NAME

unicharset_extractor − Reads box or plain text files to extract the unicharset.

SYNOPSIS

unicharset_extractor [−−output_unicharset filename] [−−norm_mode mode] box_or_text_file [...]

Where mode means: 1=combine graphemes (use for Latin and other simple scripts) 2=split graphemes (use for Indic/Khmer/Myanmar) 3=pure unicode (use for Arabic/Hebrew/Thai/Tibetan)

DESCRIPTION

Tesseract needs to know the set of possible characters it can output. To generate the unicharset data file, use the unicharset_extractor program on training pages bounding box files or a plain text file:

unicharset_extractor fontfile_1.box fontfile_2.box ...

The unicharset will be put into the file *./unicharset* if no output filename is provided.

NOTE Use the appropriate norm_mode based on the language.

SEE ALSO

tesseract(1), unicharset(5)

https://github.com/tesseract−ocr/tesseract/wiki/TrainingTesseract

HISTORY

unicharset_extractor first appeared in Tesseract 2.00.

COPYING

AUTHOR

The Tesseract OCR engine was written by Ray Smith and his research groups at Hewlett Packard (1985−1995) and Google (2006−present).