On locale-aware substring matching, either case-sensitive or case-insensitive

devblogs.microsoft.com/oldnewthing/20241101-00

November 1, 2024

Say you want to do a case-insensitive substring search in a locale-aware manner. For example, maybe you have a list of names, and you want the user to be able to search for them by typing any name fragment. A search for “ber” could find “Bert” as well as “Roberta”.

I’ve seen people solve this problem by converting the string to lowercase, and then doing a code unit-based substring search. This technique doesn’t work for multiple reasons.

One reason is that some languages (like English) do not consider diacritics significant in collation. The word *naive* and *naïve* are considered equivalent for searching purposes. But a code unit substring search considers them different.

For languages in which diacritics are significant, you have the problem of composed and decomposed characters. For example, the lowercase a with ring in the Swedish word *någon* could be represented either as

- Two code points: U+0061 (LATIN SMALL LETTER A) followed by U+030A (COMBINING RING ABOVE), or
- A single code point: U+03E5 (LATIN SMALL LETTER A WITH RING ABOVE)

The number of possibilities increases if you have characters with multiple diacritics. And then you also have ligatures, where the *fi* “fi” ligature is equivalent to two separate characters *f* and *i*.

So what’s the right thing to do?

In Windows, you can use the `FindNLSStringEx` function to do a locale-aware substring search. Use the `LINGUISTIC_IGNOREDIACRITIC` flag to say that you want to honor diacritics only when they are significant to the locale.[1] (A better name would have been `LINGUISTIC_IGNOREINSIGNIFICANTDIACRITICS`.)

On other platforms, and even on Windows,[2] you can use the ICU library’s string search service and search with primary weight. (Primary weight honors diacritics which are significant to the locale.)

Bonus reading: A popular but wrong way to convert a string to uppercase or lowercase. What has case distinction but is neither uppercase nor lowercase?

[1] Throw in one of the `IGNORECASE` flags if you want a case-sensitive substring search.

[2] The Windows globalization team now recommends that people use ICU, which has been part of Windows since Windows 10 version 1703 (build 15063). More details and gotchas here.