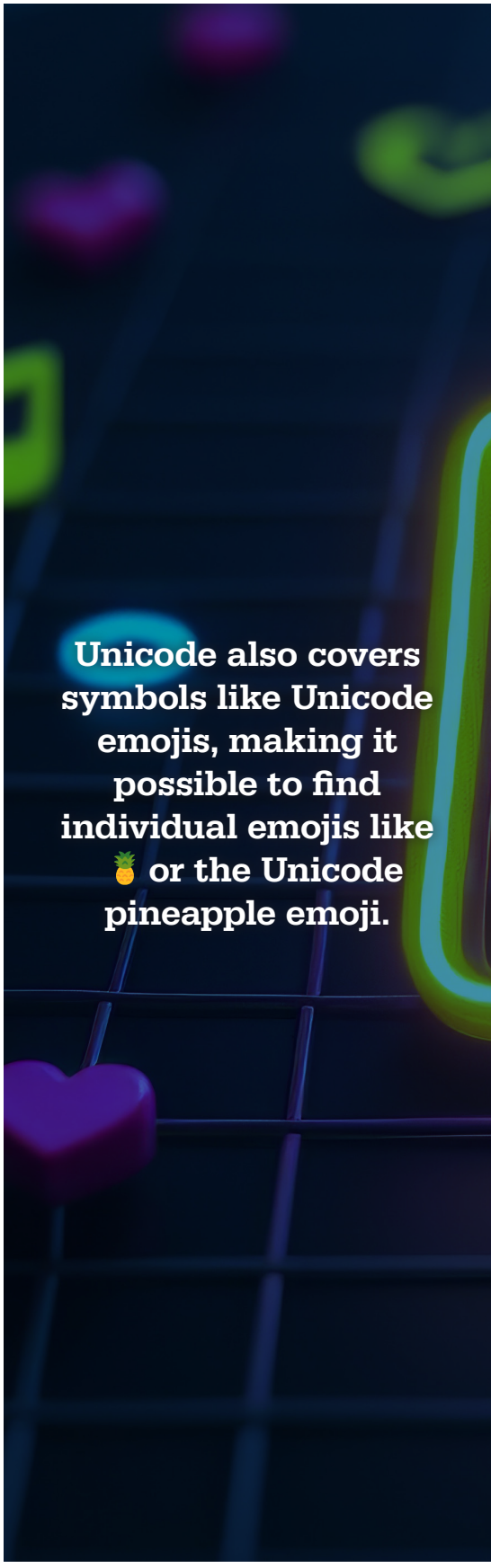
Extending Enterprise Search to International Languages, Numeric Expressions and Even Emojis

Article contributed by dtSearch®



When people think about enterprise search, they typically think about searching in English. In reality, enterprise search can extend to international languages, numeric expressions and even emojis. But let's start with English.

Take a search for *banana mango pineapple*. A basic "all words" search would find files, emails and the like containing all three fruits, while an "any words" search would look for matches with even just one mention of any of the fruits. In dtSearch®, for example, you can fine-tune searching using phrase, Boolean, proximity and metadata parameters, such as a query for the phrase *Ecuadorean banana* or *Mexican mango* or *Hawaiian pineapple* in a file or email that also has *direct ship* w/7 of *air express* and no mention of *truck route* in subject metadata.

Any other standard search options?

Concept searching in dtSearch can cover related words like *plantain* for *banana*. Fuzzy searching adjusts from 1 to 10 to sift through typographical or OCR errors like *banama* for *banana*. Date and date range searching work across all text or just certain metadata to automatically pick up popular date formats. A search for *date(10/1/25 to 11/20/26)* would pick up *10/15/26*, *Oct 15, 2026* and *October 15, 2026*.

Backing up a step, for those not familiar with enterprise search, how does it work?

Enterprise search enables instant concurrent searching across terabytes from a Windows network, from an

on-premises web server or from the cloud. But to enable this instant concurrent searching, enterprise search first has to index the data. An index pre-compiles each unique word and its place in the data. While resource-intensive at the software level, all you need to do to start enterprise search indexing is point to the relevant email archives and other folders to cover. It doesn't even matter if some of the data is remote. For example, dtSearch's off-the-shelf end-user Windows product can interchangeably index content both local and remote so long as the remote content appears as part of the Windows folder system.

