DisputeSoft
Software Dispute Experts

An Overview of Machine Learning Applications to Technology Assisted Review for E-Discovery Part Two: Optical Character Recognition and Natural Language Processing

By Sarah Bullard, formerly of DisputeSoft

This is the second article in a four-part series on technology assisted review (TAR), a process that uses machine learning to increase efficiency and decrease the cost of document review in discovery.

The last article discussed differences between supervised and unsupervised machine learning algorithms and explained why supervised learning algorithms are more commonly used with TAR for e-discovery. This article explores two machine learning techniques pertinent to TAR: optical character recognition and natural language processing.

Optical Character Recognition

Optical character recognition (OCR) is more commonly known as text recognition. Computers understand English characters and words when they are entered into the machine using an unambiguous keyboard and stored in a similarly discrete manner. Microsoft Word has recognized English words—and spell-checked, grammar-checked, and even autocorrected them—since the early 1990s. But when it comes to e-discovery, attorneys are not always sifting through neatly-typed text editor documents. In fact, one of the most popular document types reviewed during the e-discovery process is a TIFF file—a static image file that does not inherently contain anything except pixels. OCR is necessary to turn those pixels into recognizable letters and words that can be processed by a TAR algorithm.

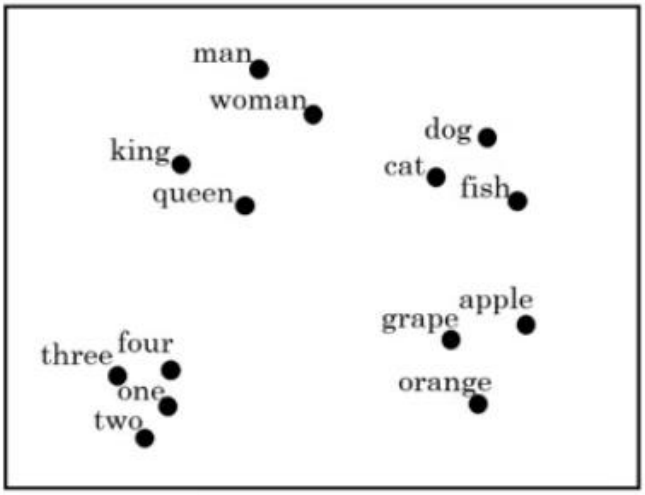
An example of an optical character recognition process run on an image of a license plate. Image source: the Visual Arts at Mason Gross website.

OCR programs usually start by applying standard sharpening and contrast-increasing filters to the photos provided. They then apply edge detection to separate dark areas from light areas and extract the letters themselves. Next, the program uses a machine learning algorithm to match the pictures of the letters to their discrete character values. The machine learning algorithm in question is usually a supervised learning algorithm. The algorithm may be ready out of the box for the purpose of identifying standard or generalized fonts, but users can also typically tailor it to their needs by providing seed sets of images of characters mapped to their discrete character values. More specialized OCR algorithms can be tailored to the languages, character sets, or fonts specific to a given use case. They can even be trained to recognize an individual's handwriting, although they are usually much less accurate in this case, and very inaccurate at recognizing cursive writing. With standard printed fonts, however, OCR engines can routinely achieve 98% accuracy or better, meaning that out of every 100 characters, the engine will correctly identify 98 of them. And according to CVisionTech, OCR engines can usually achieve up to 99% accuracy.

Natural Language Processing

Natural language processing (NLP) is an application of machine learning used to parse human language. Consider a model for supervised TAR that uses keywords to categorize documents. An attorney runs documents through OCR, and words identified using OCR are searched against a list of relevant terms. If any of the keywords are found in a given document, it is classified as relevant or potentially relevant.

However, a more efficient strategy would be to determine how each word relates to other words and to the case as a whole in a more flexible way. Large lists of relevant words and phrases, with as many synonyms as a person can think of included, will never replace a more complex "understanding" of how they are related to each other and to the case. Word2Vec is an NLP toolset that has increased in popularity since its creation in 2013. It is trained by feeding in pairs of words within the set of documents to be searched. The network then generates statistics based on these pairs, and those statistics are used to generate a weighted matrix that is used to understand which words are most closely related to each other in a given circumstance. Specifically, Word2Vec finds these relationships via a neural network; that is, a machine learning algorithm modeled after the human brain.



Tools like Word2Vec use neural networks to store words as vectors, which reveals information about their relationships to each other. Such a tool can group similar words together, as shown here. This example is from Analytics Vidhya.

As evident in the chart above, Word2Vec has used a neural network to create a two-dimensional spatial representation of several words. The chart displays words that have been grouped together

based on semantic similarities, including numbers, fruit, and animals. The proximity of data points within a group indicates a stronger correlation between the words' semantic relationship. Note that people are grouped closely together, but Word2Vec appears to distinguish between people based more strongly on royal status (king, queen vs. man, woman) than on gender (king, man vs. queen, woman).

Take a look at some real-life examples of connections Word2Vec has been able to make (examples from Skymind). Below are several connections in the old SAT analogy format. This format consists of three given terms. The first two terms, which are separated by a single colon, are related to each other in some way that is not explicitly stated. The third term is separated from the first two by a set of double colons, and followed by a single colon. The job of the answerer is to find a fourth term that relates to the third term in the same manner that the second relates to the first.

In this manner, Word2Vec was provided with the first three terms and came up with several suggestions, shown in blue, for the final term.

