

Return of the Zombies: How Enterprise Search Can Help

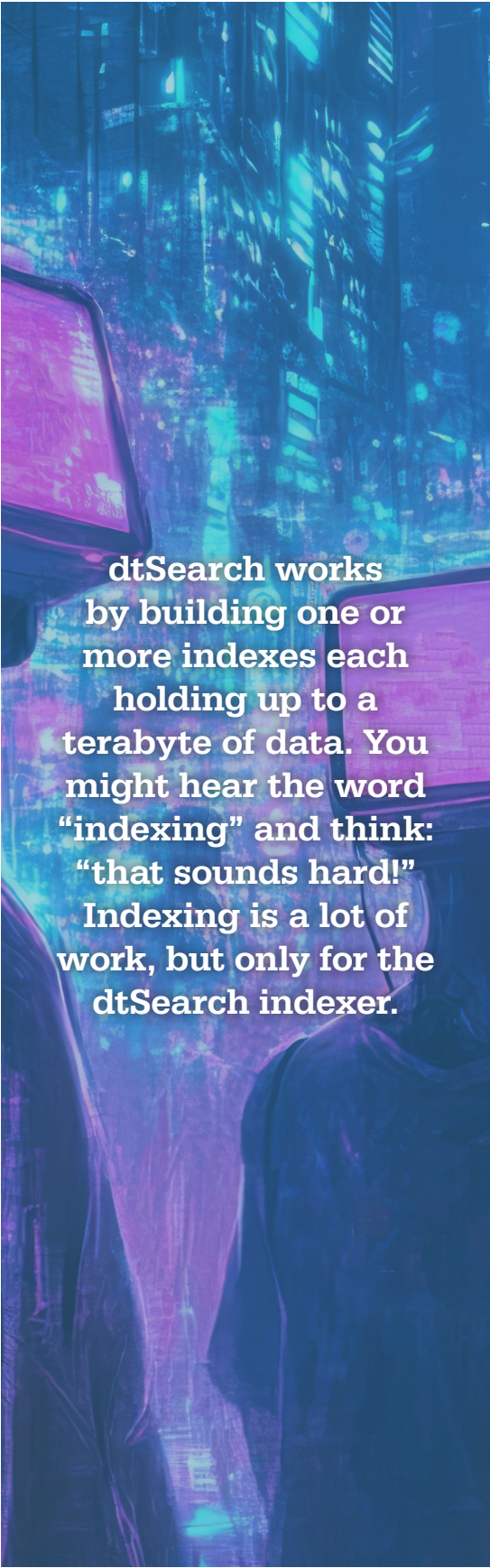
May is Zombie Awareness Month. And today's topic is the return of the zombies and how enterprise search can help. In the data world, I would define zombies as anything left undead in your data that can pop up later and cause trouble. This makes finding zombies in data a classic case for enterprise search and dtSearch.

dtSearch works by building one or more indexes each holding up to a terabyte of data. You might hear the word "indexing" and think: "that sounds hard!" Indexing is a lot of work, but only for the dtSearch indexer. All you need to do to kick off indexing is point to the folders, email archives and the like to cover, and enterprise search will take it from there.

If you went through files yourself, you'd typically review each in its associated application: Outlook for emails, Adobe Acrobat Reader for PDFs, Microsoft Word for memos, etc. Rather than reviewing files in their associated applications, the indexer works with the binary version of files. For each binary file, the indexer figures out the exact file format then applies the correct template to recognize all text and metadata. The part of the software that does this is called the document filters.

Binary file access works for parsing both standalone files and multilevel nested files. You can have an email with a ZIP or RAR attachment that itself contains a Word file with an Excel spreadsheet inside. Binary formats let the indexer and its document filters automatically work with all of that, down to the innermost component. While mismatched file extensions might throw you off if you were individually reviewing files, mismatched file extensions like an Access database saved with a PDF extension or a PowerPoint saved with a OneNote file extension, will not affect the indexer. The reason is that the document

Article contributed
by dtSearch®



**dtSearch works
by building one or
more indexes each
holding up to a
terabyte of data. You
might hear the word
"indexing" and think:
"that sounds hard!"
Indexing is a lot of
work, but only for the
dtSearch indexer.**

filters look inside the binary format to figure out the file format; the file extension is not relevant.