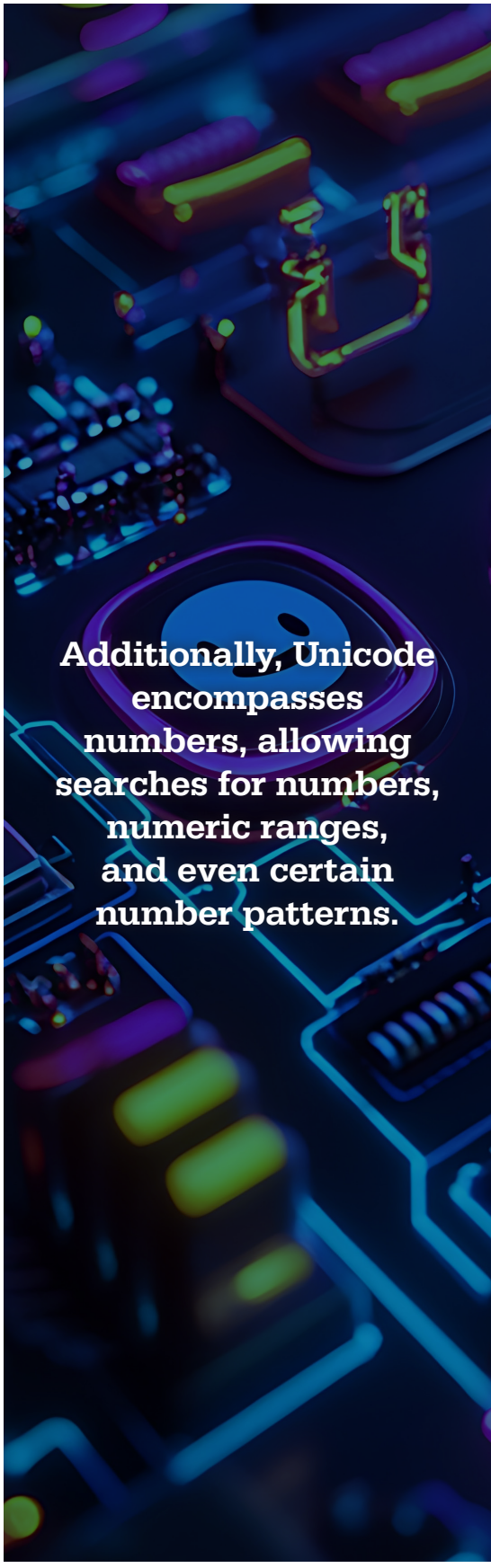
Doesn't the indexer need to identify the file type of each file?

To apply the right parsing specification, the indexer has to pinpoint the file version and file type: Microsoft Word, Access, Excel, PowerPoint, OneNote, PDF, email formats, etc. But the indexer can identify the correct file information on its own as it accesses a binary format. And it is not just standalone local and remote files that the indexer can work with. You can have an email with a ZIP or RAR attachment holding a Word document that itself contains an Excel spreadsheet and the indexer will parse all of that. Accessing local and remote files through their binary formats has multiple benefits.

Like what?

The indexer can parse files and emails a lot faster going through individual binary formats rather than retrieving each file in its associated application. Binary format access further allows the indexer to correctly determine the exact file format regardless of file extension. That way, it doesn't matter if a Word document has a .PDF extension or if a PDF a .DOCX extension. And text that an associated application display might obscure is on the same footing as any other text. Deep metadata, white writing against a white background, black writing beneath black redaction rectangles, and tracked changes still in a file are all apparent in the binary format.

Article contributed by dtSearch®



You mentioned that enterprise search can cover much more than just English words?

Leveraging Unicode in binary formats, the indexer can parse text in any of the hundreds of international languages that Unicode spans, including a broad variety of European languages and alphabets, left-to-right Middle Eastern text and double-byte Asian text. Even ancient languages like Sanskrit and Egyptian hieroglyphics have Unicode representations. A single file or email can have multiple different Unicode encodings, and the indexer will track the entire evolution. Phrase, "any words," "all words," Boolean, proximity and metadata-focused search features operate across different Unicode languages. And fuzzy searching still adjusts from 1 to 10 to sift through typographical and OCR errors in a broad range of multilingual text.