China:Taiwan::Russia:[Ukraine, Moscow, Moldova, Armenia]

house:roof::castle:[dome, bell_tower, spire, crenellations, turrets]

knee:leg::elbow:[forearm, arm, ulna_bone]

Word2Vec can also identify other types of relationships. The examples below illustrate what happens when you subtract one concept from another:

Human – Animal = Ethics

Library – Books = Hall

Tools like Word2Vec can make e-discovery experts' jobs easier when training, testing, and fine-tuning a TAR process. For example, suppose an expert wanted to use e-discovery to assess documents related to object-oriented programming in connection with a software matter. The expert gathers a seed set and testing set and begins training and tuning a model for e-discovery. The expert notices that all documents classified as responsive are correctly identified, but that some documents that are responsive are classified as unresponsive. To resolve this problem, the expert could broaden the scope of the seed set or use Word2Vec to help expand the criteria the

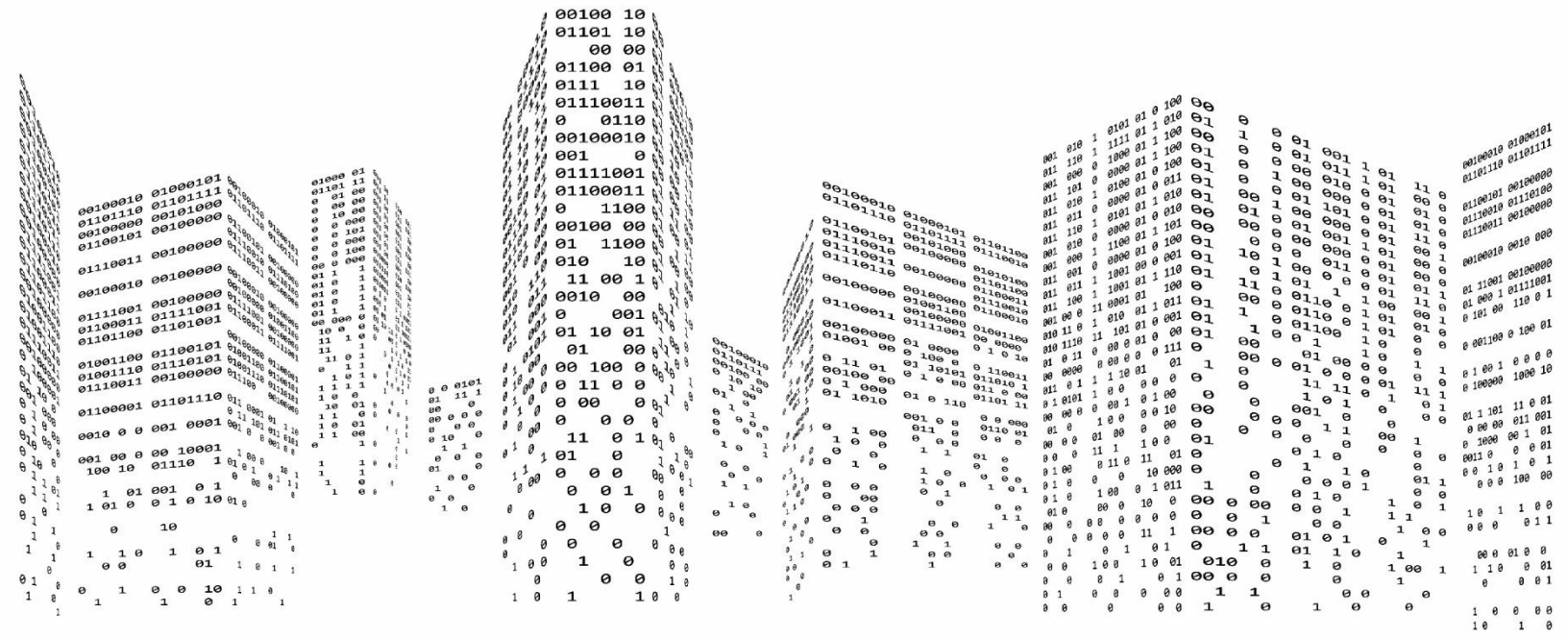
algorithm uses to determine whether a document is responsive. The expert could gather a corpus of documents related to object-oriented programming—textbooks, articles, and Stack Overflow questions might be a good place to start—and then use Word2Vec or a similar NLP tool to find relationships between the words and group them together by relevance. Lastly, the expert can adjust the machine learning algorithm to accept synonyms for words it has learned are relevant, or to combine relevant terms to find additional terms to include.

Natural language processing tools are useful for identifying intended meanings within textual data. They can be used to classify textual data into predefined categories. With minimal training, NLP models can be trained to do so with 74% accuracy, as shown in a brief demonstrative experiment from Susan Li: Multi-Class Text Classification Model Comparison and Selection. With more training, the model can reach even higher degrees of accuracy.

Conclusion

Both OCR and NLP are machine learning technologies that are beneficial to document processing for e-discovery. OCR can turn image files into text files that can be used by a TAR system. NLP allows data scientists and algorithms to observe connections between linguistic terms that may be otherwise difficult or impossible to detect. Both OCR and NLP work together to facilitate the TAR process. To learn more about the role of machine learning applications in TAR for e-discovery, check out the next article in the series, which tackles the question of whether TAR can be more effective than exhaustive manual review.

To learn more about DisputeSoft's e-discovery services including identification, recovery, preservation, and analysis of systems, databases, and other non-custodial evidence, visit our electronic discovery services page and explore a representative e-discovery case: General Electric v. Mitsubishi Heavy Industries.



If you are in need of an e-discovery expert, we invite you to consider DisputeSoft.

Contact Information

Jeff Parmet, Managing Partner

301.251.6182

jparmet@disputesoft.com

12505 Park Potomac Ave. | Suite 475 | Potomac, MD | 20854