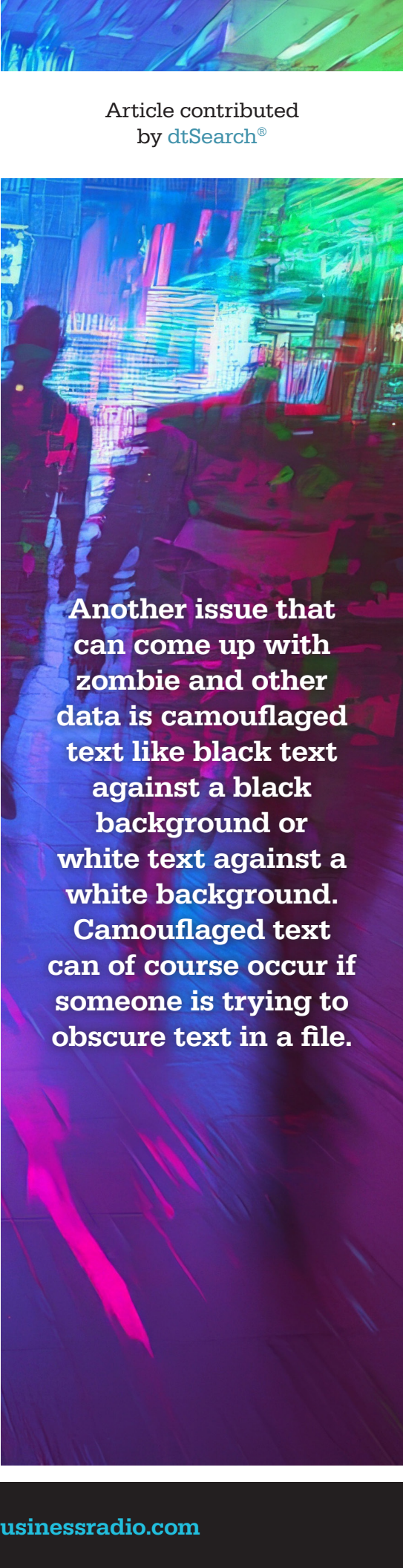
Another issue that can come up with zombie and other data is camouflaged text like black text against a black background or white text against a white background. Camouflaged text can of course occur if someone is trying to obscure text in a file. But it can also occur with certain redaction programs that show redacted text under a black rectangle where the text may look like it is gone but nonetheless remains. Similarly, with track changes on, certain text may appear deleted. But if the changes are not fully accepted, the deleted text can persist as part of the document. In all of these cases, the document filters are going to “see” the hidden text.

Remote data is also not a problem. So long as SharePoint attachments, Office 365 / OneDrive, DropBox files and the like present as part of the Windows folder system, the indexer can handle these just like ordinary local files. After recognizing all text and metadata, the indexer stores each unique word and number and the position of each in the data.

Indexing can take a bit. But the result is instant searching across terabytes. And concurrent searching works in multiple different environments, including a classic Windows network environment, a local Intranet server or a SaaS-hosted server. In any of these environments, searching can continue even while indexes automatically update to account for new, modified or deleted files, letting your team continually keep tabs on zombie and other data.

After indexing, dtSearch has over 25 different types of individual and concurrent search options. Search across all data, or limit specific search elements to particular metadata such as email sender or recipient. Search options include basic all words / any words / exact phrase searching, Boolean (and/or/not) searching, proximity searching, concept searching, number searching, numeric range searching, automatic date and date range recognition across popular date formats, and much more. dtSearch can even find and generate hash values across all data or flag any credit card numbers in the data.

Article contributed
by dtSearch®



Another issue that can come up with zombie and other data is camouflaged text like black text against a black background or white text against a white background. Camouflaged text can of course occur if someone is trying to obscure text in a file.

