Corpus of Decisions

Permanent Court of International Justice

(CD-PCIJ-Source)

Compilation Report

Version 1.1.0

DOI: 10.5281/zenodo.7051937

Title	Source Code for the 'Corpus of Decisions: Permanent Court of International Justice'
Abkürzung	CD-PCIJ-Source
Author	Seán Fobbe
Version	1.1.0
Download	https://doi.org/10.5281/zenodo.7051937
License	MIT No Attribution (MIT-0)

Citation

Seán Fobbe (2022). Source Code for the 'Corpus of Decisions: Permanent Court of International Justice' (CD-PCIJ-Source). Version 1.1.0. Zenodo. DOI: 10.5281/zenodo.7051937.

Digital Object Identifiers: Concept DOI and Version DOI

This data set is uniquely identified via the Digital Object Identifier (DOI) system. DOIs are persistent identifiers that are globally unique and can be resolved as a link by entering a DOI into the web service at www.doi.org. The DOI given in this document is a *Version DOI*, which uniquely identifies version 1.1.0. Analysts who wish to enable replication analyses are strongly advised to cite the *Version DOI* and the exact version of the data used. A *Concept DOI* is available from the page of the Zenodo record under the heading 'Cite all versions?' and will always resolve to the latest version.

License: MIT No Attribution (MIT-0)

Disclaimer

This data set is a personal academic initiative and is not associated with or endorsed by the International Court of Justice or the United Nations.

Contents

```r
cat(readLines("README.md"),
    sep = "\n")
```

1 README: Corpus of Decisions: Permanent Court of International Justice (CD-PCIJ)

1.1 Overview

This R script downloads and processes the full set of decisions and appended opinions rendered by the Permanent Court of International Justice (PCIJ) — as published in Series A, B and AB on https://www.icj-cij.org — into a rich and structured human- and machine-readable data set. It is the basis of the **Corpus of Decisions: Permanent Court of International Justice (CD-PCIJ)**.

All data sets created with this script will always be hosted permanently open access and freely available at Zenodo, the scientific repository of CERN. Each version is uniquely identified with a persistent Digitial Object Identifier (DOI), the *Version DOI*. The newest version of the data set will always available via the link of the *Concept DOI*: https://doi.org/10.5281/zenodo.3840479

1.2 Functionality

This script will produce 17 ZIP archives:

- 2 archives of CSV files containing the full machine-readable data set (English/French)
- 2 archives of CSV files containing the full machine-readable metadata (English/French)
- 2 archives of TXT files containing all machine-readable texts with a reduced set of metadata encoded in the filenames (English/French)
- 2 archives of PDF files containing all human-readable texts with enhanced OCR (English/French)
- 2 archives of PDF files containing all human-readable majority opinions with enhanced OCR (English/French)
- 2 archives of PDF files containing original documents split into monolingual documents (English/French)
- 2 archives of TXT files containing extracted text from the original documents (English/French)
- 1 archive of PDF files as originally published by the PCIJ/ICJ (multilingual)
- 1 archive of analysis data and diagrams
- 1 archive containing all source files

The integrity and veracity of each ZIP archive is documented with cryptographically secure hash signatures (SHA2-256 and SHA3-512). Hashes are stored in a separate CSV file created during the data set compilation process.

Please refer to the Codebook regarding the relative merits of each variant. Unless you have very specific needs you should only use the variants denoted 'TESSERACT' or 'ENHANCED' for serious work.

1.3 System Requirements

- You must have the R Programming Language and all **R packages** listed under the heading 'Load Packages' installed.

- You must have the system dependencies **tesseract** and **imagemagick** (on Fedora Linux, names may differ with other Linux distibutions) installed for the OCR pipeline to work.
- Due to the use of Fork Clusters and system commands the script as published will (probably) only run on Fedora Linux. The specific version of Fedora used is documented as part of the session information at the end of this script. With adjustments it may also work on other distributions.
- Parallelization will automatically be customized to your machine by detecting the maximum number of cores. A full run of this script takes approximately 90 minutes on a machine with a Ryzen 3700X CPU using 16 threads, 64 GB DDR4 RAM and a fast SSD.
- You must have the **openssl** system library installed for signature generation. If you prefer not to generate signatures this part of the script can be removed without affecting other parts, but a missing signature CSV file will result in non-fatal errors during Codebook compilation.
- Optional code to compile a high-quality PDF report adhering to standards of strict reproducibility is included. This requires the R packages **rmarkdown**, **magick**, an installation of **LaTeX** and all the packages specified in the TEX Preamble file.

1.4 Compilation

All comments are in **roxygen2-style** markup for use with *spin()* or *render()* from the **rmarkdown** package. Compiling the scripts will produce the full data set, high-quality PDF reports and save all diagrams to disk.

Both scripts can be executed as ordinary R scripts without any of the markdown and report generation elements. The Corpus creation script will also produce the full data set. No diagrams or reports will be saved to disk in this scenario.

To compile the full data set, a Compilation Report and the Codebook, copy all files provided in the Source ZIP Archive into an empty (!) folder and run the following command in an R session:

```
source("run_project.R")
```

1.5 Open Access Publications (Fobbe)

Website — https://www.seanfobbe.com

Open Data — https://zenodo.org/communities/sean-fobbe-data

Code Repository — https://zenodo.org/communities/sean-fobbe-code

Regular Publications — https://zenodo.org/communities/sean-fobbe-publications

1.6 Contact

Did you discover any errors? Do you have suggestions on how to improve the data set? You can either post these to the Issue Tracker on GitHub or write me an e-mail at fobbe-data@posteo.de

2 Preamble

2.1 Datestamp

The datestamp is set at the beginning of the script so it will be held constant even if long runtime breaks the date barrier.

```
datestamp <- Sys.Date()
print(datestamp)
```

```
## [1] "2022-09-06"
```

2.2 Date and Time (Begin)

```
begin.script <- Sys.time()
print(begin.script)
```

```
## [1] "2022-09-06 19:01:06 CEST"
```

2.3 Load Packages

```
library(RcppTOML)      # Read and write TOML files
library(httr)          # HTTP Tools
library(rvest)         # Web Scraping
library(mgsub)         # Vectorized Gsub
library(stringr)       # String Manipulation
library(pdftools)      # PDF utilities
```

```
## Using poppler version 22.01.0
```

```
library(fs)            # File Operations
library(knitr)         # Scientific Reporting
library(kableExtra)    # Enhanced Knitr Tables
library(magick)        # Required for cropping when compiling PDF
```

```
## Linking to ImageMagick 6.9.12.52
## Enabled features: cairo, fontconfig, freetype, ghostscript, lcms, pango, raw,
    rsvg, webp, x11
## Disabled features: fftw, heic
```

```
## Using 16 threads
```

```r
library(DiagrammeR)    # Graph/Network Visualization
library(DiagrammeRsvg) # Export DiagrammeR Graphs as SVG
library(rsvg)          # Render SVG to PDF
```

```
## Linking to librsvg 2.54.5
```

```r
library(ggplot2)       # Advanced Plotting
library(scales)        # Rescaling of Plots
library(viridis)       # Viridis Color Palette
```

```
## Loading required package: viridisLite
```

```
##
## Attaching package: 'viridis'
```

```
## The following object is masked from 'package:scales':
##
##     viridis_pal
```

```r
library(RColorBrewer)  # ColorBrewer Palette
library(readtext)      # Read TXT Files
library(quanteda)      # Advanced Text Analytics
```

```
## Package version: 3.2.3
## Unicode version: 13.0
## ICU version: 69.1
```

```
## Parallel computing: 16 of 16 threads used.
```

```
## See https://quanteda.io for tutorials and examples.
```

```r
library(quanteda.textstats)  # Text Statistics Tools
library(quanteda.textplots)  # Specialized Plots for Text Statistics
library(textcat)       # Classify Text Language
library(data.table)    # Advanced Data Handling
```

```
## data.table 1.14.2 using 8 threads (see ?getDTthreads).  Latest news: r-
    datatable.com
```

```r
library(doParallel)    # Parallelization
```

```
## Loading required package: foreach
```

```
## Loading required package: iterators
```

```
## Loading required package: parallel
```

2.4 Load Additional Functions

Note: Each custom function will be printed in full prior to its first use in order to enhance readability. All custom functions are prefixed with 'f.' for clarity.

```r
source("functions/f.boxplot.body.R")
source("functions/f.boxplot.outliers.R")
source("functions/f.dopar.multihashes.R")
source("functions/f.dopar.pagenums.R")
source("functions/f.dopar.pdfextract.R")
source("functions/f.dopar.pdfocr.R")
source("functions/f.fast.freqtable.R")
source("functions/f.hyphen.remove.R")
source("functions/f.lingsummarize.iterator.R")
source("functions/f.linkextract.R")
source("functions/f.special.replace.R")
source("functions/f.token.processor.R")
```

3 Parameters

3.1 Output Directory

The directory name must include a terminating slash!

```
outputdir <- paste0(getwd(),
                    "/ANALYSIS/")
```

3.2 Read Configuration File

All configuration options are set in a separate configuration file that is read here. They should only be changed in that file!

```
config <- RcppTOML::parseTOML("config.toml")
print(config)
```

```
## List of 9
##  $ cores   :List of 2
##   ..$ max   : logi TRUE
##   ..$ number: int 8
##  $ doi     :List of 2
##   ..$ data    :List of 2
##   .. ..$ concept: chr "10.5281/zenodo.3840479"
##   .. ..$ version: chr "10.5281/zenodo.7051934"
##   ..$ software:List of 2
##   .. ..$ concept: chr "10.5281/zenodo.4136955"
##   .. ..$ version: chr "10.5281/zenodo.7051937"
##  $ download:List of 1
##   ..$ timeout: int 600
##  $ fig     :List of 3
##   ..$ align : chr "center"
##   ..$ dpi   : int 300
##   ..$ format: chr [1:2] "pdf" "png"
##  $ freqvar :List of 1
##   ..$ ignore: chr [1:3] "date" "doc_id" "text"
##  $ license :List of 2
##   ..$ code: chr "MIT-0"
##   ..$ data: chr "Creative Commons Zero 1.0 Universal"
##  $ ocr     :List of 1
##   ..$ dpi: int 300
##  $ project :List of 3
##   ..$ author   : chr "Seán Fobbe"
##   ..$ fullname : chr "Corpus of Decisions: Permanent Court of International
    Justice"
##   ..$ shortname: chr "CD-PCIJ"
##  $ version :List of 2
##   ..$ dash    : chr "1-1-0"
##   ..$ semantic: chr "1.1.0"
```

3.3 Name of Data Set

```
datashort <- config$project$shortname
print(datashort)
```

```
## [1] "CD-PCIJ"
```

3.4 Version Number

```
version <- config$version$semantic
print(version)
```

```
## [1] "1.1.0"
```

3.5 Create Version Number with Dashes

This is used in output files.

```
version.dash <- gsub("\\.",
                     "-",
                     version)
print(version.dash)
```

```
## [1] "1-1-0"
```

3.6 DOI of Data Set Concept

```
doi.concept <- config$doi$data$concept
print(doi.concept)
```

```
## [1] "10.5281/zenodo.3840479"
```

3.7 DOI of Specific Version

```
doi.version <- config$doi$data$version
print(doi.version)
```

```
## [1] "10.5281/zenodo.7051934"
```

3.8 License

```
license <- config$license$data
print(license)
```

```
## [1] "Creative Commons Zero 1.0 Universal"
```

3.9 DPI for OCR

This is the resolution at which PDF files will be converted to TIFF during the OCR step. DPI values will significantly affect the quality of text ouput and file size. Higher DPI requires more RAM, means higher quality text and greater PDF file size. A value of 300 is recommended.

```
ocr.dpi <- config$ocr$dpi
print(ocr.dpi)
```

```
## [1] 300
```

3.10 Frequency Tables: Ignored Variables

This is a character vector of variable names that will be ignored in the construction of frequency tables.