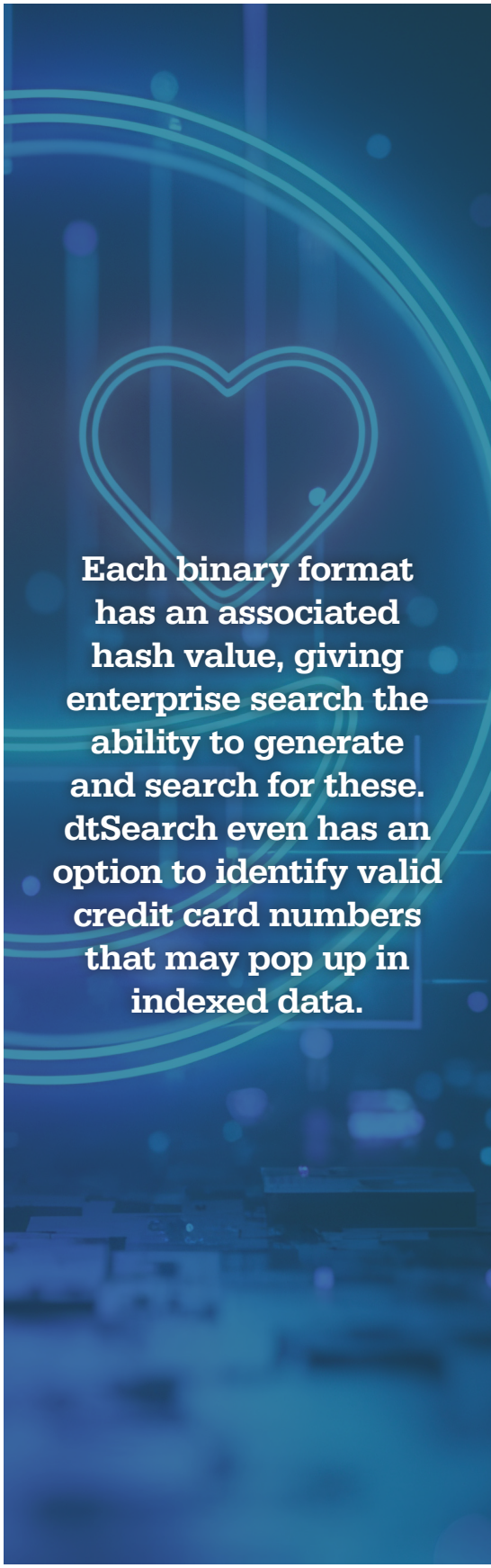
Does Unicode enable non-word-based search options?

Unicode also covers symbols like Unicode emojis, making it possible to find individual emojis like 🍍 or the Unicode pineapple emoji. Additionally, Unicode encompasses numbers, allowing searches for numbers, numeric ranges, and even certain number patterns. Each binary format has an associated hash value, giving enterprise search the ability to generate and search for these. dtSearch even has an option to identify valid credit card numbers that may pop up in indexed data. And numeric search elements can work with word-based searching. The synthesis enables the refinement of a word-based query through the addition of number or numeric range elements across all text or in specific metadata.

How does relevancy-ranking work?

By default, dtSearch will rank retrieved files by hit term density and rarity. Take an "any words" search for *banana mango pineapple*. If *banana* and *mango* are prevalent across indexed data but *pineapple* is relatively rare, *pineapple* will get a greater relevancy score, with files with the densest *pineapple* mentions

Article contributed by dtSearch®



scoring highest. dtSearch also supports positive and negative variable term weighting. Variable term weighting can apply universally to individual words or numbers across all text. Or variable term weighting can apply differentially to words or numbers in metadata or towards the top or bottom of files. For a fresh view of search results, the software can re-sort by a completely different metric like file date or file location. Regardless of ranking type, dtSearch displays retrieved files with highlighted hits for easy browsing.

Final thoughts?

Let enterprise search ramp up efficiency throughout your organization by enabling instant concurrent searching across international languages, numeric data and so much more. Visit dtSearch.com to download a fully-functional 30-day evaluation.

Article contributed by dtSearch®

For a fresh view of search results, the software can re-sort by a completely different metric like file date or file location. Regardless of ranking type, dtSearch displays retrieved files with highlighted hits for easy browsing.