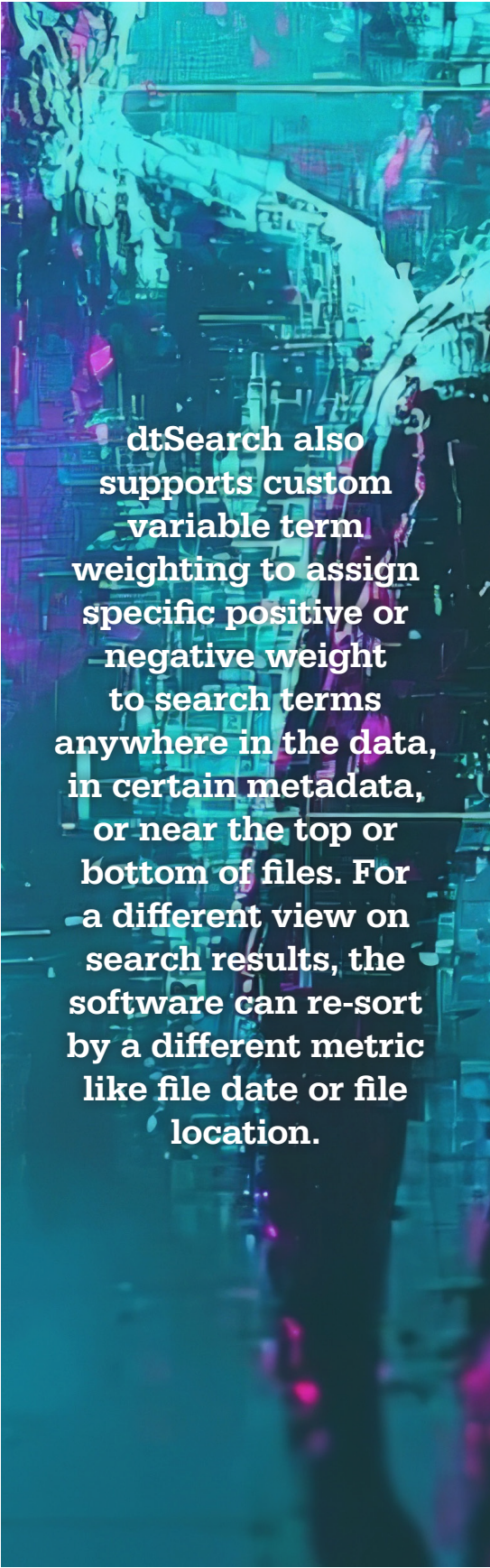
The software also work with multilingual text. Unicode is the standard for all current file types with support for hundreds of international languages. A single file or email can start out in Russian, move to some other European language, then go to double-byte Chinese, Japanese or Korean text, followed by right-to-left text like Arabic or Hebrew and then continue on in English. Unicode and enterprise search will follow the whole progression. Note that Unicode searching also lets you search for specific Unicode emojis, like the Unicode zombie emoji 🧟

Say the word ZOMBIES is misspelled in an email or mis-OCR'ed as ZOMDIES. Fuzzy searching can sift through that, adjusting from 1 to 10 to accommodate various typographical deviations. And it even works with different international languages. After a search, automatic relevancy ranking can sort retrieved files by hit term density and rarity, so files with rarer search terms in denser configurations come up at the top.

dtSearch also supports custom variable term weighting to assign specific positive or negative weight to search terms anywhere in the data, in certain metadata, or near the top or bottom of files. For a different view on search results, the software can re-sort by a different metric like file date or file location. Whatever the sorting, dtSearch shows a full copy of retrieved files with highlighted hits.

dtSearch.com has fully-functional 30-day evaluation enterprise search downloads to get you started on instant concurrent searching across terabytes. Those zombies don't stand a chance!

Article contributed
by dtSearch®



dtSearch also supports custom variable term weighting to assign specific positive or negative weight to search terms anywhere in the data, in certain metadata, or near the top or bottom of files. For a different view on search results, the software can re-sort by a different metric like file date or file location.