

Beyond HTML (1). Scraping PDF documents

Alex Sanchez and Francesc Carmona
Genetics Microbiology and Statistics Department
Universitat de Barcelona
October 2022

Outline

- 1) Introduction
- 2) Scraping PDF files
- 3) References and Resources

Introduction

Beyond HTML

- We have learned how to scrape *static* web pages which are developed mainly in HTML.
- While the most common, it is not the only format used in the web either
 - To develop web pages
 - To store data
- A global overview of web scraping should also consider *dynamic web pages*,
- As well of acquiring data from a bunch of distinct formats such as XML, JSON or PDF.
- In this chapter we will take a bird's eye view on how to extract data from PDF files.

Extracting information from PDF files

Plenty of pdf files

- PDF is a very popular format between users.
- Although it is not exactly a format for the web,
- The web is full of pdf documents giving access to
 - Text we may wish to mine
 - Data we may want to recover
- We would like to be able to easily extract text and tables from pdf documents in a similar way as we do with HTML.

The Portable Document Format

- The PDF format was created by Adobe in 1993 to facilitate *sharing* and *printing* documents.
- But it was not intended for being indexed or searched.
- Unlike HTML or XML, PDF files do not have an easy-to-parse structure which makes scraping pdf more artisanal work.

There are PDFs and PDFs

- The fact that there is no agreement on the internal structure a pdf file must have, makes it more complicated to extract information from it in a unique form.
- Some files, such as pdfs created by saving a text document or "printing" a file as pdf may have an relatively clean structure that makes them easy to parse.
- Other, such as files created by scanning text/images ("OCR") may be more complicated requiring distinct software for that goal.

PDF conversion software

- The characteristics described above make scraping pdf files a task more complicated than one would expect.
- This has also generated the availability of multiple software solutions of distinct types and distinct prices.
 - [Top 9 Free PDF Converter in 2022](#)
 - [How to Extract Data From PDF Documents](#) (*Commercial, not endorsed!*)
- Some factors when it comes to deciding which tool to use
 - Ease of use
 - Repeteability
 - Quantity

R packages

- There are a few packages specific for pdf files
 - **pdftools**
 - pdftables
 - PDF Data Extractor (PDE)
 - **tabulizer** (*Not in CRAN due to dependencies problems*)
- Also some standard packages allow reading files.
 - **textreadr**
 - A small collection of tools to read many file types.

Extracting tables with `tabulizer`

Tabulizer Example

- From the package vignette

```
library(tabulizer)
f ← system.file("examples", "data.pdf", package = "tabulizer")
out1 ← extract_tables(f)
str(out1)
## List of 4
## $ : chr [1:32, 1:10] "mpg" "21.0" "21.0" "22.8" ...
## $ : chr [1:7, 1:5] "Sepal.Length" "5.1" "4.9" "4.7" ...
## $ : chr [1:7, 1:6] "" "145" "146" "147" ...
## $ : chr [1:15, 1] "supp" "VC" "VC" "VC" ...
```

A more elaborated example (1)

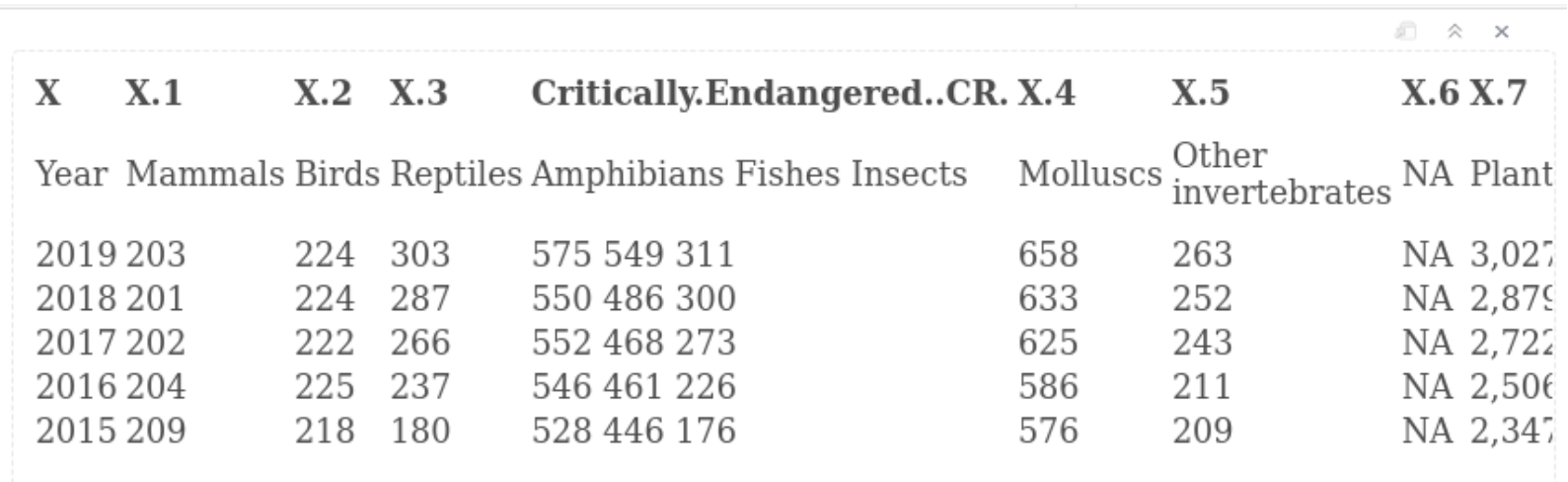
This has been adapted from the R-bloggers entry:
[PDF Scraping in R with tabulizer](#)

```
library(tabulizer)
library(tidyverse)
myFile ← "endangered_species.pdf"
endangered_species_scrape ← extract_tables(
  file = myFile,
  method = "decide",
  output = "data.frame")
endangered_species_raw_tbl ← endangered_species_scrape %>%
  pluck(1) %>%
  as_tibble()
```

