Finanzas en R

Aplicaciones de OCR

Sebastián Egaña Santibáñez Nicolás Leiva Díaz

Enlaces del profesor

https://segana.netlify.app

https://github.com/sebaegana

https://www.linkedin.com/in/sebastian-egana-santibanez/

Optical Character Recognition

OCR significa Reconocimiento Óptico de Caracteres (Optical Character Recognition, en inglés) y es una tecnología que permite convertir texto contenido en imágenes o documentos escaneados en texto editable y procesable digitalmente. Es especialmente útil cuando se trabaja con PDFs escaneados o imágenes que contienen texto, como facturas, informes o formularios.

A continuación veremos un ejemplo de aplicación de dicha tecnología.

Librerías a utilizar:

library(pdftools)

Esta librería permite leer y manipular archivos PDF. Tiene funciones como:

- pdf_text(): extrae texto de PDFs que ya contienen texto digital.
- pdf_render_page(): convierte una página del PDF en una imagen (útil cuando el PDF es escaneado y no tiene texto digital).

library(tesseract)

Tesseract es un motor OCR de código abierto muy potente. En R, esta librería permite:

- Aplicar OCR a imágenes para extraer el texto contenido.
- Reconocer múltiples idiomas (por defecto, el inglés, pero puedes instalar otros).

library(stringr)

Esta librería es parte del tidyverse y proporciona funciones para manipular y limpiar cadenas de texto, por ejemplo:

- str_detect(), str_replace(), str_extract(), str_sub() y muchas otras.
- Muy útil para procesar el texto extraído con pdftools o tesseract, limpiarlo y extraer información relevante.

Material

- Reporte

Códigos a utilizar

Caso 01

```
library(pdftools)
# Using poppler version 22.04.0

border_patrol <- pdf_text("usbp_stats_fy2017_sector_profile.pdf")
```

```r
head(border_patrol)

length(border_patrol)

border_patrol[1]

strsplit("criminology", split = "n")

sector_profile <- border_patrol[1]
sector_profile <- strsplit(sector_profile, "\n")
sector_profile <- sector_profile[[1]]

head(sector_profile)

sector_profile <- trimws(sector_profile)

sector_profile

grep("Miami", sector_profile)

grep("Nationwide Total", sector_profile)

sector_profile <- sector_profile[grep("Miami", sector_profile):
                                 grep("Nationwide Total", sector_profile)]

head(sector_profile)

library(stringr)

sector_profile <- str_split_fixed(sector_profile, " {2,}", 10)

head(sector_profile)

sector_profile <- data.frame(sector_profile)
names(sector_profile) <- c("sector",
                           "agent_staffing",
                           "apprehensions",
                           "other_than_mexican_apprehensions",
                           "marijuana_pounds",
                           "cocaine_pounds",
                           "accepted_prosecutions",
```

```
                              "assaults",
                              "rescues",
                              "deaths")
```

Caso 02

```
library(tesseract)

eng <- tesseract("eng")
text <- tesseract::ocr("http://jeroen.github.io/images/testocr.png", engine = eng)
cat(text)

pngfile <- pdftools::pdf_convert('https://jeroen.github.io/images/ocrscan.pdf', dpi = 600)

text <- tesseract::ocr(pngfile)
cat(text)
```