

NAME

dawg2wordlist – convert a Tesseract DAWG to a wordlist

SYNOPSIS

dawg2wordlist *UNICHARSET DAWG WORDLIST*

DESCRIPTION

dawg2wordlist(1) converts a Tesseract Directed Acyclic Word Graph (DAWG) to a list of words using a unicharset as key.

OPTIONS

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

DAWG The input DAWG, created by wordlist2dawg(1)

WORDLIST Plain text (output) file in UTF-8, one word per line

SEE ALSO

tesseract(1), mftraining(1), wordlist2dawg(1), unicharset(5), combine_tessdata(1)

<https://github.com/tesseract-ocr/tesseract/wiki/TrainingTesseract>

COPYING

Copyright (C) 2012 Google, Inc. Licensed under the Apache License, Version 2.0

AUTHOR

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