

The “Trick” and the “Treat” of Enterprise Search

Article contributed
by [dtSearch®](#)

Looking for Halloween treats? A search engine like dtSearch can instantly search through terabytes for *candy corn or chocolate bars, gummy bears and marshmallows, toffee and not butterscotch, red licorice* within X words of *jelly beans, caramel* appearing X or fewer words before *candy apples*, or any combination of the above. You can even check off concept searching to find *treats* generally. The search engine has multiple different options for relevancy ranking and can display the full text of retrieved files and emails with *treat* hit highlights for convenient navigation. But while treats are one side of the Halloween data equation, tricks are the other.

Like what?

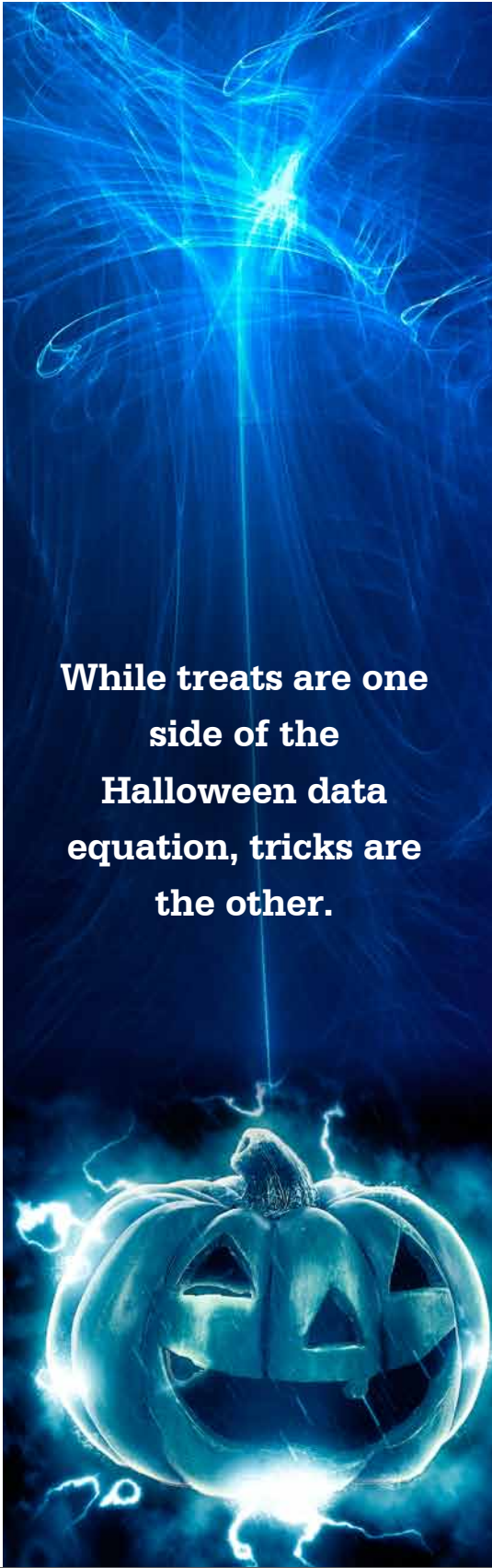
One data trick is that it is easy to save a file with a mismatched file extension, like a PDF with a .DOCX extension or a Microsoft Word document with a .PDF extension. Mismatched file extensions can make it difficult to review documents inside of their native applications. The good news is that the search engine can look right past mismatched file extensions.

How does that work?

The search engine instantly searches terabytes after first indexing the data. Each index can hold up to a terabyte, and there are no limits on the number of indexes the search engine can create and simultaneously search. Indexing is effortless. Simply point to the folders and the like to index and the search engine will take it from there. And here's the key. During indexing, the search engine approaches all files—Microsoft Office, PDFs, emails, web-ready files, etc.—in binary format. The binary format itself has the information about the file type, so the search engine can figure out the applicable file type without reference to the file extension.

So a single index can span multiple data formats, regardless of file extensions?

