Support for hOCR and Tesseract 4 in R

Jeroen Ooms | FEBRUARY 14, 2018

Earlier this month we released a new version of the tesseract package to CRAN. This package provides R bindings to Google's open source optical character recognition (OCR) engine Tesseract.

Two major new features are support for HOCR and support for the upcoming Tesseract 4.

hOCR output

Support for HOCR output was requested by one of our users on Github. The `ocr()` function gains a parameter `HOCR` which allows for returning results in hOCR format:

```
library(tesseract)

# Text output
text ← ocr("https://jeroen.github.io/images/testocr.p
cat(text)

# hOCR output
xml ← ocr("https://jeroen.github.io/images/testocr.pr
cat(xml)
```

hOCR is an open standard of data representation for formatted text obtained from OCR (wikipedia). The definition encodes text, style, layout information, recognition confidence metrics and other information using XML.

Every word in the hOCR output includes meta data such as bounding box, confidence metrics, etc. With a little xml2 and regular expression magic we can extract a beautiful data frame:

```
library(tesseract)
library(xml2)
library(stringr)
library(tibble)
xml <- ocr("https://jeroen.github.io/images/testocr.pr
doc <- read_xml(xml)
nodes <- xml_find_all(doc, ".//span[@class='ocrx_word']
words <- xml_text(nodes)
metatext <- xml_attr(nodes, 'title')
bbox <- str_replace(str_extract(meta, "bbox [\\d ]+"),
conf <- as.numeric(str_replace(str_extract(meta, "x_wo
tibble(confidence = conf, word = words, bbox = bbox)
```

```
# A tibble: 60 x 3
   confidence word  bbox
        <dbl> <chr> <chr>
 1       89.0 This  36 92 96 116
 2       89.0 is    109 92 129 116
 3       92.0 a     141 98 156 116
 4       93.0 lot   169 92 201 116
 5       91.0 of    212 92 240 116
 6       91.0 12    251 92 282 116
 7       92.0 point 296 92 364 122
 8       89.0 text  374 93 427 116
 9       93.0 to    437 93 463 116
10       90.0 test  474 93 526 116
# ... with 50 more rows
```

So this gives us a little more information about the OCR results than just the text.

Upcoming Tesseract 4

The Google folks and contributors are working very hard on the next generation the Tesseract OCR engine which uses new neural network system based on LSTMs, with major accuracy gains. The release of Tesseract 4 is scheduled for later this year but an alpha release is already available.

Our latest CRAN release of the tesseract now has the required changes to support Tesseract 4. On MacOS you can already give this try this by installing tesseract from the master branch:

```
brew remove tesseract
brew install tesseract --HEAD
```

After updating tesseract you need to reinstall the R package from source:

```
install.packages("tessract", type = "source")
```

This is still alpha, things may break. Report problems in our github repository.

Comments Community 1 Login

♡ Recommend

Share

Sort by Best

Start the discussion…

LOG IN WITH

OR SIGN UP WITH DISQUS ?

Name

Be the first to comment.

ALSO ON ROPENSCI

Overlaying climate data with species occurrence data

3 comments • a year ago

Ted Hart — rWBclimate includes future climate data in the package.

Testing packages with R Travis for OS-X

2 comments • 2 years ago

John Blischak — Thanks for

hunspell tutorial

1 comment • a year ago

TimothyBates — Nice! Is there a way to add words to a local dictionary, either within the

Australian rOpenSci Unconference

1 comment • 2 years ago

Michael Strack — Wicked

INFO

Mission

Team

Collaborators

Careers

WORK

Packages

Blog

Tech Notes

Tutorials

Use Cases

More Resources

PARTICIPATE

Contact us

Community

Contribute software

Unconference

Code of conduct

rOpenSci is a fiscally sponsored project of NumFOCUS

Donate