A more elaborated example (2)

```
# Show first 6 rows
```

```
endangered_species_raw_tbl %>% head() %>% knitr::kable()
```



The screenshot shows a R console window with a kable table. The table has 11 columns: X, X.1, X.2, X.3, Critically.Endangered..CR., X.4, X.5, X.6, X.7, and X.8. The rows represent years from 2015 to 2019. The data is as follows:

| X | X.1 | X.2 | X.3 | Critically.Endangered..CR. | X.4 | X.5 | X.6 | X.7 | X.8 |
|------|---------|-------|----------|----------------------------|--------|---------|----------|---------------------|----------|
| Year | Mammals | Birds | Reptiles | Amphibians | Fishes | Insects | Molluscs | Other invertebrates | NA Plant |
| 2019 | 203 | 224 | 303 | 575 | 549 | 311 | 658 | 263 | NA 3,027 |
| 2018 | 201 | 224 | 287 | 550 | 486 | 300 | 633 | 252 | NA 2,879 |
| 2017 | 202 | 222 | 266 | 552 | 468 | 273 | 625 | 243 | NA 2,722 |
| 2016 | 204 | 225 | 237 | 546 | 461 | 226 | 586 | 211 | NA 2,506 |
| 2015 | 209 | 218 | 180 | 528 | 446 | 176 | 576 | 209 | NA 2,347 |

Once the data has been extracted it may be (must be) further processed

- to clean and tidy the data
- to visualize and analyze it

Extracting text with pdf_tools()

The `pdf_tools` package

- Imagine we need to convert a pdf file from the "BOLETÍN OFICIAL DEL REGISTRO MERCANTIL"

<https://www.boe.es/borme/dias/2022/10/05/pdfs/BORME-A-2022-191-08.pdf>

- Start downloading the file on which we are interested.

```
myURL ← "https://www.boe.es/borme/dias/2022/10/05/pdfs/BORME-A-2022-191-08.pdf"  
myFilename ← "BORME-A-2022-191-08.pdf"  
download.file(url=myURL, destfile=myFilename)
```


Text extraction from pdf

- PDF file is extracted as a vector of character strings
- Each string contains a page of the document
- Once the data is available it may/must be cleaned using R text analysis capabilities

```
library(pdftools)
txt ← pdf_text(myFilename)
class(txt)
length(txt)
cat(txt[1])
```

Núm. 191
Pág. 46120

BOLETÍN OFICIAL DEL REGISTRO MERCANTIL
Miércoles 5 de octubre de 2022

SECCIÓN PRIMERA
Empresarios
Actos Insritos
BARCELONA

438642 - AUTHO LTD SUCURSAL EN ESPAÑA.
Fe de erratas: En el BORME no.8/2022 y con referencia 13527 SE OMITIO LA PUBLICACION DE "OTROS ACTOS: CONSTANCIA DE LA COMPOSICION DEL ORGANO DE ADMINISTRACION DE LA SOCIEDAD MATRIZ ". Datos registrales. T 47748 , F 215, S 8, H B 561206, I/A 2 (5.01.22).

438643 - SISTEMAS DE AUTOGAS DE ESPAÑA S.L.
Fe de erratas: En el BORME no.74/2022 y con referencia 144233 SE OMITIOLA PULICACION DEL ACTO DE AUMENTO DE CAPITAL EN "3100,00 " SIENDO EL CAPITAL RESULTANTE DE " 6100,00 ". Datos registrales. T 47301 , F 195, S 8, H B 547852, I/A 4 (27.03.20).

438644 - PROMVIAS XXI SA.
Nombramientos: CONS.DEL.MAN: ANGULO CORDERO RAFAEL. Fe de erratas: En el BORME no.01/2015 y con referencia 178510 SE RECTIFICA EN CUANTO A LA OMISION DEL NOMBRAMIENTO DE RAFAEL ANGULO CORDERO EN EL CARGO DE CONS.DEL.MAN. Datos registrales. T 30215 , F 199, S 8, H B 318461, I/A 17 (22.04.15).

Getting more information

- In addition to reading in our .pdf file, we may want to extract certain metadata about it as well.
- pdftools has a few handy functions that can be used to extract things
 - the number of pages: `pdf_info()`
 - the fonts being used: `pdf_fonts()`
 - the table of contents: `pdf_toc()`
 - whether there are any attachments: `pdf_attachments()`
 - or the original date and time the document was created:
`pdf_pagesize()`

References and Resources

Resources

- PDF Scraping in R with tabulizer
- Converting PDFs to txt files with R -Getting your .pdfs into R - Parsing your .pdfs in R
- <https://learningactors.com/how-to-extract-and-clean-data-from-pdf-files-in-r/> -Scrape tables from PDF