relevancy judgments, the sheer size the WebAP dataset renders it a useful—albeit imperfect—metric against which to gauge the *relative* effectiveness of different search engines. However noisy the WebAP's labels are believed to be, that subjective "noise" poses a similar impediment to any and all search engines. Thus, however flawed the WebAP's labels may be as a means of measuring objective relevance, they remain extremely useful as a means of *comparing* the effectiveness of different search engines.

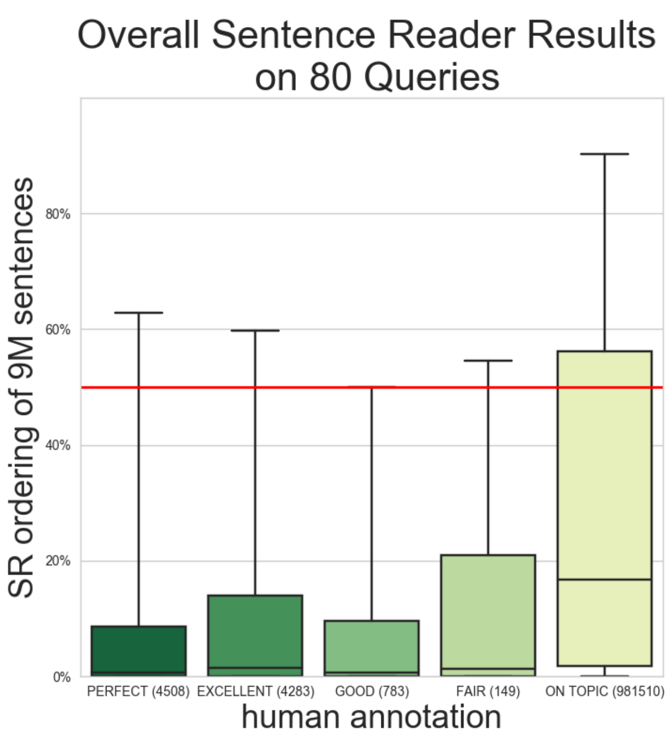
One might analogize the subjective "noise" within the WebAP labels to a strong headwind slowing the times of runners in an annual footrace: although the windy conditions make it impossible to compare times across different years (races run in different conditions), the fact that all participants from a given race confronted the same adverse conditions *still allows for meaningful relative comparisons*. So too with the WebAP dataset: though based on highly subjective human judgments, insofar as the dataset's noise affects all competitors equally, the WebAP dataset remains a useful comparative metric.

And when one compares the relative performance of LSA and Sentence Reader in terms of their ability to locate the hidden Easter Egg sentences within our test dataset, the contrast is stark. If one asks "What percentage of the material, as ranked by the search engine, would a human searcher need to read in order to find half of the Easter Egg sentences deemed PERFECT by WebAP's human annotators?", our experimental results reveal that a reviewer using LSA would be forced to wade through more than 17 times as much material as a searcher using Sentence Reader:

Percentage to Review in Order to Reach Median for Each Label Type

	PERFECT	EXCELLENT	GOOD	FAIR
Sentence Reader	.72%	1.41%	.69%	1.28%
LSA	12.45%	13.03%	13.30%	16.51%
Improvement as multiple	17.29 times	9.24 times	19.28 times	12.90 times


Overall LSA Results
on 80 Queries
LSA ordering of 517k documents
80%
60%
40%
20%
0%
PERFECT (4508)
EXCELLENT (4283)
GOOD (783)
FAIR (149)
ON TOPIC (981510)
human annotation


Overall Sentence Reader Results
on 80 Queries
SR ordering of 9M sentences
80%
60%
40%
20%
0%
PERFECT (4508)
EXCELLENT (4283)
GOOD (783)
FAIR (149)
ON TOPIC (981510)
human annotation

Conclusion

For clients whose litigators bill them on an hourly basis (as nearly all do), improving the efficiency of the lawyers' search engine by a multiple of 17 translates directly into a 17-fold reduction in those costs. Alternately, that same 17-fold improvement can be interpreted as increasing by a factor of 17 the number of "hot" documents that one's lawyers are able to locate within a fixed time interval.

Either way, given that document review typically comprises the single greatest expense incurred by litigants, a 17-fold increase in search power represents a massive benefit to clients and litigators alike. If, from the client side, Sentence Reader's power represents a 17-fold reduction in search costs, from the lawyers' side it suggests a 17-fold increase in power, quality, and confidence.

Indeed, as the economics of document review force many litigants to outsource the task of front-line review to overseas attorneys, the question of how to ensure the quality and integrity of the review process becomes ever more salient. In circumstances where front-line reviewers have likely never even met the U.S.-based attorneys who are ultimately responsible for overseeing their work and for providing quality control (i.e. a malpractice defense), a coarse, illiterate search engine like LSA provides little reassurance. Because LSA was never designed or intended for search—let alone for quality assurance—it lacks the fine-grained semantic intelligence, the linguistic expertise, and the sentence-level rigor that make Sentence Reader 17 times more effective.

By first absorbing the entirety of Wikipedia and then applying that knowledge to evaluate the meaning of each individual sentence in the target corpus, Sentence Reader's understanding of English is broader, deeper, and more precise than LSA's:

	LSA's Problem	Sentence Reader's Solution
Broader Knowledge	LSA frequently draws a blank because it has no idea what a word means unless that word happens to appear in the target corpus.	Sentence Reader understands all words that appear at least 8 times in Wikipedia (over 1 million terms), even if the word does not appear anywhere in the target corpus.
Deeper Knowledge	Even when LSA *has* seen a word before, it fails to learn accurate meanings because LSA rarely sees enough examples from which to generalize.	Because Wikipedia is such a reliable and comprehensive knowledge base, Sentence Reader has, like an experienced human reader, encountered most words in thousands of different contexts; Sentence Reader is genuinely literate.
Greater Precision	Even when LSA does know what a word means, it looks only at documents and never at individual sentences. As a result, LSA often misses highly relevant documents.	Like a diligent human being, Sentence Reader reads every single sentence of every single document to ensure that it never misses anything that the user might want to see.

These three advantages give Sentence Reader a leg up at every stage in the reading process: Sentence Reader is better than LSA at recognizing words, better at knowing what words mean, and better at applying its knowledge to grasp the meanings of individual sentences. Because Sentence Reader outperforms the competition at every step of the reading process, its net results are 17 times better.

Whether used for initial explorations in anticipation of litigation or as back-end QC on a manual review process that has already been completed, Sentence Reader's breadth of knowledge and power of discernment are unmatched. By incorporating the very latest research from the fields of Information Retrieval and Natural Language Processing, Haave has created a genuinely state-of-the-art tool. To learn more about Sentence Reader and its potential to help not only lawyers, but anyone requiring a state-of-the-art search engine, please contact Nora Pykonnen at **nora@haave.io**.