It is a good idea to add variables to this list that are unlikely to produce useful frequency tables. This is often the case for variables with a very large proportion of unique values. Use this option judiciously, as frequency tables are useful for detecting anomalies in the metadata.

```
freq.var.ignore <- config$freqvar$ignore
print(freq.var.ignore)
```

```
## [1] "date"   "doc_id" "text"
```

3.11 Set Download Timeout

```
options(timeout = config$download$timeout)
```

3.12 Knitr Options

3.12.1 Image Output File Formats

```
fig.format <- config$fig$format
print(fig.format)
```

```
## [1] "pdf" "png"
```

3.12.2 DPI for Raster Graphics

```
fig.dpi <- config$fig$dpi
print(fig.dpi)
```

```
## [1] 300
```

3.12.3 Alignment of Diagrams in Report

```
fig.align <- config$fig$align
print(fig.align)
```

```
## [1] "center"
```

3.12.4 Set Knitr Options

```
knitr::opts_chunk$set(fig.path = outputdir,
                      dev = fig.format,
                      dpi = fig.dpi,
                      fig.align = fig.align)
```

4 Manage Directories

4.1 Define Set of Data Directories

```
dirset <- c("MULT_PDF_ORIGINAL_FULL",
            "EN_PDF_ENHANCED_FULL",
            "FR_PDF_ENHANCED_FULL",
            "EN_PDF_ORIGINALSPLIT_FULL",
            "FR_PDF_ORIGINALSPLIT_FULL",
            "EN_PDF_ENHANCED_MajorityOpinions",
            "FR_PDF_ENHANCED_MajorityOpinions",
            "EN_TXT_TESSERACT_FULL",
            "FR_TXT_TESSERACT_FULL",
            "EN_TXT_EXTRACTED_FULL",
            "FR_TXT_EXTRACTED_FULL")
```

4.2 Clean up files from previous runs

```
delete <- list.files(pattern = "\\.pdf|\\.zip|\\.pdf|\\.csv|\\.tex")
unlink(delete)

for (dir in dirset){
    unlink(dir, recursive = TRUE)
}

unlink(outputdir, recursive = TRUE)
unlink("temp", recursive = TRUE)
```

4.3 Create directories

```
for (dir in dirset){
    dir.create(dir)
}


dir.create("temp")
dir.create(outputdir)
```

5 LaTeX Configuration

These LaTeX definitions are used for the cover and inside cover.

5.0.1 Construct LaTeX Definitions

```
latexdefs <- c("%==========================\n% Definitions\n
    %==========================",
               "\n% NOTE: This file was created automatically during the
    compilation process.\n",
               "\n%-----Version-----",
               paste0("\\newcommand{\\version}{",
                      config$version$semantic,
                      "}"),
               "\n%-----Titles-----",
               paste0("\\newcommand{\\datatitle}{",
                      config$project$fullname,
                      "}"),
               paste0("\\newcommand{\\datashort}{",
                      config$project$shortname,
                      "}"),
               paste0("\\newcommand{\\softwaretitle}{Source Code for the \\
    enquote{",
                      config$project$fullname,
                      "}}"),
               paste0("\\newcommand{\\softwareshort}{",
                      config$project$shortname,
                      "-Source}"),
               "\n%-----Data DOIs-----",
               paste0("\\newcommand{\\dataconceptdoi}{",
                      config$doi$data$concept,
                      "}"),
               paste0("\\newcommand{\\dataversiondoi}{",
                      config$doi$data$version,
                      "}"),
               paste0("\\newcommand{\\dataconcepturldoi}{https://doi.org/",
                      config$doi$data$concept,
                      "}"),
               paste0("\\newcommand{\\dataversionurldoi}{https://doi.org/",
                      config$doi$data$version,
                      "}"),
               "\n%-----Software DOIs-----",
               paste0("\\newcommand{\\softwareconceptdoi}{",
                      config$doi$software$concept,
                      "}"),
               paste0("\\newcommand{\\softwareversiondoi}{",
                      config$doi$software$version,
                      "}"),

               paste0("\\newcommand{\\softwareconcepturldoi}{https://doi.org/",
                      config$doi$software$concept,
                      "}"),
               paste0("\\newcommand{\\softwareversionurldoi}{https://doi.org/",
                      config$doi$software$version,
```

```
                        "}"))
```

5.0.2 Write LaTeX Definitions

```
writeLines(latexdefs,
           "temp/CD-PCIJ_Source_TEX_Definitions.tex")
```

5.1 Write Package Citations

```
knitr::write_bib(c(.packages()),
                 "temp/packages.bib")
```

```
## tweaking foreach
```

6 Parallelization

Parallelization is used for many tasks in this script, e.g. for accelerating the conversion from PDF to TXT, OCR, analysis with **quanteda** and with **data.table**. The maximum number of cores will automatically be detected and used.

The download of decisions from the ICJ website is not parallelized to ensure respectful use of the Court's bandwidth.

The use of **fork clusters** is significantly more efficient than PSOCK clusters, although it restricts use of this script to Linux systems.

6.1 Detect Number of Logical Cores

This will detect the maximum number of threads (= logical cores) available on the system or set them according to the config file.

```
if(config$cores$max == TRUE){

    fullCores <- detectCores()

}else{

    fullCores <- config$cores$number

}

print(fullCores)
```

```
## [1] 16
```

6.1.1 Set Number of OCR Control Cores

Note: Reduced number of control cores for OCR, as Tesseract calls up to four threads by itself.

```
ocrCores <- round((fullCores / 4)) + 1
print(ocrCores)
```

```
## [1] 5
```

6.1.2 Data.table

```
setDTthreads(threads = fullCores)
```

6.1.3 Quanteda

```
quanteda_options(threads = fullCores)
```

7 Visualize Corpus Creation Process

```
workflow <- "
digraph workflow {

  # Graph Statement
  graph [layout = dot, overlap = false]

  # Legend

  subgraph cluster1{
      peripheries=1
  9991 [label = 'Data Nodes', shape = 'ellipse', fontsize = 22]
  9992 [label = 'Action Nodes', shape = 'box', fontsize = 22]
}


  # Data Nodes

  node[shape = 'ellipse', fontsize = 22]

  A [label = 'www.icj-cij.org']
  B [label = 'Raw PDF Files']
  C [label = 'Labelling Information']
  D [label = 'MULT_PDF_ORIGINAL_FULL']
  E [label = 'Split Instructions']
  F [label = 'EN_PDF_ORIGINALSPLIT']
  G [label = 'FR_PDF_ORIGINALSPLIT']
  H [label = 'EN_TXT_EXTRACTED']
  I [label = 'FR_TXT_EXTRACTED']
  J [label = 'EN_TXT_TESSERACT']
  K [label = 'FR_TXT_TESSERACT']
  L [label = 'EN_PDF_ENHANCED_FULL']
  M [label = 'FR_PDF_ENHANCED_FULL']
  N [label = 'EN_PDF_ENHANCED_MajorityOpinions']
  O [label = 'FR_PDF_ENHANCED_MajorityOpinions']
  P [label = 'Frequency Tables']
  Q [label = 'EN_CSV_TESSERACT_FULL']
  R [label = 'FR_CSV_TESSERACT_FULL']
  S [label = 'EN_CSV_TESSERACT_META']
  T [label = 'FR_CSV_TESSERACT_META']
  U  [label = 'ANALYSIS']


  # Action Nodes

  node[shape = 'box', fontsize = 22]

  0 [label = 'Download Module']
  1 [label = 'Labelling Module']
  2 [label = 'Strict REGEX Validation: Codebook File Name Schema']
  3 [label = 'File Split Module']
  4 [label = 'Detect Missing Language Counterparts']
  5 [label = 'Text Extraction Module']
  6 [label = 'Tesseract OCR Module']
```

```
  7 [label = 'Create Majority Variant']
  8 [label = 'OCR Quality Control Module']
  81 [label = 'Clean Texts']
  9 [label = 'Language Purity Module']
  10 [label = 'Add Metadata']
  11 [label = 'Calculate Frequency Tables']
  12 [label = 'Visualize Frequency Tables']
  13 [label = 'Calculate and Add Summary Statistics']
  14 [label = 'Calculate Token Frequencies']
  15 [label = 'Calculate Document Similarity']
  16 [label = 'Write CSV Files']

  # Edge Statements
  A -> 0 -> B
  {B,C} -> 1 -> 2 -> D
  {D,E} -> 3
  3 -> {F,G} -> 4
  4 -> 5 ->{H,I}
  4 -> 6 -> {J,K,L,M}
  {L,M} -> 7 -> {N,O}
  {H, I, J, K} -> 8
  {J,K} -> 81 -> 9 -> 10 -> 11 -> P -> 12
  10 -> {14,15}
  10 -> 13 -> 16
  16 ->{Q,R,S,T}
  {P, 11, 12, 13, 14, 15} -> U
}
"



grViz(workflow) %>% export_svg %>% charToRaw %>% rsvg_pdf("ANALYSIS/CD-PCIJ_
    Workflow.pdf")
grViz(workflow) %>% export_svg %>% charToRaw %>% rsvg_png("ANALYSIS/CD-PCIJ_
    Workflow.png")
```

