NAME

mftraining − feature training for Tesseract

SYNOPSIS

mftraining −U *unicharset* −O *lang.unicharset FILE...*

DESCRIPTION

mftraining takes a list of .tr files, from which it generates the files **inttemp** (the shape prototypes), **shapetable**, and **pffmtable** (the number of expected features for each character). (A fourth file called Microfeat is also written by this program, but it is not used.)

OPTIONS

−U *FILE*

(Input) The unicharset generated by unicharset_extractor(1)

−F *font_properties_file*

(Input) font properties file, 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*

−D *dir*

Directory to write output files to.

−O *FILE*

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

SEE ALSO

tesseract(1), cntraining(1), unicharset_extractor(1), combine_tessdata(1), shapeclustering(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).