NAME

combine_lang_model − generate starter traineddata

SYNOPSIS

combine_lang_model −−input_unicharset *filename* −−script_dir *dirname* −−output_dir *rootdir* −−lang *lang* [−−lang_is_rtl] [pass_through_recoder] [−−words file −−puncs file −−numbers file]

DESCRIPTION

combine_lang_model(1) generates a starter traineddata file that can be used to train an LSTM−based neural network model. It takes as input a unicharset and an optional set of wordlists. It eliminates the need to run set_unicharset_properties(1), wordlist2dawg(1), some non−existent binary to generate the recoder (unicode compressor), and finally combine_tessdata(1).

OPTIONS

−−lang lang
The language to use. Tesseract uses 3−character ISO 639−2 language codes. (See LANGUAGES)

−−script_dir PATH
Directory name for input script unicharsets. It should point to the location of langdata (github repo) directory. (type:string default:)

−−input_unicharset FILE
Unicharset to complete and use in encoding. It can be a hand−created file with incomplete fields. Its basic and script properties will be set before it is used. (type:string default:)

−−lang_is_rtl BOOL
True if language being processed is written right−to−left (eg Arabic/Hebrew). (type:bool default:false)

−−pass_through_recoder BOOL
If true, the recoder is a simple pass−through of the unicharset. Otherwise, potentially a compression of it by encoding Hangul in Jamos, decomposing multi−unicode symbols into sequences of unicodes, and encoding Han using the data in the radical_table_data, which must be the content of the file: langdata/radical−stroke.txt. (type:bool default:false)

−−version_str STRING
An arbitrary version label to add to traineddata file (type:string default:)

−−words FILE
(Optional) File listing words to use for the system dictionary (type:string default:)

−−numbers FILE
(Optional) File listing number patterns (type:string default:)

−−puncs FILE
(Optional) File listing punctuation patterns. The words/puncs/numbers lists may be all empty. If any are non−empty then puncs must be non−empty. (type:string default:)

−−output_dir PATH
Root directory for output files. Output files will be written to <output_dir>/<lang>/<lang>.* (type:string default:)

HISTORY

combine_lang_model(1) was first made available for tesseract4.00.00alpha.

RESOURCES

Main web site: **https://github.com/tesseract−ocr** Information on training tesseract LSTM: **https://github.com/tesseract−ocr/tesseract/wiki/TrainingTesseract−4.00**

SEE ALSO

tesseract(1)

COPYING

AUTHOR

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