NAME

wordlist2dawg – convert a wordlist to a DAWG for Tesseract

SYNOPSIS

wordlist2dawg *WORDLIST DAWG lang.unicharset*

wordlist2dawg –t *WORDLIST DAWG lang.unicharset*

wordlist2dawg –r 1 *WORDLIST DAWG lang.unicharset*

wordlist2dawg –r 2 *WORDLIST DAWG lang.unicharset*

wordlist2dawg –l <short> <long> *WORDLIST DAWG lang.unicharset*

DESCRIPTION

wordlist2dawg(1) converts a wordlist to a Directed Acyclic Word Graph (DAWG) for use with Tesseract. A DAWG is a compressed, space and time efficient representation of a word list.

OPTIONS

–t Verify that a given dawg file is equivalent to a given wordlist.

–r 1 Reverse a word if it contains an RTL character.

–r 2 Reverse all words.

–l <short> <long> Produce a file with several dawgs in it, one each for words of length <short>, <short+1>,... <long>

ARGUMENTS

WORDLIST A plain text file in UTF–8, one word per line.

DAWG The output DAWG to write.

lang.unicharset The unicharset of the language. This is the unicharset generated by mftraining(1).

SEE ALSO

tesseract(1), combine_tessdata(1), dawg2wordlist(1)

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).