NAME

combine_tessdata − combine/extract/overwrite/list/compact Tesseract data

SYNOPSIS

combine_tessdata [*OPTION*] *FILE...*

DESCRIPTION

combine_tessdata(1) is the main program to combine/extract/overwrite/list/compact tessdata components in [lang].traineddata files.

To combine all the individual tessdata components (unicharset, DAWGs, classifier templates, ambiguities, language configs) located at, say, /home/$USER/temp/eng.* run:

combine_tessdata /home/$USER/temp/eng.

The result will be a combined tessdata file /home/$USER/temp/eng.traineddata

Specify option −e if you would like to extract individual components from a combined traineddata file. For example, to extract language config file and the unicharset from tessdata/eng.traineddata run:

combine_tessdata −e tessdata/eng.traineddata \
/home/$USER/temp/eng.config /home/$USER/temp/eng.unicharset

The desired config file and unicharset will be written to /home/$USER/temp/eng.config /home/$USER/temp/eng.unicharset

Specify option −o to overwrite individual components of the given [lang].traineddata file. For example, to overwrite language config and unichar ambiguities files in tessdata/eng.traineddata use:

combine_tessdata −o tessdata/eng.traineddata \
/home/$USER/temp/eng.config /home/$USER/temp/eng.unicharambigs

As a result, tessdata/eng.traineddata will contain the new language config and unichar ambigs, plus all the original DAWGs, classifier templates, etc.

Note: the file names of the files to extract to and to overwrite from should have the appropriate file suffixes (extensions) indicating their tessdata component type (.unicharset for the unicharset, .unicharambigs for unichar ambigs, etc). See k*FileSuffix variable in ccutil/tessdatamanager.h.

Specify option −u to unpack all the components to the specified path:

combine_tessdata −u tessdata/eng.traineddata /home/$USER/temp/eng.

This will create /home/$USER/temp/eng.* files with individual tessdata components from tessdata/eng.traineddata.

OPTIONS

−c *.traineddata FILE...*: Compacts the LSTM component in the .traineddata file to int.

−d *.traineddata FILE...*: Lists directory of components from the .traineddata file.

−e *.traineddata FILE...*: Extracts the specified components from the .traineddata file

−o *.traineddata FILE...*: Overwrites the specified components of the .traineddata file with those provided on the command line.

−u *.traineddata PATHPREFIX* Unpacks the .traineddata using the provided prefix.

CAVEATS

Prefix refers to the full file prefix, including period (.)

COMPONENTS

The components in a Tesseract lang.traineddata file as of Tesseract 4.0 are briefly described below; For more information on many of these files, see
https://github.com/tesseract−ocr/tesseract/wiki/TrainingTesseract and
https://github.com/tesseract−ocr/tesseract/wiki/TrainingTesseract−4.00

lang.config
: (Optional) Language−specific overrides to default config variables. For 4.0 traineddata files, lang.config provides control parameters which can affect layout analysis, and sub−languages.

lang.unicharset
: (Required − 3.0x legacy tesseract) The list of symbols that Tesseract recognizes, with properties. See unicharset(5).

lang.unicharambigs
: (Optional − 3.0x legacy tesseract) This file contains information on pairs of recognized symbols which are often confused. For example, *rn* and *m*.

lang.inttemp
: (Required − 3.0x legacy tesseract) Character shape templates for each unichar. Produced by mftraining(1).

lang.pffmtable
: (Required − 3.0x legacy tesseract) The number of features expected for each unichar. Produced by mftraining(1) from **.tr** files.

lang.normproto
: (Required − 3.0x legacy tesseract) Character normalization prototypes generated by cntraining(1) from **.tr** files.

lang.punc−dawg
: (Optional − 3.0x legacy tesseract) A dawg made from punctuation patterns found around words. The "word" part is replaced by a single space.

lang.word−dawg
: (Optional − 3.0x legacy tesseract) A dawg made from dictionary words from the language.

lang.number−dawg
: (Optional − 3.0x legacy tesseract) A dawg made from tokens which originally contained digits. Each digit is replaced by a space character.

lang.freq−dawg
: (Optional − 3.0x legacy tesseract) A dawg made from the most frequent words which would have gone into word−dawg.

lang.fixed−length−dawgs
: (Optional − 3.0x legacy tesseract) Several dawgs of different fixed lengths — useful for languages like Chinese.

lang.shapetable
: (Optional − 3.0x legacy tesseract) When present, a shapetable is an extra layer between the character classifier and the word recognizer that allows the character classifier to return a collection of unichar ids and fonts instead of a single unichar−id and font.

lang.bigram−dawg
: (Optional − 3.0x legacy tesseract) A dawg of word bigrams where the words are separated by a space and each digit is replaced by a *?*.

lang.unambig−dawg

(Optional – 3.0x legacy tesseract) .

lang.params−model
(Optional – 3.0x legacy tesseract) .

lang.lstm
(Required – 4.0 LSTM) Neural net trained recognition model generated by lstmtraining.

lang.lstm−punc−dawg
(Optional – 4.0 LSTM) A dawg made from punctuation patterns found around words. The "word" part is replaced by a single space. Uses lang.lstm−unicharset.

lang.lstm−word−dawg
(Optional – 4.0 LSTM) A dawg made from dictionary words from the language. Uses lang.lstm−unicharset.

lang.lstm−number−dawg
(Optional – 4.0 LSTM) A dawg made from tokens which originally contained digits. Each digit is replaced by a space character. Uses lang.lstm−unicharset.

lang.lstm−unicharset
(Required – 4.0 LSTM) The unicode character set that Tesseract recognizes, with properties. Same unicharset must be used to train the LSTM and build the lstm−*−dawgs files.

lang.lstm−recoder
(Required – 4.0 LSTM) Unicharcompress, aka the recoder, which maps the unicharset further to the codes actually used by the neural network recognizer. This is created as part of the starter traineddata by combine_lang_model.

lang.version
(Optional) Version string for the traineddata file. First appeared in version 4.0 of Tesseract. Old version of traineddata files will report Version string:Pre−4.0.0. 4.0 version of traineddata files may include the network spec used for LSTM training as part of version string.

HISTORY

combine_tessdata(1) first appeared in version 3.00 of Tesseract

SEE ALSO

tesseract(1), wordlist2dawg(1), cntraining(1), mftraining(1), unicharset(5), unicharambigs(5)

COPYING

AUTHOR

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