KID
Handling Scanned Files

Fabian Blasch

08.03.2022

Tesseract

Tesseract is an optical character recognition engine, it can be used to convert pictures into machine readable characters. The sourcecode is available on GitHub and fortunately the API was implemented in an R package.

First Steps

Tesseract offers the possibility to alter the engine based on the problem at hand. In the case of KIDs, we want to extract the position of the scale of each SRRI entry. Thus one parameter of great importance is the character white list. It allows us to narrow the classification into the characters or in this case digits handed over to tesseract via the white list.

```
# load KiDs
setwd("./../KIDs")
devtools::load_all()
```

```
## i Loading KIDs
```

```
# example pdf
setwd("./../../../KIDs/Security")
pd <- list.files(pattern = ".pdf")[1]

# first convert to png so we can use tesseract directly
pdftools::pdf_convert(pd, pages = 1, filenames = c("testscanread.png"), dpi = 600)
```

```
## Converting page 1 to testscanread.png... done!
```

```
## [1] "testscanread.png"
```

```
# set whitelist to digits 1-7 and whitespace
tesseract::ocr_data("Scale_notext.jpeg",
                engine = tesseract::tesseract(options = list(tessedit_char_whitelist = " 1234567")))
```

```
##   word confidence           bbox
## 1    1   0.000000    97,5,165,84
## 2    2  88.425575   342,5,402,85
## 3    4   7.978073   816,5,869,84
## 4    5  60.284004 1061,5,1121,84
## 5    6  77.367195 1313,5,1366,85
## 6    7   0.000000 1528,5,1574,85
```

The readout worked good for all digits on white background, next we may use other engine parameters as well as image processing to obtain the scale in its entirety.