Yes. Indexed data can include any number of different formats from any number of different data locations. A single search or multiple concurrent searches can run across all indexed data. Instant concurrent searching can even continue while indexes automatically update to reflect new, deleted and modified content.



**While treats are one
side of the
Halloween data
equation, tricks are
the other.**

What is the next trick?

You can't always readily spot all of a file's metadata inside a file's native application. Let's say you are looking at a Word document inside of the Microsoft Word application. Some metadata may require a lot of clicking around before you even know that it is there. But the binary format makes all metadata fully apparent to the search engine. A general indexed search will pick up all hits in the full-text or metadata. Alternatively, the search engine can identify all metadata covered by an index, and let you refine a search or search element to focus in on just that metadata.

What's the next data trick?

While most word processing documents are standalone Microsoft Word files, most spreadsheets are standalone Excel files, etc., it is also possible to have multilevel nested file structures. For example, you can have an email with a compressed ZIP or RAR attachment, and inside that attachment is a PowerPoint with an embedded Excel spreadsheet. If you viewed the presentation file in its native PowerPoint application, by default you may not even see the whole embedded Excel file. However, in binary format, the full embedded structure inside the compressed email attachment is available, letting the search engine seamlessly drill down to the innermost nested data.

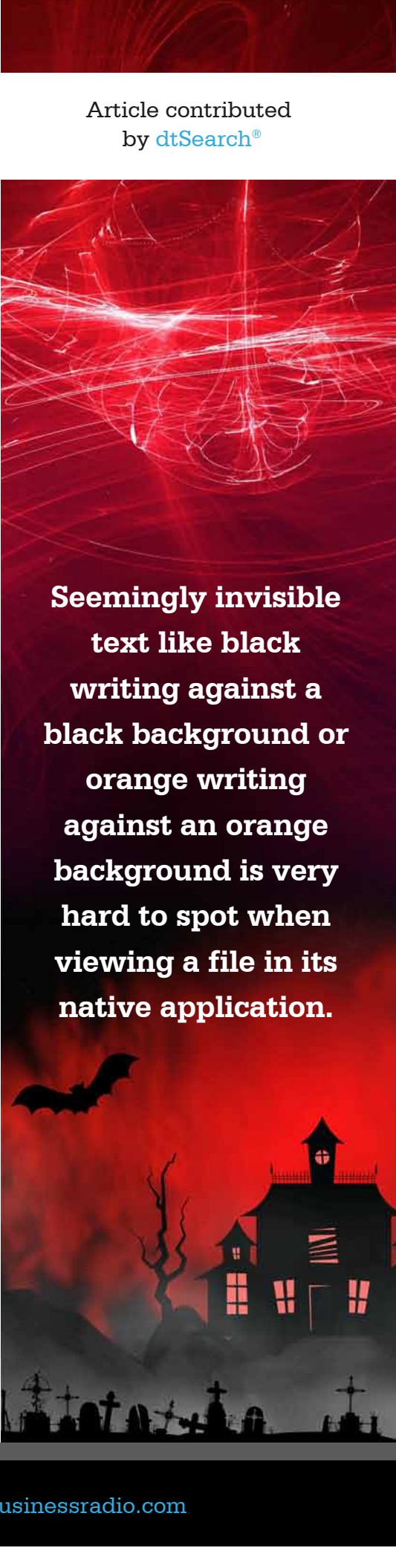
What is the next trick?

PDFs come in two basic flavors. The first flavor is the standard text-based PDFs that you create when you print a file to PDF or OCR an image that has text in it to PDF. Because this PDF is text-based, you can copy and paste passages from the PDF into a different file. The second flavor of PDF is "image only." Looking at PDFs in the file system, image-only PDFs will look indistinguishable from standard text-based PDFs. But when you get inside an image-only PDF, you won't be able to perform basic text functions like copy and paste.

How does a search engine work with these "image only" PDFs?

