NAME

gendict − Compiles word list into ICU string trie dictionary

SYNOPSIS

gendict [**−−uchars | −−bytes −−transform** *transform*] [**−h**, **−?**, **−−help**] [**−V**, **−−version**] [**−c**, **−−copyright**] [**−v**, **−−verbose**] [**−i**, **−−icudatadir** *directory*] *input-file output−file*

DESCRIPTION

gendict reads the word list from *dictionary-file* and creates a string trie dictionary file. Normally this data file has the **.dict** extension.

Words begin at the beginning of a line and are terminated by the first whitespace. Lines that begin with whitespace are ignored.

OPTIONS

−h, **−?**, **−−help**
Print help about usage and exit.

−V, **−−version**
Print the version of **gendict** and exit.

−c, **−−copyright**
Embeds the standard ICU copyright into the *output-file*.

−v, **−−verbose**
Display extra informative messages during execution.

−i, **−−icudatadir** *directory*
Look for any necessary ICU data files in *directory*. For example, the file **pnames.icu** must be located when ICU's data is not built as a shared library. The default ICU data directory is specified by the environment variable **ICU_DATA**. Most configurations of ICU do not require this argument.

−−uchars
Set the output trie type to UChar. Mutually exclusive with **--bytes.**

−−bytes
Set the output trie type to Bytes. Mutually exclusive with **--uchars.**

−−transform
Set the transform type. Should only be specified with **--bytes.** Currently supported transforms are: **offset-<hex-number>,** which specifies an offset to subtract from all input characters. It should be noted that the offset transform also maps U+200D to 0xFF and U+200C to 0xFE, in order to offer compatibility to languages that require these characters. A transform must be specified for a bytes trie, and when applied to the non-value characters in the *input-file* must produce output between 0x00 and 0xFF.

input−file
The source file to read.

output−file
The file to write the output dictionary to.

CAVEATS

The *input-file* is assumed to be encoded in UTF-8. The integers in the *input-file* that are used as values must be made up of ASCII digits. They may be specified either in hex, by using a 0x prefix, or in decimal. Either **--bytes** or **--uchars** must be specified.

ENVIRONMENT

ICU_DATA
Specifies the directory containing ICU data. Defaults to **${prefix}/share/icu/63.2/**. Some tools in ICU depend on the presence of the trailing slash. It is thus important to make sure that it is present if **ICU_DATA** is set.

AUTHORS

Maxime Serrano

VERSION

1.0

COPYRIGHT

SEE ALSO

http://www.icu-project.org/userguide/boundaryAnalysis.html