Globalizing Text Search

Article contributed
by dtSearch®

We can't all speak the same language, but we can all search text across the world's Unicode-based languages. And there are hundreds of these. According to the Unicode Consortium, which defines the Unicode standard: "The Unicode Consortium is the standards body for the internationalization of software and services. Deployed on more than 20 billion devices around the world, Unicode also provides the solution for internationalization and the architecture to support localization."

Executive summary. You can learn more about the Unicode Consortium and Unicode generally at Unicode.org. But the executive summary is that Unicode takes the world's languages – English; other European languages; double-byte Chinese, Japanese and Korean; right-to-left Arabic and Hebrew; and even ancient languages like Egyptian hieroglyphics – and turns these into a standard text representation. Each year, the Unicode Consortium also comes out with a new series of emojis.

Microsoft Office documents, emails, PDFs, etc. all leverage Unicode text representation. A Unicode subrange is a range of characters representing text in a specific language or family of languages. A single document or email can have numerous subranges, allowing the same document to flip from say English to Russian to Korean and back to English.

Unicode caution. A big caution with Unicode is that many letters and numbers in one character set can look visually indistinguishable from letters and numbers in other character sets. This makes it easy for a link in an email or document or online to look "legit" but actually take you to a destination that is very different from what you would expect. If you see a link, unless you are 100% sure that the source is verified, don't click on that link. Instead, re-enter the name of the website yourself in your browser to make sure that it is taking you to the actual site and not a phony Unicode look-alike.

Parsing Unicode text. A search engine parses the Unicode in documents and emails in their binary format where it can follow the progression of different languages. But first, to correctly parse a file, a search engine has to identify the exact right file format specification to apply. File format specifications can be hundreds of pages long, and applying an incorrect file format specification to a binary format can result in text parsing that is effectively gibberish.

Microsoft Word, Access, Excel, PowerPoint, OneNote, PDF, Outlook/Exchange, etc. each have unique and very different file format specifications. The document filters have to correctly figure out the file type, and in many cases, the correct file type version as these too can result in very different parsing specifications. After applying the correct parsing specification, the indexer can proceed through the Unicode text, recording each unique Unicode word or number and its location.

Indexed search. Each index can hold up to a terabyte of Unicode text and there are no limits on the number of terabyte-size indexes the search engine can create and simultaneously search. Indexed search is typically instantaneous whether operating on a single search thread or multiple

We can't all speak the same language, but we can all search text across the world's Unicode-based languages.

Reprinted with permission of USA Business Radio, usabusinessradio.com

search threads. Online, search can operate statelessly, allowing each concurrent search thread to proceed independently without slowing down other search threads. Index updates to accommodate new, modified or deleted items do not block out searching. Indexes can automatically update, for example through the Windows Task Scheduler, as often as desired with individual or multiuser concurrent searching proceeding unaffected.

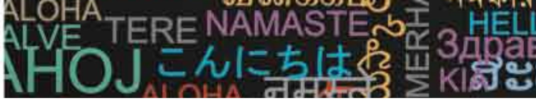
Indexed search across Unicode is very comprehensive.

- The file parsing and indexing works even if a file has a misapplied file extension, like a PDF file saved with a .DOCX extension or vice versa. The search engine looks inside the binary format to determine the correct file parsing standard to apply so a misapplied file extension is not an issue.
- No matter how obscure certain metadata may be inside the application that created a file, all metadata is equally accessible in binary format.
- Visually tricky text in a generating application like red text against a red background or green text against a green background is as readily available as any other text in binary format.
- Finally, the binary format allows the search engine to seamlessly parse multlilevel nested formats, like an email with a ZIP or RAR file containing a PowerPoint with an embedded Excel spreadsheet.

Unicode search options. A search engine like dtSearch has over 25 different types of search options. Of course, numeric-oriented searching like numeric range, date range, credit card identification, file hash value generation and search are generally language neutral. And most word-oriented search types are as well, including unstructured natural language searching, Boolean and/or/not searching, phrase searching, proximity searching, metadata-specific searching, searches for specific Unicode emojis, etc. Even fuzzy searching adjustable from 1 to 10 to sift through minor spelling or OCR errors works regardless of the underlying Unicode language.

While most search options are universal, a few are language-specific. Stemming which looks for different endings on the same root word like *jump, jumping, jumped,* etc. is different for different languages. And noise words which are words you typically want to overlook in indexing as they tend to just clutter things up like *the* and *an* for English are different for different languages. But dtSearch now has these noise word lists and stemming rules pre-packaged and selectable via dropdown for over 25 European languages.

About dtSearch®. dtSearch has enterprise and developer products that run “on premises” or on cloud platforms to instantly search terabytes of “Office” files, PDFs, emails along with nested attachments, databases and online data. Because dtSearch can instantly search terabytes with over 25 different search features, many dtSearch customers are Fortune 100 companies and government agencies. But anyone with lots of data to search can download a fully-functional 30-day evaluation copy from dtSearch.com



Article contributed by dtSearch®

Unicode takes the world’s languages – English; other European languages; double-byte Chinese, Japanese and Korean; right-to-left Arabic and Hebrew; and even ancient languages like Egyptian hieroglyphics – and turns these into a standard text representation.

Reprinted with permission of USA Business Radio, usabusinessradio.com