NAME

utfpatgen – generate patterns for TeX hyphenation

SYNOPSIS

utfpatgen *dictionary_file pattern_file patout_file translate_file*

DESCRIPTION

UTFpatgen is an extension to **patgen**(1) for generating patterns from large input alphabets, with an extended hyphenation level range and native dynamic memory management.

The program reads a *dictionary_file* containing a list of hyphenated words and a *pattern_file* containing previously-generated patterns (if any) for a particular language (not a complete TEX source file; see below), and produces the *patout_file* with (previously- plus newly-generated) hyphenation patterns for that language.

The *translate_file* defines language specific values for the parameters *left_hyphen_min* and *right_hyphen_min* used by TEX's hyphenation algorithm and the external representation of the lower and upper case version(s) of all 'letters' of that language.

Further details of the pattern generation process such as hyphenation levels and pattern lengths are requested interactively from the user's terminal. Optionally, *UTFpatgen* creates a new dictionary file **pattmp.***n* showing the good and bad hyphens found by the generated patterns, where *n* is the highest hyphenation level.

All filenames must be complete; no adding of default extensions or path searching is done.

INPUT FORMATS

Letters

UTFpatgen is able to process any UTF-8 encoded character, or more generally, any encoding that is prefix-free (no letter is a prefix of another) and does not use the '0xFF' byte, which has a special meaning in *UTFpatgen*), described next:

Levels and weights

Non-character parts of the text, such as hyphenation levels or weights, should be represented as a 2-byte sequence '0xFF <value>'. If a file uses the **patgen**(1) encoding (ASCII numerals), we recommend using **sed**(1) for conversion.

File formats

The formats and conventions required in the 4 input files (*dictionary_file, pattern_file, patout_file, translate_file*) are identical to those in **patgen**(1) with the only exception of level and weight encoding described earlier.

SEE ALSO

Frank Liang, *Word hy-phen-a-tion by com-puter*, STAN-CS-83-977, Stanford University Ph.D. thesis, 1983, http://tug.org/docs/liang.

Donald E. Knuth, *The TEXbook*, Addison-Wesley, Appendix H.

https://ctan.org/pkg/patgen
: The original patgen program, by Frank Liang, with system updates by Peter Breitenlohner.

https://ctan.org/pkg/hyph-utf8
: Collected hyphenation patterns for many languages in many formats.

https://ctan.org/tex-archive/language/
: General CTAN directory for patterns and support for many other languages.

https://tug.org/TUGboat/Contents/listkeyword.html#CatTAGMultilingualDocumentProcessing
: *TUGboat* articles on hyphenation and other aspects of language-specific document processing.

AUTHORS

Ondřej Metelka

https://ctan.org/pkg/utfpatgen