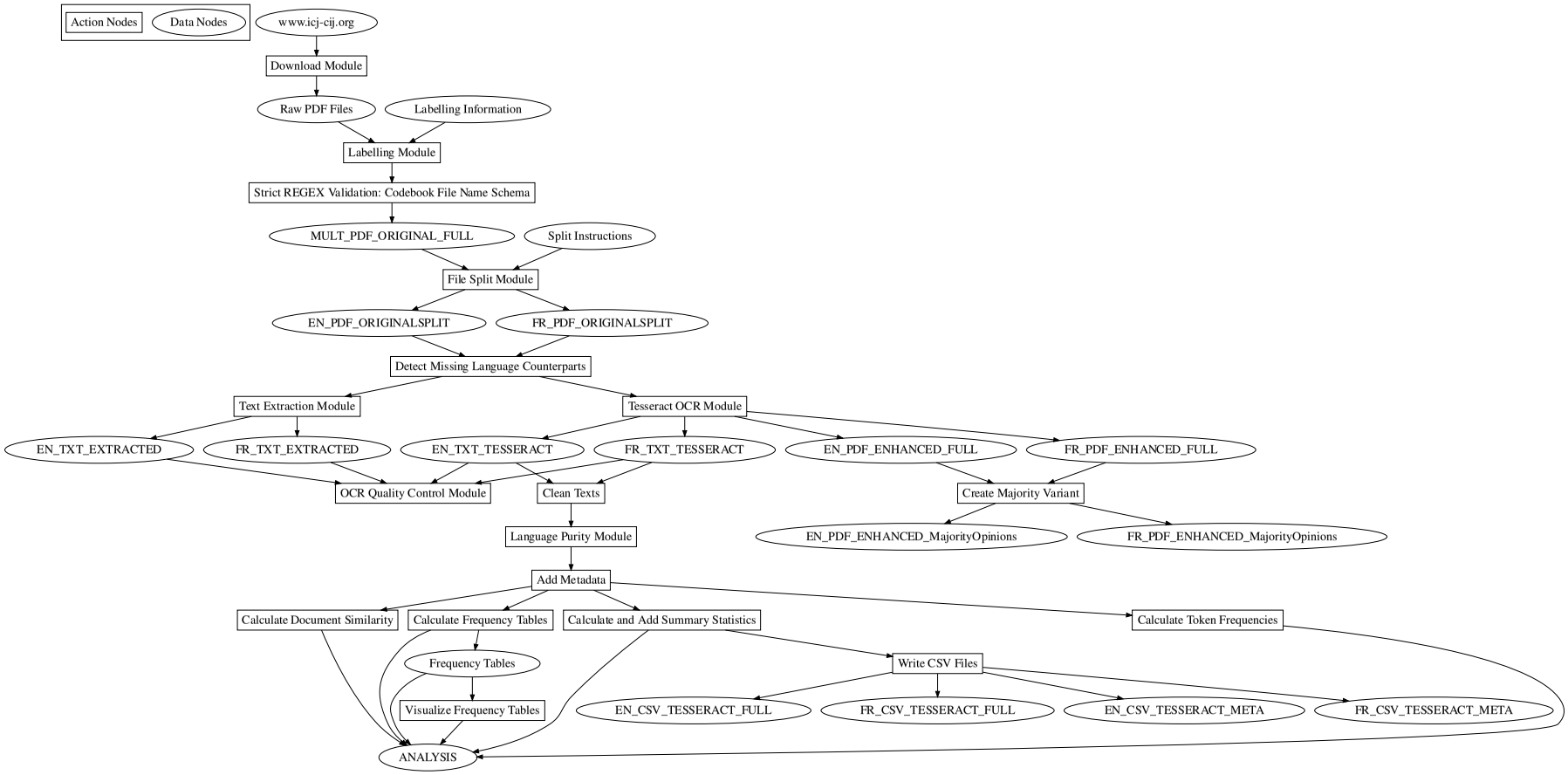


Figure 1: CD-PCIJ: Workflow Schematic

8 Download Files

8.1 Show Function: f.linkextract

```
print(f.linkextract)
```

```
## function(URL){
##     tryCatch({
##         read_html(URL) %>%
##             html_nodes("a")%>%
##             html_attr('href')},
##         error = function(cond) {
##             return(NA)}
##         )
## }
```

8.2 Acquire Download Links

8.2.1 Series A

```
URL1 <- c("https://www.icj-cij.org/en/pcij-series-a")

volatile <- f.linkextract(URL1)

links.a <- grep("files",
                volatile,
                ignore.case = TRUE,
                value = TRUE)

Sys.sleep(runif(1, 1, 3))
```

8.2.2 Series B

```
URL2 <- c("https://www.icj-cij.org/en/pcij-series-b")

volatile <- f.linkextract(URL2)

links.b <- grep("files",
                volatile,
                ignore.case = TRUE,
                value = TRUE)

Sys.sleep(runif(1, 1, 3))
```

8.2.3 Series AB

```
URL3 <- c("https://www.icj-cij.org/en/pcij-series-ab")

volatile <- f.linkextract(URL3)

links.ab <- grep("files",
                 volatile,
                 ignore.case = TRUE,
                 value = TRUE)
```

8.3 Combine

```
links.pcij <- c(links.a,
                links.b,
                links.ab)
```

8.4 Clean Links and Names

```
links.unique <- unique(links.pcij)

links.download <- paste0("https://www.icj-cij.org",
                         links.unique)

names.download <- make.unique(basename(links.download),
                              sep = "--")
```

8.5 Create Download Table

```
dt <- data.table(links.download,
                 names.download)
```

8.6 Timestamp (Download Begin)

```
begin.download <- Sys.time()
print(begin.download)
```

```
## [1] "2022-09-06 19:01:13 CEST"
```

8.7 Execute Download

```
for (i in sample(dt[,.N])){

    download.file(dt$links.download[i],
                  dt$names.download[i])

    Sys.sleep(runif(1, 0.5, 1.5))

}
```

8.8 Timestamp (Download End)

```
end.download <- Sys.time()
print(end.download)
```

```
## [1] "2022-09-06 19:06:42 CEST"
```

8.9 Duration (Download)

```
end.download - begin.download
```

```
## Time difference of 5.486681 mins
```

8.10 Download Result

8.10.1 Number of Files to Download

```
download.expected.N <- dt[,.N]
print(download.expected.N)
```

```
## [1] 264
```

8.10.2 Number of Files Successfully Downloaded

```
files.pdf <- list.files(pattern = "\\.pdf",
                        ignore.case = TRUE)

download.success.N <- length(files.pdf)
print(download.success.N)
```

```
## [1] 264
```

8.10.3 Number of Missing Files

```
missing.N <- download.expected.N - download.success.N
print(missing.N)
```

```
## [1] 0
```

8.10.4 Names of Missing Files

```
missing.names <- setdiff(dt$names.download,
                         files.pdf)
print(missing.names)
```

```
## character(0)
```

8.11 Timestamp (Retry Download Begin)

```
begin.download <- Sys.time()
print(begin.download)
```

```
## [1] "2022-09-06 19:06:42 CEST"
```

8.12 Retry Download

```
if(missing.N > 0){

    dt.retry <- dt[names.download %in% missing.names]

    for (i in 1:dt.retry[,.N]){
        response <- GET(dt.retry$links.download[i])
        Sys.sleep(runif(1, 0.25, 0.75))
        if (response$headers$"content-type" == "application/pdf" & response$
    status_code == 200){
            tryCatch({download.file(url = dt.retry$links.download[i], destfile =
    dt.retry$names.download[i])
            },
            error=function(cond) {
                return(NA)}
            )
        }else{
            print(paste0(dt.retry$names.download[i], " : no PDF available"))
        }
        Sys.sleep(runif(1, 0.5, 1.5))
    }
}
```

8.13 Timestamp (Retry Download End)

```
end.download <- Sys.time()
print(end.download)
```

```
## [1] "2022-09-06 19:06:42 CEST"
```

8.14 Duration (Retry Download)

```
end.download - begin.download
```

```
## Time difference of 0.008901358 secs
```

8.15 Retry Result

```
files.pdf <- list.files(pattern = "\\.pdf",
                        ignore.case = TRUE)
```

8.15.1 Successful during Retry

```
retry.success.names <- files.pdf[files.pdf %in% missing.names]
print(retry.success.names)
```

```
## character(0)
```

8.15.2 Missing after Retry

```
retry.missing.names <- setdiff(retry.success.names,
                               missing.names)
print(retry.missing.names)
```

```
## character(0)
```

8.16 Final Download Result

8.16.1 Number of Files to Download

```
download.expected.N <- dt[,.N]
print(download.expected.N)
```

```
## [1] 264
```

8.16.2 Number of Files Successfully Downloaded

```
files.pdf <- list.files(pattern = "\\.pdf",
                        ignore.case = TRUE)

download.success.N <- length(files.pdf)
print(download.success.N)
```

```
## [1] 264
```

8.16.3 Number of Missing Files

```
missing.N <- download.expected.N - download.success.N
print(missing.N)
```

```
## [1] 0
```

8.16.4 Names of Missing Files

```
missing.names <- setdiff(dt$names.download,
                         files.pdf)
print(missing.names)
```

```
## character(0)
```

9 Labelling Module

While the files *are* labelled semantically, the filenames contain almost no useful machine-readable information. This module applies complex hand-coded filenames to all documents in the collection. Filenames are designed to be fully interoperable with the "Corpus of Decisions: International Court of Justice (CD-ICJ)."

9.1 Manual Coding

```
##########################################
###  HAND CODING OF  FILENAMES
##########################################
```

9.2 Read Enhanced Filenames

```
filenames.enhanced <- fread("data/CD-PCIJ_Source_Filenames-FullNames-
    SplitInstructions.csv",
                            header = TRUE)[,.(oldname, newname)]
```

9.3 Strictly Validate Naming Scheme with REGEX

Test strict compliance with variable types described in Codebook. This should be an empty character vector!

