Beyond Words – A Search Engine Can Also Find Items Like Numbers and Emojis in Your Data

Article contributed by dtSearch®

When most people think about text retrieval, they think of word searches. A natural language search like *get me the Smith memo on Project Urgent in South Dakota* would be one example. A structured search request like *Smith and project urgent and South Dakota and not (project irrelevant w/12 vacation)* would be another, with w/ designating word proximity.

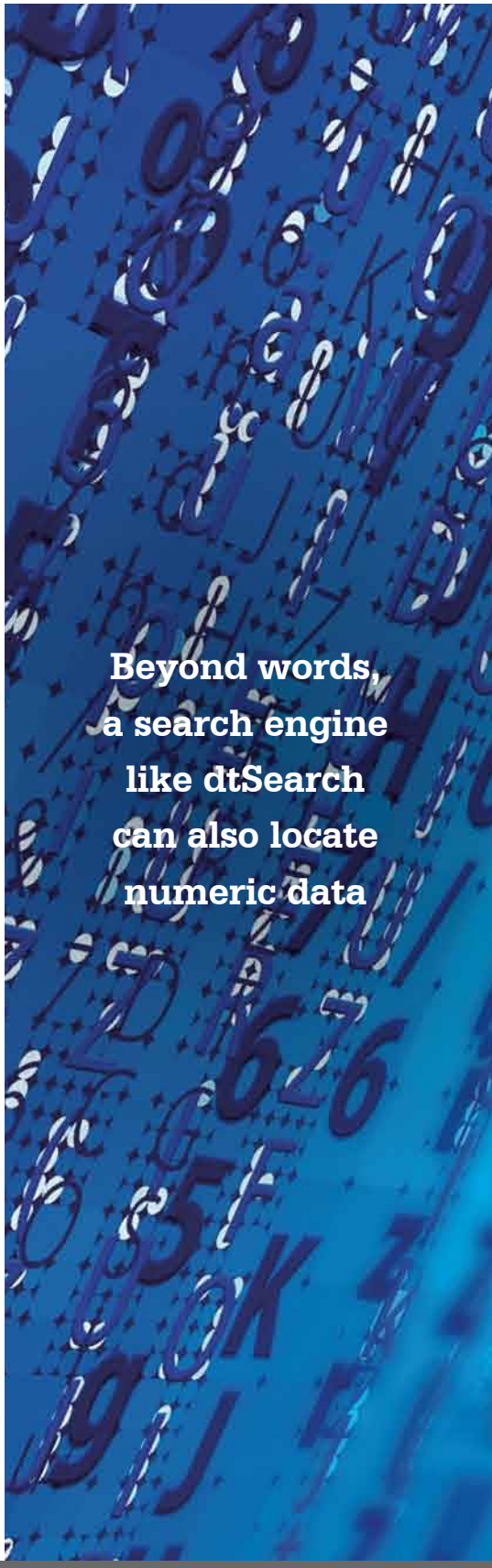
For both of these examples, a search engine like dtSearch® will instantly sift through terabytes to find matching items in files, emails and other data. You can then individually browse the full text of retrieved items with highlighted word hits. Multiple concurrent searchers can also do the same.

But beyond words, a search engine like dtSearch can also locate numeric data. You can search for specific numbers like 147 or numeric ranges like 64 to 998. Date searching supports different date formats. For example, you could search for 3/1/21 and find not only 3/1/21 but also different formats of the same date such as March 1, 2021. The reverse is true as well, so if you enter March 1, 2021, the search engine can also find 3/1/21. And you can also search for date ranges like 3/1/21 to April 15, 2021.

The application can further combine numeric and word searching. For example, you could search for *project urgent w/15 South Dakota* and then add a date range of *2/1/20 to 5/31/21*. Matching files, emails and emails attachments can have any of these elements anywhere in the full content. Or you can limit elements to specific fields. This is particularly helpful with dates, so you could look for *project urgent w/15 South Dakota* anywhere, and *2/1/20 to 5/31/21* limited to email date sent metadata.

The application further lets you search for credit card numbers. Of course, you can find a specific credit card number you are looking for. But dtSearch can also go beyond that and identify any credit card numbers that may appear anywhere in data. Here is how that works. Every time dtSearch sees X digits together that could be a credit card, it runs those numbers through a standard credit card verification algorithm. This will check to see if those X digits do in fact represent a valid credit card number and if so flag them as a credit card number.

Unfortunately, there is no neat algorithmic validity check to see if a series of numbers represents a valid Social Security number like there is for checking if numbers represent a credit card. But you can use numeric patterns matching. With that, you could look for number, number, number – number, number – number, number, number, number, for example, and find anything that matches.



Additionally, Unicode covers hundreds of international character sets. You could enter an English or other European-language search, a double-byte character Chinese, Japanese or Korean search or a right-to-left Hebrew or Arabic search. Beyond words, there are Unicode mathematical character sets, for example. And you can also search for emoji – smiley face 😊

Also beyond words, you can search for hash values. Each document and each email has a unique hash value associated with it that acts as a forensics-oriented fingerprint for that file. The application can automatically generate hash values for all documents and emails. Then the application can search for those hash values as part of a search request, or simply display them in search results as part of an unrelated search request.

dtSearch instantly searches terabytes by first building an index across the dataset. Indexing works as an internal guide, storing each unique word and number and the location of both. After indexing, individual searching or even multithreaded concurrent searching is typically instantaneous. Online searching can operate in a completely stateless manner, meaning that there are no limits on the number of concurrent search threads that dtSearch can process at once.

Indexing is easy. Just point to the directories you want to index, and the software will do the rest. No need to tell dtSearch what type of data you have. The software will on its own automatically recognize and work with PDF files; Microsoft Word, Access, Excel, PowerPoint, OneNote files; popular email formats; and even compressed email attachments.

Each index can hold up to a terabyte of text consisting of words and numbers. There are no limits on the number of terabyte-size indexes that dtSearch can automatically build and search across. While indexing itself can take some time, after indexing, even high-volume concurrent search across terabytes can operate instantly.

dtSearch has enterprise and developer products that run “on premises” or on cloud platforms to instantly search terabytes of “Office” files, PDFs, emails along with nested attachments, databases and online data. Because dtSearch can instantly search terabytes with over 25 precision search options, many dtSearch customers are Fortune 100 companies and government agencies. But anyone with lots of data to search can download a fully-functional 30-day evaluation copy from dtSearch.com

Article contributed by dtSearch®