Like a Black Cat Melting into the Night: A Search Engine's Guide to Text Tricks

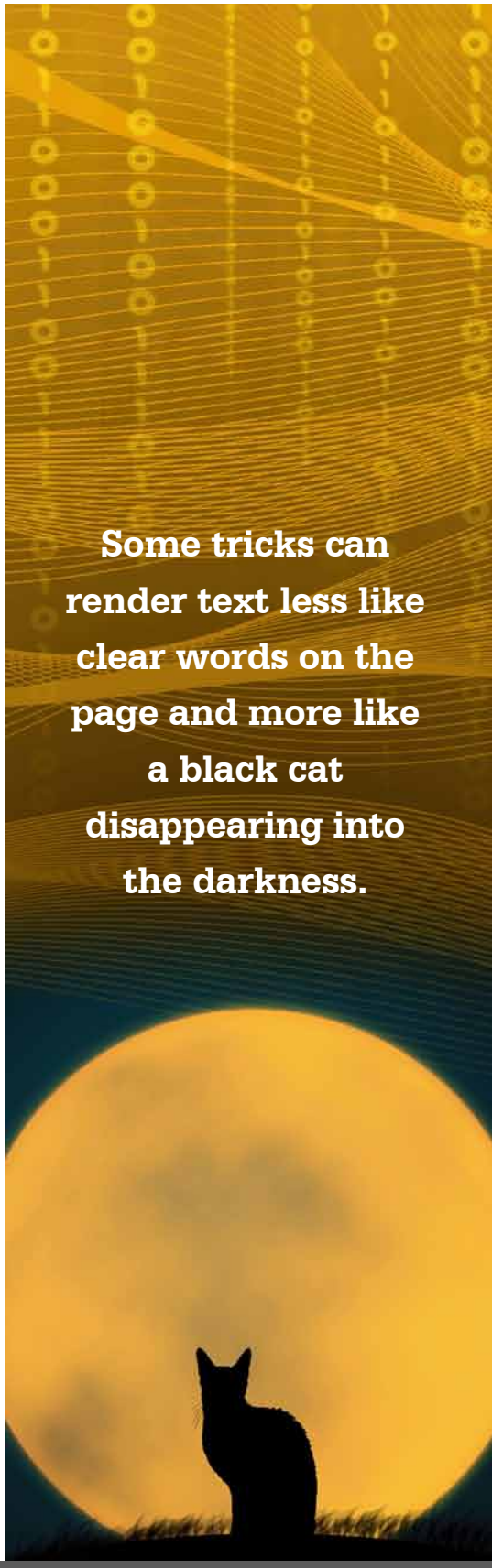
Article contributed by dtSearch®

Most files and emails – Word, Excel, Access, PowerPoint, OneNote, PDF, Outlook/Exchange, etc. – present as easily readable text when you view them in their native applications. But some tricks can render text less like clear words on the page and more like a black cat disappearing into the darkness. This article details several such text tricks and shows how an enterprise search engine like dtSearch® (as distinct from an Internet search engine like Google) can shine a light on them.

Before we illuminate our black cat of text tricks, it is essential to understand how everything fits together under normal circumstances. Suppose you are looking through thousands or even millions of files and emails to see whether they contain one or more of the following terms: costume, candy, or cauldron. If you have unlimited time, you could individually retrieve each file in its native application and scan the text. We'll call this the eyeball method. You could also deploy an enterprise search engine to search through terabytes at once instantly.

An enterprise search engine instantly searches terabytes after first indexing the data. The index itself is just an internal tool holding each unique word, number and location of each in the data. Indexing is not difficult for the end-user; just tell the search engine the folders to cover, and the search engine's indexer does the rest.

In building its index, the search engine approaches each file in its binary format, bypassing the need to pull up each in native application. Looking directly at a file in binary format, it can be hard to make out any words at all in the sea of binary codes. To process the text inside, the search engine must first figure out the applicable parsing specification. Parsing specifications can be hundreds of pages long and vary greatly among different file formats. Matching the right specification to the right file type is, therefore, critical.



Article contributed
by dtSearch®

In addition to its much faster speed, indexed search can go far beyond the eyeball method in terms of the search types it enables. In fact, indexed search makes available over 25 different search features, including complex phrases, Boolean and proximity expressions: (candy corn or black cat) and (cauldron w/ 12 witch costume) and not Christmas. For multilingual text, indexed search supports any of the hundreds of international languages that Unicode supports. Beyond word searches, a search engine can also look for specific numbers and numeric ranges, as well as dates or date, ranges spanning different date formats like 10/31/22 vs. October 31, 2022. A search engine can even flag any credit card numbers in data.

Finally, unlike the eyeball method, indexed search enables multiuser concurrent queries across a network or online. Running from an "on-premises" web server or from the cloud such as on Azure or AWS, an online search can execute in a stateless manner with no built-in limits on the number of concurrent search threads. Concurrent search can continue even while an index updates to reflect files that have been added, deleted, or modified since the last index build, so there is no "downtime."

Now for our black cat tricks.

Trick 1: Black Text against a Black Background, White Text against a White Background

Talk about a black cat melting into the night. This type of text is very hard to spot when viewing a file in its native application. However, because a search engine parses files in their binary formats, black against black text and the like is on the same plane as any other text to a search engine.

Trick 2: Deeply Buried Metadata

A native application view of files can hide certain metadata, so it can take a lot of clicking around even to know that it is there. But all metadata is fully apparent in the binary format view of a file that a search engine sees in indexing, and thus fully searchable.

Before we illuminate our black cat of text tricks, it is essential to understand how everything fits together under normal circumstances.

Trick 3: Multilevel Nested File Structure

Sometimes files don't present as standalone items. You can have a Word document with an Excel spreadsheet buried inside, and the same duo can be part of a larger nested structure, such as an email with a ZIP or RAR attachment. When you view a nested file inside the outer file's native application, sometimes just a fraction of the embedded file is visible by default. But when a search engine goes through files in binary format, it sees everything. Plus, a search engine can even let you individually copy a file out of a larger ZIP or RAR archive or an individual email out of a larger email archive.

Trick 4: Files With Mismatched Extensions

It is all too easy to save an Access database with an Excel .XLSX extension or a Word file with a .PDF extension. However, a search engine can go directly into the binary format to determine the correct file type, bypassing the file extension entirely.

Trick 5: Image-Only PDFs

Sometimes a PDF can have what looks like ordinary text. But then, when you try to copy and paste it, you get no text out. (This example is the opposite of the others, in that the text may be clearly visible inside a PDF viewer like Adobe Reader, but not otherwise accessible.) But a search engine can flag image-only PDFs when it builds its index, letting you know that you need to run them through an OCR program like Adobe Acrobat to turn them into "searchable image" PDFs.

Trick 6: Mistypings

Dipping into that bag of treats can lead to sticky fingers resulting in even more typos than usual. Fuzzy searching adjustable from 1 to 10 can find a word even if it is misspelled. So if Halloween is mistyped as Hallomeen in an email, a search engine can still find that in a search for Halloween with a low level of fuzziness.

Keep an eye out for that black cat, and Happy Halloween!

Article contributed by dtSearch®