Because there is no main text in "image only" PDFs, the search engine will by default only be able to index the filename and any metadata. But when the search engine builds its index, it can flag these "image only" PDFs so you know to run them through an OCR program like Adobe Acrobat to turn them into text-based PDFs. At that point, the full text will be accessible just like ordinary PDFs.

Article contributed
by [dtSearch®](#)



**Seemingly invisible
text like black
writing against a
black background or
orange writing
against an orange
background is very
hard to spot when
viewing a file in its
native application.**

What is the next trick?

Seemingly invisible text like black writing against a black background or orange writing against an orange background is very hard to spot when viewing a file in its native application. However, when the search engine approaches a file in its binary format, all text is on the same footing, regardless of contrast with the background color inside a file's native application.

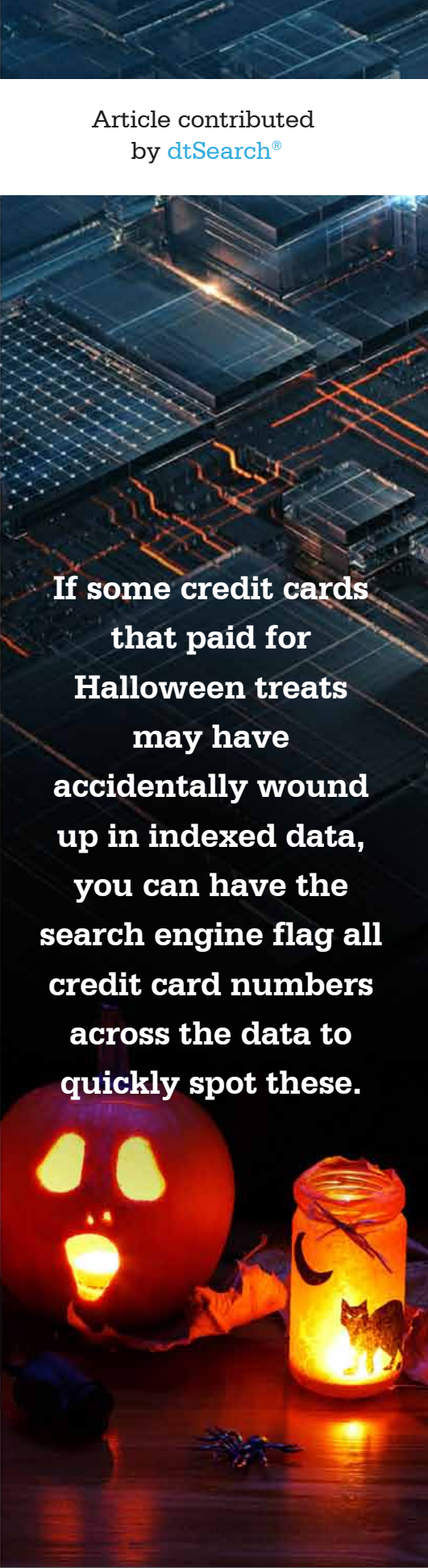
What's next?

The next tricks relate to special search features the search engine can deploy for advanced text recognition. Suppose that *Halloween* is mistyped *Hallomeen* in an email, or mis-OCR'ed as *Hallaween* in a PDF. Fuzzy searching adjustable from 1 to 10 can still find both of those misspellings in a search for Halloween. Other advanced techniques go beyond standard word search, such as regular expression searches, number and numeric range searches, and dates and date range searches even across different formats like October 31, 2022 and 10/31/22. Finally, if some credit cards that paid for Halloween treats may have accidentally wound up in indexed data, you can have the search engine flag all credit card numbers across the data to quickly spot these.

Anything else you'd like to add?

Happy Halloween! And please go to [dtSearch.com](https://dtsearch.com) to download a fully-functional 30-day evaluation version to instantly search for treats and tricks in terabytes of your own enterprise data.

Article contributed
by [dtSearch®](https://dtsearch.com)



**If some credit cards
that paid for
Halloween treats
may have
accidentally wound
up in indexed data,
you can have the
search engine flag all
credit card numbers
across the data to
quickly spot these.**