9.3.1 Execute Validation

```
regex.test <- grep(paste0("^PCIJ", # var: court
                          "_",
                          "(A|B|AB)", # var: series
                          "_",
                          "[0-9]{2}", # var: seriesno
                          "_",
                          "[A-Za-z0-9-]+", # var: shortname
                          "_",
                          "[A-Z-]+", # var: applicant
                          "_",
                          "[A-Z]+", # var: respondent
                          "_",
                          "[0-9]{4}-[0-9]{2}-[0-9]{2}", # var: date
                          "_",
                          "[A-Z]{3}", # var: doctype
                          "_",
                          "[0-9]{2}", # var: collision
                          "_",
                          "(SE|IN|AJ|EV|EX|JO|IM|TL|DH|PR|DI|PO|ME|NA|EV-SE|TL-DH
    |JO-TL)", # var: stage
                          "_",
```

```
                            "[A-Z0-9]{2}", # var: opinion
                            "_",
                            "(EN|FR|DE|BI)", # var: language
                            ".pdf$"),
                    filenames.enhanced$newname,
                    value = TRUE,
                    invert = TRUE)
```

9.3.2 Results of Validation

```
print(regex.test)
```

```
## character(0)
```

9.3.3 Stop Script on Failure

```
if (length(regex.test) != 0){
    stop("REGEX VALIDATION FAILED: FILE NAMES NOT IN COMPLIANCE WITH CODEBOOK
    SCHEMA!")
    }
```

9.4 Execute Rename

```
file.rename(filenames.enhanced$oldname,
            filenames.enhanced$newname)
```

10 File Split Module

Multilingual original files need to be split into monolingual documents as most current natural language processing techniques work best with monolingual data.

10.1 Manual Coding

```
##########################################
###  HANDCODING OF  Split Instructions
##########################################
```

10.2 Read Instructions

```
split <- fread("data/CD-PCIJ_Source_Filenames-FullNames-SplitInstructions.csv",
                            header = TRUE)[,.(newname, split)]
```

10.3 No Split

These files will not be split. They are monolingual in the original.

```
nosplit <- split[split == "nosplit"]$newname
print(nosplit)
```

```
## [1] "PCIJ_A_03_Neuilly_BGR_GRC_1924-09-12_ANX_01_NA_NA_EN.pdf"
## [2] "PCIJ_A_07_GermanInterestsUpperSilesia_DEU_POL_1926-05-25_ANX_01_NA_NA_FR.
    pdf"
## [3] "PCIJ_AB_41_CustomsRegime_LNC_NA_1931-09-05_ANX_01_NA_NA_DE.pdf"
## [4] "PCIJ_AB_70_Meuse_NLD_BEL_1937-06-28_ANX_01_NA_NA_FR.pdf"
```

10.4 Custom Parameters for Split

These files require custom parameters for splitting due to their non-standard composition.

```
customsplit <- split[split == "customsplit"]$newname
print(customsplit)
```

```
## [1] "PCIJ_A_07_GermanInterestsUpperSilesia_DEU_POL_1926-02-05_DEC_01_JO_00_BI.
    pdf"
## [2] "PCIJ_A_07_GermanInterestsUpperSilesia_DEU_POL_1926-03-22_ORD_01_DH_00_BI.
    pdf"
## [3] "PCIJ_A_08_SinoBelgianTreaty_BEL_CHN_1926-11-25_APP_01_NA_NA_BI.pdf"
## [4] "PCIJ_AB_53_EasternGreenland_DNK_NOR_1933-04-05_ANX_01_NA_NA_BI.pdf"
## [5] "PCIJ_B_15_DanzigCourts_LNC_NA_1928-03-03_ANX_02_NA_NA_BI.pdf"
```

10.4.1 Greenland File

Remove PDF page 15 (internal page 130 is duplicate).

```
filename <- "PCIJ_AB_53_EasternGreenland_DNK_NOR_1933-04-05_ANX_01_NA_NA_BI.pdf"

file.temp <- paste0(filename,
                    "-temp")

file_move(filename,
          file.temp)

pdf_subset(file.temp,
           c(1:14,
             16:49),
           filename)
```

```
## [1] "PCIJ_AB_53_EasternGreenland_DNK_NOR_1933-04-05_ANX_01_NA_NA_BI.pdf"
```

```
unlink(file.temp)
```

10.4.2 Sino-Belgian Treaty File

This file should contain only the application to institute proceedings, but also contains the subsequent Order of 8 January 1927. As the Order is already contained in a separate file it will be removed from this one. The result is then listed for splitting with English on odd pages in the following step.

```
filename <- "PCIJ_A_08_SinoBelgianTreaty_BEL_CHN_1926-11-25_APP_01_NA_NA_BI.pdf"

file.temp <- paste0(filename, "-temp")

file_move(filename, file.temp)

pdf_subset(file.temp,
           1:5,
           filename)
```

```
## [1] "PCIJ_A_08_SinoBelgianTreaty_BEL_CHN_1926-11-25_APP_01_NA_NA_BI.pdf"
```

```
unlink(file.temp)
```

10.4.3 Silesia Files

These two files need to be manually recombined and then split again. The first file is missing the last English page, the second file is missing the first French page.

```
file.temp <- "combine-temp.pdf"

files.combine <- c("PCIJ_A_07_GermanInterestsUpperSilesia_DEU_POL_1926-02-05_DEC_
    01_JO_00_BI.pdf",
                   "PCIJ_A_07_GermanInterestsUpperSilesia_DEU_POL_1926-03-22_ORD_
    01_DH_00_BI.pdf")


pdf_combine(files.combine,
            file.temp)
```

```
## [1] "combine-temp.pdf"
```

```
pdf_subset(file.temp,
           c(1, 3, 5),
           "PCIJ_A_07_GermanInterestsUpperSilesia_DEU_POL_1926-02-05_DEC_01_JO_00
    _FR.pdf")
```

```
## [1] "PCIJ_A_07_GermanInterestsUpperSilesia_DEU_POL_1926-02-05_DEC_01_JO_00_FR.
    pdf"
```

```
pdf_subset(file.temp,
           c(2, 4, 6),
           "PCIJ_A_07_GermanInterestsUpperSilesia_DEU_POL_1926-02-05_DEC_01_JO_00
    _EN.pdf")
```

```
## [1] "PCIJ_A_07_GermanInterestsUpperSilesia_DEU_POL_1926-02-05_DEC_01_JO_00_EN.
    pdf"
```

```
pdf_subset(file.temp,
           c(5, 7),
           "PCIJ_A_07_GermanInterestsUpperSilesia_DEU_POL_1926-03-22_ORD_01_DH_00
    _FR.pdf")
```

```
## [1] "PCIJ_A_07_GermanInterestsUpperSilesia_DEU_POL_1926-03-22_ORD_01_DH_00_FR.
    pdf"
```

```
pdf_subset(file.temp,
           c(6, 8),
           "PCIJ_A_07_GermanInterestsUpperSilesia_DEU_POL_1926-03-22_ORD_01_DH_00
    _EN.pdf")
```

```
## [1] "PCIJ_A_07_GermanInterestsUpperSilesia_DEU_POL_1926-03-22_ORD_01_DH_00_EN.
    pdf"
```

```
unlink(file.temp)
```

10.4.4 Danzig File

French and German alternate in this file, not English and French. Split accordingly.

```
file <- "PCIJ_B_15_DanzigCourts_LNC_NA_1928-03-03_ANX_02_NA_NA_BI.pdf"

temp1 <- seq(1, pdf_length(file), 1)

even <- temp1[lapply(seq(1, max(temp1), 1), "%%", 2) == 0]
even.name <- gsub("BI\\.pdf",
                  "DE\\.pdf",
                  file)
pdf_subset(file,
           pages = even,
           output = even.name)
```

```
## [1] "PCIJ_B_15_DanzigCourts_LNC_NA_1928-03-03_ANX_02_NA_NA_DE.pdf"
```

```
odd <- temp1[lapply(seq(1, max(temp1), 1), "%%", 2) != 0]
odd.name <- gsub("BI\\.pdf",
                 "FR\\.pdf",
                 file)
pdf_subset(file,
           pages = odd,
           output = odd.name)
```

```
## [1] "PCIJ_B_15_DanzigCourts_LNC_NA_1928-03-03_ANX_02_NA_NA_FR.pdf"
```

10.5 Start Fork Cluster

```
cl <- makeForkCluster(fullCores)
registerDoParallel(cl)
```

10.6 English on Odd Pages

The following files will be split on the assumption that the English version is on odd-numbered pages:

```
odd.english <- split[split == "odd-english"]$newname

odd.english <- c(odd.english,
                 "PCIJ_A_08_SinoBelgianTreaty_BEL_CHN_1926-11-25_APP_01_NA_NA_BI.
    pdf")
```

