

# For the Love of Text Retrieval Software: Find What You're Looking for This Valentine's

*Text retrieval software instantly searches the contents of documents, PDFs, emails, and even online data and databases, whether from an individual PC, across a network, or from an Internet or Intranet site and finds the correct output.*

An individual or multiple concurrent users can search terabytes of online and offline data for specific Unicode words, phrases, numbers and even emojis with text retrieval software. Let's look at the process and scope of text retrieval and how it could benefit you.

## 1. Unicode

The first step is understanding what makes up today's office documents, emails, databases and online data. What fills these data types is Unicode characters as defined by the Unicode Consortium. Some Unicode subranges cover English and European character sets, some cover double-byte character sets like Chinese, Japanese, and Korean, and others cover "left to right" text like Arabic and Hebrew. There is even a subrange for ancient Egyptian hieroglyphics. Beyond classical alphabets, Unicode also includes numbers, symbols and emojis.

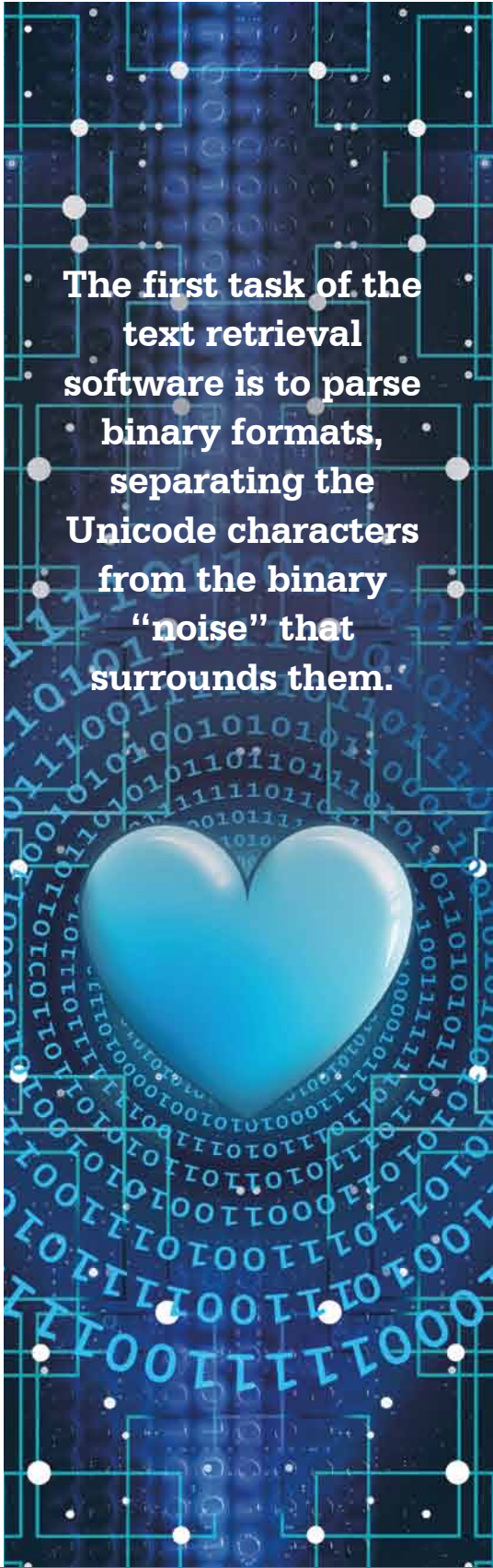
When you think of a Word file, an Excel spreadsheet, an Access database, a PowerPoint, a OneNote file, a PDF, or an email, you usually think about what these look like inside their native applications. It is easy to see the Unicode text when viewing a spreadsheet in Excel or a PDF in Adobe Reader. While this approach works well on a file-by-file basis, if you need to review millions of items, pulling up each in its native application is not a feasible approach.

## 2. Binary formats

Rather than retrieving each file in its native applications, text retrieval software goes straight to the binary format of everything. However, Unicode text easily readable in its native application can look nothing short of gibberish in binary form. The first task of the text retrieval software is to parse these binary formats, separating the Unicode characters from the binary "noise" that surrounds them.

To parse a binary file correctly, the text retrieval software needs to figure out the suitable file format standard to be applied, as these differ dramatically among different file formats. Applying an email parsing specification to a OneNote file will get you nowhere. You may think that the text retrieval software can just look at the file format extension and that a .PDF would indicate a PDF file, and a .DOCX would suggest a Word document. However, it is all too easy to save a PDF file with a .DOCX extension and a Word document with a .PDF extension.

Article contributed  
by [dtSearch®](#)



**The first task of the text retrieval software is to parse binary formats, separating the Unicode characters from the binary "noise" that surrounds them.**

To figure out the correct file format specification to apply, the text retrieval software has to look inside each binary file. Moreover, it is not just the main body text of each file that the software needs to identify but also all metadata. The component of the text retrieval software that makes this file format identification and parsing is called the document filter.

### 3. Indexing

After parsing the text and metadata of a binary format, the next step for text retrieval software is to enable instant single or multithreaded searching across terabytes is to index the data. An index stores each unique word, number and emoji and its location in the data. But isn't indexing a lot of work, you ask? Yes, but only for the software and not for individuals using the software. The end-user has only to point to the various folders and websites to index.

The Windows Task Scheduler can even launch indexing tasks on its own. And while indexing is resource-intensive, after indexing, the search operates with a very light footprint, allowing multiple search threads to operate without affecting each other across a network or online. Even updating an index to reflect the addition of new files and additional information does not impact concurrent searching.

### 4. Searching

So, what can you search for? At one end of the spectrum, you can enter a natural language unstructured search request like *valentine's* 🍷 and see what comes up. At the other end, you could also enter a highly structured search request like: ( 🍷 and "valentine's day") and (candy w/22 of "easter egg") and not (spring fling or Florida vacation).

You can make the structured search even more precise, such as limiting (candy w/ 22 of "easter egg") to specific metadata or using directed proximity search to find *candy* only if it appears a certain number of words before *easter egg*.

Fuzzy searching adjustable from 1 to 10 sifts through potential misspellings like finding in a search for *valentine's* the typo *valentime's* in an email or the OCR error *valontime's* day in a PDF. Concept searching locates synonyms in the existing thesaurus or self-defines it, like *chocolates* for *candy*. Whether your search is structured or unstructured, the text retrieval software offers multiple options for relevancy-ranking and displays a full copy of retrieved items with highlighted hits.

### Find What You Seek

Beyond word and emoji searching, numeric searches include searching for number patterns, specific numbers or number ranges, and specific dates or date ranges. Advanced forensics-oriented users can also generate hash values for each indexed file and search on those hash values. Text retrieval software can even identify any credit card numbers that may appear in the Unicode text. There is much scope for text retrieval at present, and companies can leverage many benefits, especially when it comes to sifting through the massive amounts and diverse formats of data that we generate today.

So happy Valentine's Day, and here's to finding all that you're searching for in your data!

Article contributed  
by [dtSearch®](#)



**There is much  
scope for text  
retrieval at  
present, and  
companies can  
leverage many  
benefits,  
especially when  
it comes to  
sifting through  
the massive  
amounts and  
diverse formats  
of data that we  
generate today.**