Package 'tesseract'

December 7, 2016

Type Package

Title Open Source OCR Engine

Version 1.3

Author Jeroen Ooms

Maintainer Jeroen Ooms <jeroen.ooms@stat.ucla.edu>

Description An OCR engine with unicode (UTF-8) support that can recognize over 100 languages out of the box.

License MIT + file LICENSE

URL https://github.com/ropensci/tesseract

BugReports https://github.com/ropensci/tesseract/issues

SystemRequirements Tesseract >= 3.03 (libtesseract-dev / tesseract-devel) and Leptonica (libleptonica-dev / leptonica-devel). On Debian you need to install the English training data separately (tesseract-ocr-eng)

Imports Rcpp (>= 0.12.0), curl, digest

LinkingTo Rcpp

RoxygenNote 5.0.1.9000

Suggests magick, pdftools, tiff

NeedsCompilation yes

Repository CRAN

Date/Publication 2016-12-07 16:47:15

R topics documented:

`ocr` *Tesseract OCR*

Description

Extract text from an image. Requires that you have training data for the language you are reading. Works best for images with high contrast, little noise and horizontal text.

Usage

```
ocr(image, engine = tesseract("eng"))

tesseract(language = NULL, datapath = NULL, options = NULL,
  cache = TRUE)
```

Arguments

`image`	file path, url, or raw vector to image (png, tiff, jpeg, etc)
`engine`	a tesseract engine created with `tesseract()`
`language`	string with language for training data. Usually defaults to `eng`
`datapath`	path with the training data for this language. Default uses the system library.
`options`	a named list with tesseract engine options
`cache`	use a cached version of this training data if available

Details

Tesseract uses training data to perform OCR. Most systems default to English training data. To improve OCR performance for other langauges you can to install the training data from your distribution. For example to install the spanish training data:

- tesseract-ocr-spa (Debian, Ubuntu)
- tesseract-langpack-spa (Fedora, EPEL)

On other platforms you can manually download training data from github and store it in a path on disk that you pass in the `datapath` parameter. Alternatively you can set a default path via the `TESSDATA_PREFIX` environment variable.

References

Tesseract training data

Examples

```
# Simple example
text <- ocr("http://jeroenooms.github.io/images/testocr.png")
cat(text)

# Roundtrip test: render PDF to image and OCR it back to text
library(pdftools)
library(tiff)

# A PDF file with some text
setwd(tempdir())
news <- file.path(Sys.getenv("R_DOC_DIR"), "NEWS.pdf")
orig <- pdf_text(news)[1]

# Render pdf to jpeg/tiff image
bitmap <- pdf_render_page(news, dpi = 300, numeric = TRUE)
tiff::writeTIFF(bitmap, "page.tiff")

# Extract text from images
out <- ocr("page.tiff")
cat(out)

engine <- tesseract(options = list(tessedit_char_whitelist = "0123456789"))
```

---

tesseract_download *Tesseract Training Data*

Description

Helper function to download training data from the official tessdata repository. Only use this function on Windows and OS-X. On Linux, training data can be installed directly with yum or apt-get.

Usage

```
tesseract_download(lang, datapath = NULL, progress = TRUE)

tesseract_info()
```

Arguments

lang	three letter code for language, see tessdata repository.
datapath	destination directory where to download store the file
progress	print progress while downloading

Examples

```
## Not run:
tesseract_download("fra")
french <- tesseract("fra")
text <- ocr("http://ocrapiservice.com/static/images/examples/french_text.png", engine = french)
cat(text)

## End(Not run)
```

Index