10.6.1 Number of Files to Split

```
length(odd.english)
```

```
## [1] 76
```

10.6.2 Names of Files to Split

```
print(odd.english)
```

```
##  [1] "PCIJ_A_01_Wimbledon_GBR-FRA-ITA-JPN_DEU_1923-01-16_APP_01_NA_NA_BI.pdf"
##  [2] "PCIJ_A_02_MavrommatisPalestine_GRC_GBR_1924-08-30_JUD_01_PO_00_BI.pdf"
##  [3] "PCIJ_A_03_Neuilly_BGR_GRC_1924-09-12_JUD_01_ME_00_BI.pdf"
##  [4] "PCIJ_A_04_InterpretationNo3_BGR_GRC_1925-03-26_JUD_01_ME_00_BI.pdf"
##  [5] "PCIJ_A_05_MavrommatisJerusalem_GRC_GBR_1925-03-26_JUD_01_ME_00_BI.pdf"
##  [6] "PCIJ_A_06_GermanInterestsUpperSilesia_DEU_POL_1925-08-25_JUD_01_PO_00_BI
    .pdf"
##  [7] "PCIJ_A_07_GermanInterestsUpperSilesia_DEU_POL_1926-05-25_JUD_01_ME_00_BI
    .pdf"
##  [8] "PCIJ_A_09_ChorzowFactory_DEU_POL_1927-07-26_JUD_01_PO_00_BI.pdf"
##  [9] "PCIJ_A_10_Lotus_FRA_TUR_1927-09-07_JUD_01_ME_00_BI.pdf"
## [10] "PCIJ_A_11_MavrommatisJerusalem-Readaptation_GRC_GBR_1927-10-10_JUD_01_PO
    _00_BI.pdf"
## [11] "PCIJ_A_12_ChorzowFactory-Indemnities_DEU_POL_1927-11-21_ORD_01_IM_00_BI.
    pdf"
## [12] "PCIJ_A_12_ChorzowFactory-Indemnities_DEU_POL_1927-10-14_APP_01_NA_NA_BI.
    pdf"
## [13] "PCIJ_A_13_ChorzowFactory-Interpretation_DEU_POL_1927-12-16_JUD_01_ME_00_
    BI.pdf"
## [14] "PCIJ_A_14_SinoBelgianTreaty_BEL_CHN_1928-02-21_ORD_01_TL_00_BI.pdf"
## [15] "PCIJ_A_15_MinoritySchoolsUpperSilesia_DEU_POL_1928-04-26_JUD_01_ME_00_BI
    .pdf"
## [16] "PCIJ_A_16_SinoBelgianTreaty_BEL_CHN_1928-08-13_ORD_01_TL_00_BI.pdf"
## [17] "PCIJ_A_17_ChorzowFactory-Indemnities_DEU_POL_1928-09-13_JUD_01_ME_00_BI.
    pdf"
```

```
## [18] "PCIJ_A_18_SinoBelgianTreaty_BEL_CHN_1929-05-25_ORD_01_DI_00_BI.pdf"
## [19] "PCIJ_A_20_SerbianLoans_FRA_YUG_1929-07-12_JUD_01_ME_00_BI.pdf"
## [20] "PCIJ_A_22_FreeZonesUpperSavoyGex_FRA_CHE_1929-08-19_ORD_01_EV-SE_00_BI.
    pdf"
## [21] "PCIJ_A_23_OderCommission_GBR-CSK-DNK-FRA-DEU_POL_1929-09-10_JUD_01_ME
    _00_BI.pdf"
## [22] "PCIJ_A_24_FreeZonesUpperSavoyGex-SecondPhase_FRA_CHE_1930-12-06_ORD_01_
    SE_00_BI.pdf"
## [23] "PCIJ_AB_40_GermanMinoritySchools_DEU_POL_1931-05-15_ADV_01_NA_00_BI.pdf"
## [24] "PCIJ_AB_41_CustomsRegime_LNC_NA_1931-09-05_ADV_01_NA_00_BI.pdf"
## [25] "PCIJ_AB_42_RailwayTraffic_LNC_NA_1931-10-15_ADV_01_NA_00_BI.pdf"
## [26] "PCIJ_AB_43_PolishWarVessels_LNC_NA_1931-12-11_ADV_01_NA_00_BI.pdf"
## [27] "PCIJ_AB_44_TreatmentPolishNationals_LNC_NA_1932-02-04_ADV_01_NA_00_BI.
    pdf"
## [28] "PCIJ_AB_45_GrecoBulgarianAgreement_LNC_NA_1932-03-08_ADV_01_NA_00_BI.pdf
    "
## [29] "PCIJ_AB_46_FreeZonesUpperSavoyGex_FRA_CHE_1932-06-07_JUD_01_ME_00_BI.pdf
    "
## [30] "PCIJ_AB_48_EasternGreenland_DNK_NOR_1932-08-02_ORD_01_JO-TL_00_BI.pdf"
## [31] "PCIJ_AB_48_EasternGreenland_DNK_NOR_1932-08-03_ORD_01_IM_00_BI.pdf"
## [32] "PCIJ_AB_49_InterpretationMemelStatute_GBR-FRA-ITA-JPN_LTU_1932-08-11_JUD
    _01_ME_00_BI.pdf"
## [33] "PCIJ_AB_50_WomenNightWork_LNC_NA_1932-11-15_ADV_01_NA_00_BI.pdf"
## [34] "PCIJ_AB_51_Castellorizo_TUR_ITA_1933-01-26_ORD_01_DI_00_BI.pdf"
## [35] "PCIJ_AB_52_PrinceVonPless_DEU_POL_1933-02-04_ORD_01_PO_00_BI.pdf"
## [36] "PCIJ_AB_53_EasternGreenland_DNK_NOR_1933-04-05_JUD_01_ME_00_BI.pdf"
## [37] "PCIJ_AB_54_PrinceVonPless_DEU_POL_1933-05-11_ORD_01_IM_00_BI.pdf"
## [38] "PCIJ_AB_55_EasternGreenland_DNK_NOR_1933-05-11_ORD_01_DI_00_BI.pdf"
## [39] "PCIJ_AB_56_HungaroCzechoslovakMixedTribunal_CSK_HUN_1933-05-12_ORD_01_DI
    _00_BI.pdf"
## [40] "PCIJ_AB_57_PrinceVonPless_DEU_POL_1933-07-04_ORD_01_PR_00_BI.pdf"
## [41] "PCIJ_AB_58_PolishAgrarianReform_DEU_POL_1933-07-29_ORD_01_IM_00_BI.pdf"
## [42] "PCIJ_AB_59_PrinceVonPless_DEU_POL_1933-12-02_ORD_01_DI_00_BI.pdf"
## [43] "PCIJ_AB_60_PolishAgrarianReform_DEU_POL_1933-12-02_ORD_01_DI_00_BI.pdf"
## [44] "PCIJ_AB_61_Pazmany_CSK_HUN_1933-12-15_JUD_01_ME_00_BI.pdf"
## [45] "PCIJ_AB_62_Lighthouses_FRA_GRC_1934-03-17_JUD_01_ME_00_BI.pdf"
## [46] "PCIJ_AB_63_OscarChinn_GBR_BEL_1934-12-12_JUD_01_ME_00_BI.pdf"
## [47] "PCIJ_AB_64_MinoritySchoolsAlbania_LNC_NA_1935-04-06_ADV_01_NA_00_BI.pdf"
## [48] "PCIJ_AB_65_DanzigLegislativeDecrees_LNC_NA_1935-12-04_ADV_01_NA_00_BI.
    pdf"
## [49] "PCIJ_AB_66_PajzsCsakyEsterhazy_HUN_YUG_1936-05-23_ORD_01_PO_00_BI.pdf"
## [50] "PCIJ_AB_67_Losinger_CHE_YUG_1936-06-27_ORD_01_PO_00_BI.pdf"
## [51] "PCIJ_AB_68_PajzsCsakyEsterhazy_HUN_YUG_1936-12-16_JUD_01_ME_00_BI.pdf"
## [52] "PCIJ_AB_69_Losinger_CHE_YUG_1936-12-14_ORD_01_DI_00_BI.pdf"
## [53] "PCIJ_AB_70_Meuse_NLD_BEL_1937-06-28_JUD_01_ME_00_BI.pdf"
## [54] "PCIJ_AB_71_LighthousesCreteSamos_FRA_GRC_1937-10-08_JUD_01_ME_00_BI.pdf"
## [55] "PCIJ_AB_72_Borchgrave_BEL_ESP_1937-11-06_JUD_01_PO_00_BI.pdf"
## [56] "PCIJ_AB_73_Borchgrave_BEL_ESP_1938-04-30_ORD_01_DI_00_BI.pdf"
## [57] "PCIJ_AB_74_PhosphatesMarocco_ITA_FRA_1938-06-14_JUD_01_PO_00_BI.pdf"
## [58] "PCIJ_AB_75_PanevezysSaldutiskisRailway_EST_LTU_1938-06-30_ORD_01_PO_00_
    BI.pdf"
## [59] "PCIJ_AB_76_PanevezysSaldutiskisRailway_EST_LTU_1939-02-28_JUD_01_ME_00_
    BI.pdf"
## [60] "PCIJ_AB_77_ElectricityCompanySofiaBulgaria_BEL_BGR_1939-04-04_JUD_01_PO
    _00_BI.pdf"
## [61] "PCIJ_AB_78_SocieteCommercialeBelgique_BEL_GRC_1939-06-15_JUD_01_ME_00_BI
```

```
    .pdf"
## [62] "PCIJ_AB_79_ElectricityCompanySofiaBulgaria_BEL_BGR_1939-12-05_ORD_01_IM
    _00_BI.pdf"
## [63] "PCIJ_AB_80_ElectricityCompanySofiaBulgaria_BEL_BGR_1940-02-26_ORD_01_DH
    _00_BI.pdf"
## [64] "PCIJ_B_06_GermanSettlers_LNC_NA_1923-09-10_ADV_01_NA_00_BI.pdf"
## [65] "PCIJ_B_07_AcquisitionPolishNationality_LNC_NA_1923-09-15_ADV_01_NA_00_BI
    .pdf"
## [66] "PCIJ_B_08_Jaworzina_LNC_NA_1923-12-06_ADV_01_NA_00_BI.pdf"
## [67] "PCIJ_B_10_ExchangeGreekTurkishPopulations_LNC_NA_1925-02-21_ADV_01_NA
    _00_BI.pdf"
## [68] "PCIJ_B_11_PostalServiceDanzig_LNC_NA_1925-05-16_ADV_01_NA_00_BI.pdf"
## [69] "PCIJ_B_12_TreatyLausanne_LNC_NA_1925-11-21_ADV_01_NA_00_BI.pdf"
## [70] "PCIJ_B_13_ILOCompetenceEmployer_LNC_NA_1926-07-23_ADV_01_NA_00_BI.pdf"
## [71] "PCIJ_B_14_DanubeCommission_LNC_NA_1927-12-08_ADV_01_NA_00_BI.pdf"
## [72] "PCIJ_B_15_DanzigCourts_LNC_NA_1928-03-03_ADV_01_NA_00_BI.pdf"
## [73] "PCIJ_B_16_GrecoTurkishAgreement_LNC_NA_1928-08-28_ADV_01_NA_00_BI.pdf"
## [74] "PCIJ_B_17_GrecoBulgarianCommunities_LNC_NA_1930-07-31_ADV_01_NA_00_BI.
    pdf"
## [75] "PCIJ_B_18_DanzigILO_LNC_NA_1930-08-26_ADV_01_NA_00_BI.pdf"
## [76] "PCIJ_A_08_SinoBelgianTreaty_BEL_CHN_1926-11-25_APP_01_NA_NA_BI.pdf"
```

10.6.3 Execute Split

```
out <- foreach(file = odd.english,
               .errorhandling = 'pass',
               .combine = c) %dopar% {

    out1 <- vector(mode = "list",
                   length = 2)

    temp1 <- seq(1, pdf_length(file), 1)


    even <- temp1[lapply(seq(1, max(temp1), 1), "%%", 2) == 0]
    even.name <- gsub("BI\\.pdf",
                      "FR\\.pdf",
                      file)
    out1[[1]] <- pdf_subset(file,
                            pages = even,
                            output = even.name)


    odd <- temp1[lapply(seq(1, max(temp1), 1), "%%", 2) != 0]
    odd.name <- gsub("BI\\.pdf",
                     "EN\\.pdf",
                     file)
    out1[[2]] <- pdf_subset(file,
                            pages = odd,
                            output = odd.name)

    return(out1)
}
```

