ustrto() — Convert a Unicode string to or from a string in a specified encoding

**Description**

ustrto(\(s, \text{enc}, \text{mode}\)) converts the Unicode string \(s\) to a string encoded in \text{enc}. Any invalid UTF-8 sequence in \(s\) is replaced with a Unicode replacement character \texttt{\ufffd}. \text{mode} controls how unsupported Unicode characters in the encoding \text{enc} are handled. The possible values for \text{mode} are:

- 1, which substitutes any unsupported characters with the \text{enc}’s substitution string;
- 2, which skips any unsupported characters;
- 3, which stops at the first unsupported character and returns an empty string; or
- 4, which replaces any unsupported character with an escaped hex digit sequence \texttt{\uhhhh} or \texttt{\Uhhhhhhhh}. The hex digit sequence contains either four or eight hex digits depending on the Unicode character’s code-point value. Any other values are treated as 1.

ustrfrom(\(s, \text{enc}, \text{mode}\)) converts a string \(s\) in encoding \text{enc} to a UTF-8 encoded Unicode string. \text{mode} controls how invalid byte sequences in \(s\) are handled. The possible values for \text{mode} are:

- 1, which substitutes an invalid byte sequence with a Unicode replacement character \texttt{\ufffd};
- 2, which skips any invalid byte sequences;
- 3, which stops at the first invalid byte sequence and returns an empty string; or
- 4, which replaces any byte in an invalid sequence with an escaped hex digit sequence \texttt{\%Xhh}. Any other values are treated as 1.

When arguments are not scalar, ustrto() returns element-by-element results.

**Syntax**

```
string matrix ustrto(string matrix s, string scalar enc, real scalar mode)

string matrix ustrfrom(string matrix s, string scalar enc, real scalar mode)
```

**Remarks and examples**

Type unicode encoding list to list available encodings. See [U] 12.4.2.3 Encodings and see the unicode encoding command in [D] unicode for details.

The substitution character for both ASCII and Latin-1 encoding is char(26)

A good use of \text{mode}=4 (escape) is to check what invalid bytes a Unicode string \texttt{ust} contains by examining the result of ustrfrom(\texttt{ust}, "utf-8", 4).
Conformability

\[
\text{ustrto}(s, \text{enc}, \text{mode}), \text{ustrfrom}(s, \text{enc}, \text{mode}):
\]

\[
s: \quad r \times c
\]

\[
\text{enc}: \quad 1 \times 1
\]

\[
\text{mode}: \quad 1 \times 1
\]

\[
\text{result}: \quad r \times c
\]

Diagnostics

\text{ustrto}(s, \text{enc}, \text{mode}) \text{ and } \text{ustrfrom}(s, \text{enc}, \text{mode}) \text{ return an empty string if an error occurs.}

Also see

[M-5] \text{ustrfix()} — Replace invalid UTF-8 sequences in Unicode string

[M-5] \text{ustrunescape()} — Convert escaped hex sequences to Unicode strings

[M-4] \text{String} — String manipulation functions

[U] 12.4.2 Handling Unicode strings

[U] 12.4.2.3 Encodings