NAME

shapeclustering − shape clustering training for Tesseract

SYNOPSIS

shapeclustering −D *output_dir* −U *unicharset* −O *mfunicharset* −F *font_props* −X *xheights FILE...*

DESCRIPTION

shapeclustering(1) takes extracted feature .tr files (generated by tesseract(1) run in a special mode from box files) and produces a file **shapetable** and an enhanced unicharset. This program is still experimental, and is not required (yet) for training Tesseract.

OPTIONS

−U *FILE*

The unicharset generated by unicharset_extractor(1).

−D *dir*

Directory to write output files to.

−F *font_properties_file*

(Input) font properties file, where each line is of the following form, where each field other than the font name is 0 or 1:

'font_name' 'italic' 'bold' 'fixed_pitch' 'serif' 'fraktur'

−X *xheights_file*

(Input) x heights file, each line is of the following form, where xheight is calculated as the pixel x height of a character drawn at 32pt on 300 dpi. [That is, if base x height + ascenders + descenders = 133, how much is x height?]

'font_name' 'xheight'

−O *FILE*

The output unicharset that will be given to combine_tessdata(1).

SEE ALSO

tesseract(1), cntraining(1), unicharset_extractor(1), combine_tessdata(1), unicharset(5)

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

COPYING

AUTHOR

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