10.6.4 Print Split Results

```r
print(unlist(out))
```

```
##   [1] "PCIJ_A_01_Wimbledon_GBR-FRA-ITA-JPN_DEU_1923-01-16_APP_01_NA_NA_FR.pdf"
##   [2] "PCIJ_A_01_Wimbledon_GBR-FRA-ITA-JPN_DEU_1923-01-16_APP_01_NA_NA_EN.pdf"
##   [3] "PCIJ_A_02_MavrommatisPalestine_GRC_GBR_1924-08-30_JUD_01_PO_00_FR.pdf"
##   [4] "PCIJ_A_02_MavrommatisPalestine_GRC_GBR_1924-08-30_JUD_01_PO_00_EN.pdf"
##   [5] "PCIJ_A_03_Neuilly_BGR_GRC_1924-09-12_JUD_01_ME_00_FR.pdf"
##   [6] "PCIJ_A_03_Neuilly_BGR_GRC_1924-09-12_JUD_01_ME_00_EN.pdf"
##   [7] "PCIJ_A_04_InterpretationNo3_BGR_GRC_1925-03-26_JUD_01_ME_00_FR.pdf"
##   [8] "PCIJ_A_04_InterpretationNo3_BGR_GRC_1925-03-26_JUD_01_ME_00_EN.pdf"
##   [9] "PCIJ_A_05_MavrommatisJerusalem_GRC_GBR_1925-03-26_JUD_01_ME_00_FR.pdf"
##  [10] "PCIJ_A_05_MavrommatisJerusalem_GRC_GBR_1925-03-26_JUD_01_ME_00_EN.pdf"
##  [11] "PCIJ_A_06_GermanInterestsUpperSilesia_DEU_POL_1925-08-25_JUD_01_PO_00_
    FR.pdf"
##  [12] "PCIJ_A_06_GermanInterestsUpperSilesia_DEU_POL_1925-08-25_JUD_01_PO_00_
    EN.pdf"
##  [13] "PCIJ_A_07_GermanInterestsUpperSilesia_DEU_POL_1926-05-25_JUD_01_ME_00_
    FR.pdf"
##  [14] "PCIJ_A_07_GermanInterestsUpperSilesia_DEU_POL_1926-05-25_JUD_01_ME_00_
    EN.pdf"
##  [15] "PCIJ_A_09_ChorzowFactory_DEU_POL_1927-07-26_JUD_01_PO_00_FR.pdf"
##  [16] "PCIJ_A_09_ChorzowFactory_DEU_POL_1927-07-26_JUD_01_PO_00_EN.pdf"
##  [17] "PCIJ_A_10_Lotus_FRA_TUR_1927-09-07_JUD_01_ME_00_FR.pdf"
##  [18] "PCIJ_A_10_Lotus_FRA_TUR_1927-09-07_JUD_01_ME_00_EN.pdf"
##  [19] "PCIJ_A_11_MavrommatisJerusalem-Readaptation_GRC_GBR_1927-10-10_JUD_01_
    PO_00_FR.pdf"
##  [20] "PCIJ_A_11_MavrommatisJerusalem-Readaptation_GRC_GBR_1927-10-10_JUD_01_
    PO_00_EN.pdf"
##  [21] "PCIJ_A_12_ChorzowFactory-Indemnities_DEU_POL_1927-11-21_ORD_01_IM_00_FR
    .pdf"
##  [22] "PCIJ_A_12_ChorzowFactory-Indemnities_DEU_POL_1927-11-21_ORD_01_IM_00_EN
    .pdf"
##  [23] "PCIJ_A_12_ChorzowFactory-Indemnities_DEU_POL_1927-10-14_APP_01_NA_NA_FR
    .pdf"
##  [24] "PCIJ_A_12_ChorzowFactory-Indemnities_DEU_POL_1927-10-14_APP_01_NA_NA_EN
    .pdf"
##  [25] "PCIJ_A_13_ChorzowFactory-Interpretation_DEU_POL_1927-12-16_JUD_01_ME
    _00_FR.pdf"
##  [26] "PCIJ_A_13_ChorzowFactory-Interpretation_DEU_POL_1927-12-16_JUD_01_ME
    _00_EN.pdf"
##  [27] "PCIJ_A_14_SinoBelgianTreaty_BEL_CHN_1928-02-21_ORD_01_TL_00_FR.pdf"
##  [28] "PCIJ_A_14_SinoBelgianTreaty_BEL_CHN_1928-02-21_ORD_01_TL_00_EN.pdf"
##  [29] "PCIJ_A_15_MinoritySchoolsUpperSilesia_DEU_POL_1928-04-26_JUD_01_ME_00_
    FR.pdf"
##  [30] "PCIJ_A_15_MinoritySchoolsUpperSilesia_DEU_POL_1928-04-26_JUD_01_ME_00_
    EN.pdf"
##  [31] "PCIJ_A_16_SinoBelgianTreaty_BEL_CHN_1928-08-13_ORD_01_TL_00_FR.pdf"
##  [32] "PCIJ_A_16_SinoBelgianTreaty_BEL_CHN_1928-08-13_ORD_01_TL_00_EN.pdf"
##  [33] "PCIJ_A_17_ChorzowFactory-Indemnities_DEU_POL_1928-09-13_JUD_01_ME_00_FR
    .pdf"
```

```
## [34] "PCIJ_A_17_ChorzowFactory-Indemnities_DEU_POL_1928-09-13_JUD_01_ME_00_EN
    .pdf"
## [35] "PCIJ_A_18_SinoBelgianTreaty_BEL_CHN_1929-05-25_ORD_01_DI_00_FR.pdf"
## [36] "PCIJ_A_18_SinoBelgianTreaty_BEL_CHN_1929-05-25_ORD_01_DI_00_EN.pdf"
## [37] "PCIJ_A_20_SerbianLoans_FRA_YUG_1929-07-12_JUD_01_ME_00_FR.pdf"
## [38] "PCIJ_A_20_SerbianLoans_FRA_YUG_1929-07-12_JUD_01_ME_00_EN.pdf"
## [39] "PCIJ_A_22_FreeZonesUpperSavoyGex_FRA_CHE_1929-08-19_ORD_01_EV-SE_00_FR.
    pdf"
## [40] "PCIJ_A_22_FreeZonesUpperSavoyGex_FRA_CHE_1929-08-19_ORD_01_EV-SE_00_EN.
    pdf"
## [41] "PCIJ_A_23_OderCommission_GBR-CSK-DNK-FRA-DEU_POL_1929-09-10_JUD_01_ME
    _00_FR.pdf"
## [42] "PCIJ_A_23_OderCommission_GBR-CSK-DNK-FRA-DEU_POL_1929-09-10_JUD_01_ME
    _00_EN.pdf"
## [43] "PCIJ_A_24_FreeZonesUpperSavoyGex-SecondPhase_FRA_CHE_1930-12-06_ORD_01_
    SE_00_FR.pdf"
## [44] "PCIJ_A_24_FreeZonesUpperSavoyGex-SecondPhase_FRA_CHE_1930-12-06_ORD_01_
    SE_00_EN.pdf"
## [45] "PCIJ_AB_40_GermanMinoritySchools_DEU_POL_1931-05-15_ADV_01_NA_00_FR.pdf
    "
## [46] "PCIJ_AB_40_GermanMinoritySchools_DEU_POL_1931-05-15_ADV_01_NA_00_EN.pdf
    "
## [47] "PCIJ_AB_41_CustomsRegime_LNC_NA_1931-09-05_ADV_01_NA_00_FR.pdf"
## [48] "PCIJ_AB_41_CustomsRegime_LNC_NA_1931-09-05_ADV_01_NA_00_EN.pdf"
## [49] "PCIJ_AB_42_RailwayTraffic_LNC_NA_1931-10-15_ADV_01_NA_00_FR.pdf"
## [50] "PCIJ_AB_42_RailwayTraffic_LNC_NA_1931-10-15_ADV_01_NA_00_EN.pdf"
## [51] "PCIJ_AB_43_PolishWarVessels_LNC_NA_1931-12-11_ADV_01_NA_00_FR.pdf"
## [52] "PCIJ_AB_43_PolishWarVessels_LNC_NA_1931-12-11_ADV_01_NA_00_EN.pdf"
## [53] "PCIJ_AB_44_TreatmentPolishNationals_LNC_NA_1932-02-04_ADV_01_NA_00_FR.
    pdf"
## [54] "PCIJ_AB_44_TreatmentPolishNationals_LNC_NA_1932-02-04_ADV_01_NA_00_EN.
    pdf"
## [55] "PCIJ_AB_45_GrecoBulgarianAgreement_LNC_NA_1932-03-08_ADV_01_NA_00_FR.
    pdf"
## [56] "PCIJ_AB_45_GrecoBulgarianAgreement_LNC_NA_1932-03-08_ADV_01_NA_00_EN.
    pdf"
## [57] "PCIJ_AB_46_FreeZonesUpperSavoyGex_FRA_CHE_1932-06-07_JUD_01_ME_00_FR.
    pdf"
## [58] "PCIJ_AB_46_FreeZonesUpperSavoyGex_FRA_CHE_1932-06-07_JUD_01_ME_00_EN.
    pdf"
## [59] "PCIJ_AB_48_EasternGreenland_DNK_NOR_1932-08-02_ORD_01_JO-TL_00_FR.pdf"
## [60] "PCIJ_AB_48_EasternGreenland_DNK_NOR_1932-08-02_ORD_01_JO-TL_00_EN.pdf"
## [61] "PCIJ_AB_48_EasternGreenland_DNK_NOR_1932-08-03_ORD_01_IM_00_FR.pdf"
## [62] "PCIJ_AB_48_EasternGreenland_DNK_NOR_1932-08-03_ORD_01_IM_00_EN.pdf"
## [63] "PCIJ_AB_49_InterpretationMemelStatute_GBR-FRA-ITA-JPN_LTU_1932-08-11_
    JUD_01_ME_00_FR.pdf"
## [64] "PCIJ_AB_49_InterpretationMemelStatute_GBR-FRA-ITA-JPN_LTU_1932-08-11_
    JUD_01_ME_00_EN.pdf"
## [65] "PCIJ_AB_50_WomenNightWork_LNC_NA_1932-11-15_ADV_01_NA_00_FR.pdf"
## [66] "PCIJ_AB_50_WomenNightWork_LNC_NA_1932-11-15_ADV_01_NA_00_EN.pdf"
## [67] "PCIJ_AB_51_Castellorizo_TUR_ITA_1933-01-26_ORD_01_DI_00_FR.pdf"
## [68] "PCIJ_AB_51_Castellorizo_TUR_ITA_1933-01-26_ORD_01_DI_00_EN.pdf"
## [69] "PCIJ_AB_52_PrinceVonPless_DEU_POL_1933-02-04_ORD_01_PO_00_FR.pdf"
## [70] "PCIJ_AB_52_PrinceVonPless_DEU_POL_1933-02-04_ORD_01_PO_00_EN.pdf"
## [71] "PCIJ_AB_53_EasternGreenland_DNK_NOR_1933-04-05_JUD_01_ME_00_FR.pdf"
## [72] "PCIJ_AB_53_EasternGreenland_DNK_NOR_1933-04-05_JUD_01_ME_00_EN.pdf"
```

```
##  [73] "PCIJ_AB_54_PrinceVonPless_DEU_POL_1933-05-11_ORD_01_IM_00_FR.pdf"
##  [74] "PCIJ_AB_54_PrinceVonPless_DEU_POL_1933-05-11_ORD_01_IM_00_EN.pdf"
##  [75] "PCIJ_AB_55_EasternGreenland_DNK_NOR_1933-05-11_ORD_01_DI_00_FR.pdf"
##  [76] "PCIJ_AB_55_EasternGreenland_DNK_NOR_1933-05-11_ORD_01_DI_00_EN.pdf"
##  [77] "PCIJ_AB_56_HungaroCzechoslovakMixedTribunal_CSK_HUN_1933-05-12_ORD_01_
    DI_00_FR.pdf"
##  [78] "PCIJ_AB_56_HungaroCzechoslovakMixedTribunal_CSK_HUN_1933-05-12_ORD_01_
    DI_00_EN.pdf"
##  [79] "PCIJ_AB_57_PrinceVonPless_DEU_POL_1933-07-04_ORD_01_PR_00_FR.pdf"
##  [80] "PCIJ_AB_57_PrinceVonPless_DEU_POL_1933-07-04_ORD_01_PR_00_EN.pdf"
##  [81] "PCIJ_AB_58_PolishAgrarianReform_DEU_POL_1933-07-29_ORD_01_IM_00_FR.pdf"
##  [82] "PCIJ_AB_58_PolishAgrarianReform_DEU_POL_1933-07-29_ORD_01_IM_00_EN.pdf"
##  [83] "PCIJ_AB_59_PrinceVonPless_DEU_POL_1933-12-02_ORD_01_DI_00_FR.pdf"
##  [84] "PCIJ_AB_59_PrinceVonPless_DEU_POL_1933-12-02_ORD_01_DI_00_EN.pdf"
##  [85] "PCIJ_AB_60_PolishAgrarianReform_DEU_POL_1933-12-02_ORD_01_DI_00_FR.pdf"
##  [86] "PCIJ_AB_60_PolishAgrarianReform_DEU_POL_1933-12-02_ORD_01_DI_00_EN.pdf"
##  [87] "PCIJ_AB_61_Pazmany_CSK_HUN_1933-12-15_JUD_01_ME_00_FR.pdf"
##  [88] "PCIJ_AB_61_Pazmany_CSK_HUN_1933-12-15_JUD_01_ME_00_EN.pdf"
##  [89] "PCIJ_AB_62_Lighthouses_FRA_GRC_1934-03-17_JUD_01_ME_00_FR.pdf"
##  [90] "PCIJ_AB_62_Lighthouses_FRA_GRC_1934-03-17_JUD_01_ME_00_EN.pdf"
##  [91] "PCIJ_AB_63_OscarChinn_GBR_BEL_1934-12-12_JUD_01_ME_00_FR.pdf"
##  [92] "PCIJ_AB_63_OscarChinn_GBR_BEL_1934-12-12_JUD_01_ME_00_EN.pdf"
##  [93] "PCIJ_AB_64_MinoritySchoolsAlbania_LNC_NA_1935-04-06_ADV_01_NA_00_FR.pdf
    "
##  [94] "PCIJ_AB_64_MinoritySchoolsAlbania_LNC_NA_1935-04-06_ADV_01_NA_00_EN.pdf
    "
##  [95] "PCIJ_AB_65_DanzigLegislativeDecrees_LNC_NA_1935-12-04_ADV_01_NA_00_FR.
    pdf"
##  [96] "PCIJ_AB_65_DanzigLegislativeDecrees_LNC_NA_1935-12-04_ADV_01_NA_00_EN.
    pdf"
##  [97] "PCIJ_AB_66_PajzsCsakyEsterhazy_HUN_YUG_1936-05-23_ORD_01_PO_00_FR.pdf"
##  [98] "PCIJ_AB_66_PajzsCsakyEsterhazy_HUN_YUG_1936-05-23_ORD_01_PO_00_EN.pdf"
##  [99] "PCIJ_AB_67_Losinger_CHE_YUG_1936-06-27_ORD_01_PO_00_FR.pdf"
## [100] "PCIJ_AB_67_Losinger_CHE_YUG_1936-06-27_ORD_01_PO_00_EN.pdf"
## [101] "PCIJ_AB_68_PajzsCsakyEsterhazy_HUN_YUG_1936-12-16_JUD_01_ME_00_FR.pdf"
## [102] "PCIJ_AB_68_PajzsCsakyEsterhazy_HUN_YUG_1936-12-16_JUD_01_ME_00_EN.pdf"
## [103] "PCIJ_AB_69_Losinger_CHE_YUG_1936-12-14_ORD_01_DI_00_FR.pdf"
## [104] "PCIJ_AB_69_Losinger_CHE_YUG_1936-12-14_ORD_01_DI_00_EN.pdf"
## [105] "PCIJ_AB_70_Meuse_NLD_BEL_1937-06-28_JUD_01_ME_00_FR.pdf"
## [106] "PCIJ_AB_70_Meuse_NLD_BEL_1937-06-28_JUD_01_ME_00_EN.pdf"
## [107] "PCIJ_AB_71_LighthousesCreteSamos_FRA_GRC_1937-10-08_JUD_01_ME_00_FR.pdf
    "
## [108] "PCIJ_AB_71_LighthousesCreteSamos_FRA_GRC_1937-10-08_JUD_01_ME_00_EN.pdf
    "
## [109] "PCIJ_AB_72_Borchgrave_BEL_ESP_1937-11-06_JUD_01_PO_00_FR.pdf"
## [110] "PCIJ_AB_72_Borchgrave_BEL_ESP_1937-11-06_JUD_01_PO_00_EN.pdf"
## [111] "PCIJ_AB_73_Borchgrave_BEL_ESP_1938-04-30_ORD_01_DI_00_FR.pdf"
## [112] "PCIJ_AB_73_Borchgrave_BEL_ESP_1938-04-30_ORD_01_DI_00_EN.pdf"
## [113] "PCIJ_AB_74_PhosphatesMarocco_ITA_FRA_1938-06-14_JUD_01_PO_00_FR.pdf"
## [114] "PCIJ_AB_74_PhosphatesMarocco_ITA_FRA_1938-06-14_JUD_01_PO_00_EN.pdf"
## [115] "PCIJ_AB_75_PanevezysSaldutiskisRailway_EST_LTU_1938-06-30_ORD_01_PO_00_
    FR.pdf"
## [116] "PCIJ_AB_75_PanevezysSaldutiskisRailway_EST_LTU_1938-06-30_ORD_01_PO_00_
    EN.pdf"
## [117] "PCIJ_AB_76_PanevezysSaldutiskisRailway_EST_LTU_1939-02-28_JUD_01_ME_00_
    FR.pdf"
```

```
## [118] "PCIJ_AB_76_PanevezysSaldutiskisRailway_EST_LTU_1939-02-28_JUD_01_ME_00_
    EN.pdf"
## [119] "PCIJ_AB_77_ElectricityCompanySofiaBulgaria_BEL_BGR_1939-04-04_JUD_01_PO
    _00_FR.pdf"
## [120] "PCIJ_AB_77_ElectricityCompanySofiaBulgaria_BEL_BGR_1939-04-04_JUD_01_PO
    _00_EN.pdf"
## [121] "PCIJ_AB_78_SocieteCommercialeBelgique_BEL_GRC_1939-06-15_JUD_01_ME_00_
    FR.pdf"
## [122] "PCIJ_AB_78_SocieteCommercialeBelgique_BEL_GRC_1939-06-15_JUD_01_ME_00_
    EN.pdf"
## [123] "PCIJ_AB_79_ElectricityCompanySofiaBulgaria_BEL_BGR_1939-12-05_ORD_01_IM
    _00_FR.pdf"
## [124] "PCIJ_AB_79_ElectricityCompanySofiaBulgaria_BEL_BGR_1939-12-05_ORD_01_IM
    _00_EN.pdf"
## [125] "PCIJ_AB_80_ElectricityCompanySofiaBulgaria_BEL_BGR_1940-02-26_ORD_01_DH
    _00_FR.pdf"
## [126] "PCIJ_AB_80_ElectricityCompanySofiaBulgaria_BEL_BGR_1940-02-26_ORD_01_DH
    _00_EN.pdf"
## [127] "PCIJ_B_06_GermanSettlers_LNC_NA_1923-09-10_ADV_01_NA_00_FR.pdf"
## [128] "PCIJ_B_06_GermanSettlers_LNC_NA_1923-09-10_ADV_01_NA_00_EN.pdf"
## [129] "PCIJ_B_07_AcquisitionPolishNationality_LNC_NA_1923-09-15_ADV_01_NA_00_
    FR.pdf"
## [130] "PCIJ_B_07_AcquisitionPolishNationality_LNC_NA_1923-09-15_ADV_01_NA_00_
    EN.pdf"
## [131] "PCIJ_B_08_Jaworzina_LNC_NA_1923-12-06_ADV_01_NA_00_FR.pdf"
## [132] "PCIJ_B_08_Jaworzina_LNC_NA_1923-12-06_ADV_01_NA_00_EN.pdf"
## [133] "PCIJ_B_10_ExchangeGreekTurkishPopulations_LNC_NA_1925-02-21_ADV_01_NA
    _00_FR.pdf"
## [134] "PCIJ_B_10_ExchangeGreekTurkishPopulations_LNC_NA_1925-02-21_ADV_01_NA
    _00_EN.pdf"
## [135] "PCIJ_B_11_PostalServiceDanzig_LNC_NA_1925-05-16_ADV_01_NA_00_FR.pdf"
## [136] "PCIJ_B_11_PostalServiceDanzig_LNC_NA_1925-05-16_ADV_01_NA_00_EN.pdf"
## [137] "PCIJ_B_12_TreatyLausanne_LNC_NA_1925-11-21_ADV_01_NA_00_FR.pdf"
## [138] "PCIJ_B_12_TreatyLausanne_LNC_NA_1925-11-21_ADV_01_NA_00_EN.pdf"
## [139] "PCIJ_B_13_ILOCompetenceEmployer_LNC_NA_1926-07-23_ADV_01_NA_00_FR.pdf"
## [140] "PCIJ_B_13_ILOCompetenceEmployer_LNC_NA_1926-07-23_ADV_01_NA_00_EN.pdf"
## [141] "PCIJ_B_14_DanubeCommission_LNC_NA_1927-12-08_ADV_01_NA_00_FR.pdf"
## [142] "PCIJ_B_14_DanubeCommission_LNC_NA_1927-12-08_ADV_01_NA_00_EN.pdf"
## [143] "PCIJ_B_15_DanzigCourts_LNC_NA_1928-03-03_ADV_01_NA_00_FR.pdf"
## [144] "PCIJ_B_15_DanzigCourts_LNC_NA_1928-03-03_ADV_01_NA_00_EN.pdf"
## [145] "PCIJ_B_16_GrecoTurkishAgreement_LNC_NA_1928-08-28_ADV_01_NA_00_FR.pdf"
## [146] "PCIJ_B_16_GrecoTurkishAgreement_LNC_NA_1928-08-28_ADV_01_NA_00_EN.pdf"
## [147] "PCIJ_B_17_GrecoBulgarianCommunities_LNC_NA_1930-07-31_ADV_01_NA_00_FR.
    pdf"
## [148] "PCIJ_B_17_GrecoBulgarianCommunities_LNC_NA_1930-07-31_ADV_01_NA_00_EN.
    pdf"
## [149] "PCIJ_B_18_DanzigILO_LNC_NA_1930-08-26_ADV_01_NA_00_FR.pdf"
## [150] "PCIJ_B_18_DanzigILO_LNC_NA_1930-08-26_ADV_01_NA_00_EN.pdf"
## [151] "PCIJ_A_08_SinoBelgianTreaty_BEL_CHN_1926-11-25_APP_01_NA_NA_FR.pdf"
## [152] "PCIJ_A_08_SinoBelgianTreaty_BEL_CHN_1926-11-25_APP_01_NA_NA_EN.pdf"
```

10.7 English on Even Pages

The following files will be split on the assumption that the English version is on even-numbered pages:

```r
even.english <- split[split == "even-english"]$newname
```

10.7.1 Number of Files to Split

```r
length(even.english)
```

```
## [1] 180
```

10.7.2 Names of Files to Split

```r
print(even.english)
```

```
##   [1] "PCIJ_A_01_Wimbledon_GBR-FRA-ITA-JPN_DEU_1923-05-22_APP_01_NA_NA_BI.pdf"
##   [2] "PCIJ_A_01_Wimbledon_GBR-FRA-ITA-JPN_DEU_1923-06-28_JUD_01_IN_00_BI.pdf"
##   [3] "PCIJ_A_01_Wimbledon_GBR-FRA-ITA-JPN_DEU_1923-08-17_JUD_01_ME_00_BI.pdf"
##   [4] "PCIJ_A_01_Wimbledon_GBR-FRA-ITA-JPN_DEU_1923-08-17_JUD_01_ME_01_BI.pdf"
##   [5] "PCIJ_A_01_Wimbledon_GBR-FRA-ITA-JPN_DEU_1923-08-17_JUD_01_ME_02_BI.pdf"
##   [6] "PCIJ_A_02_MavrommatisPalestine_GRC_GBR_1924-08-30_JUD_01_PO_01_BI.pdf"
##   [7] "PCIJ_A_02_MavrommatisPalestine_GRC_GBR_1924-08-30_JUD_01_PO_02_BI.pdf"
##   [8] "PCIJ_A_02_MavrommatisPalestine_GRC_GBR_1924-08-30_JUD_01_PO_03_BI.pdf"
##   [9] "PCIJ_A_02_MavrommatisPalestine_GRC_GBR_1924-08-30_JUD_01_PO_04_BI.pdf"
##  [10] "PCIJ_A_02_MavrommatisPalestine_GRC_GBR_1924-08-30_JUD_01_PO_05_BI.pdf"
##  [11] "PCIJ_A_06_GermanInterestsUpperSilesia_DEU_POL_1925-08-25_JUD_01_PO_01_
    BI.pdf"
##  [12] "PCIJ_A_06_GermanInterestsUpperSilesia_DEU_POL_1925-08-25_JUD_01_PO_02_
    BI.pdf"
##  [13] "PCIJ_A_07_GermanInterestsUpperSilesia_DEU_POL_1926-05-25_JUD_01_ME_01_
    BI.pdf"
##  [14] "PCIJ_A_07_GermanInterestsUpperSilesia_DEU_POL_1926-05-25_JUD_01_ME_02_
    BI.pdf"
##  [15] "PCIJ_A_08_SinoBelgianTreaty_BEL_CHN_1927-01-08_ORD_01_IM_00_BI.pdf"
##  [16] "PCIJ_A_08_SinoBelgianTreaty_BEL_CHN_1927-02-15_ORD_01_IM_00_BI.pdf"
##  [17] "PCIJ_A_08_SinoBelgianTreaty_BEL_CHN_1927-06-18_ORD_01_TL_00_BI.pdf"
##  [18] "PCIJ_A_09_ChorzowFactory_DEU_POL_1927-07-26_JUD_01_PO_01_BI.pdf"
##  [19] "PCIJ_A_10_Lotus_FRA_TUR_1927-09-07_JUD_01_ME_01_BI.pdf"
##  [20] "PCIJ_A_10_Lotus_FRA_TUR_1927-09-07_JUD_01_ME_02_BI.pdf"
##  [21] "PCIJ_A_10_Lotus_FRA_TUR_1927-09-07_JUD_01_ME_03_BI.pdf"
##  [22] "PCIJ_A_10_Lotus_FRA_TUR_1927-09-07_JUD_01_ME_04_BI.pdf"
##  [23] "PCIJ_A_10_Lotus_FRA_TUR_1927-09-07_JUD_01_ME_05_BI.pdf"
##  [24] "PCIJ_A_10_Lotus_FRA_TUR_1927-09-07_JUD_01_ME_06_BI.pdf"
##  [25] "PCIJ_A_10_Lotus_FRA_TUR_1927-09-07_ANX_01_NA_NA_BI.pdf"
##  [26] "PCIJ_A_11_MavrommatisJerusalem-Readaptation_GRC_GBR_1927-10-10_JUD_01_
    PO_01_BI.pdf"
##  [27] "PCIJ_A_11_MavrommatisJerusalem-Readaptation_GRC_GBR_1927-10-10_JUD_01_
    PO_02_BI.pdf"
##  [28] "PCIJ_A_11_MavrommatisJerusalem-Readaptation_GRC_GBR_1927-10-10_JUD_01_
    PO_03_BI.pdf"
```

```
## [29] "PCIJ_A_13_ChorzowFactory-Interpretation_DEU_POL_1927-12-16_JUD_01_ME
    _01_BI.pdf"
## [30] "PCIJ_A_15_MinoritySchoolsUpperSilesia_DEU_POL_1928-04-26_JUD_01_ME_01_
    BI.pdf"
## [31] "PCIJ_A_15_MinoritySchoolsUpperSilesia_DEU_POL_1928-04-26_JUD_01_ME_02_
    BI.pdf"
## [32] "PCIJ_A_15_MinoritySchoolsUpperSilesia_DEU_POL_1928-04-26_JUD_01_ME_03_
    BI.pdf"
## [33] "PCIJ_A_15_MinoritySchoolsUpperSilesia_DEU_POL_1928-04-26_JUD_01_ME_04_
    BI.pdf"
## [34] "PCIJ_A_15_MinoritySchoolsUpperSilesia_DEU_POL_1928-04-26_ANX_01_NA_NA_
    BI.pdf"
## [35] "PCIJ_A_15_MinoritySchoolsUpperSilesia_DEU_POL_1928-04-26_ANX_02_NA_NA_
    BI.pdf"
## [36] "PCIJ_A_17_ChorzowFactory-Indemnities_DEU_POL_1928-09-13_JUD_01_ME_01_BI
    .pdf"
## [37] "PCIJ_A_17_ChorzowFactory-Indemnities_DEU_POL_1928-09-13_JUD_01_ME_02_BI
    .pdf"
## [38] "PCIJ_A_17_ChorzowFactory-Indemnities_DEU_POL_1928-09-13_JUD_01_ME_03_BI
    .pdf"
## [39] "PCIJ_A_17_ChorzowFactory-Indemnities_DEU_POL_1928-09-13_JUD_01_ME_04_BI
    .pdf"
## [40] "PCIJ_A_17_ChorzowFactory-Indemnities_DEU_POL_1928-09-13_ORD_01_EX_00_BI
    .pdf"
## [41] "PCIJ_A_17_ChorzowFactory-Indemnities_DEU_POL_1928-09-13_ANX_01_NA_NA_BI
    .pdf"
## [42] "PCIJ_A_19_ChorzowFactory-Indemnities_DEU_POL_1929-05-25_ORD_01_EX_00_BI
    .pdf"
## [43] "PCIJ_A_19_ChorzowFactory-Indemnities_DEU_POL_1928-12-15_ORD_01_DI_00_BI
    .pdf"
## [44] "PCIJ_A_20_SerbianLoans_FRA_YUG_1929-07-12_JUD_01_ME_01_BI.pdf"
## [45] "PCIJ_A_20_SerbianLoans_FRA_YUG_1929-07-12_JUD_01_ME_02_BI.pdf"
## [46] "PCIJ_A_20_SerbianLoans_FRA_YUG_1929-07-12_JUD_01_ME_03_BI.pdf"
## [47] "PCIJ_A_20_SerbianLoans_FRA_YUG_1929-07-12_ANX_01_NA_NA_BI.pdf"
## [48] "PCIJ_A_21_BrazilianLoans_FRA_BRA_1929-07-12_JUD_01_ME_00_BI.pdf"
## [49] "PCIJ_A_21_BrazilianLoans_FRA_BRA_1929-07-12_JUD_01_ME_01_BI.pdf"
## [50] "PCIJ_A_21_BrazilianLoans_FRA_BRA_1929-07-12_JUD_01_ME_02_BI.pdf"
## [51] "PCIJ_A_21_BrazilianLoans_FRA_BRA_1929-07-12_ANX_01_NA_NA_BI.pdf"
## [52] "PCIJ_A_22_FreeZonesUpperSavoyGex_FRA_CHE_1929-08-19_ORD_01_EV-SE_01_BI.
    pdf"
## [53] "PCIJ_A_22_FreeZonesUpperSavoyGex_FRA_CHE_1929-08-19_ORD_01_EV-SE_02_BI.
    pdf"
## [54] "PCIJ_A_22_FreeZonesUpperSavoyGex_FRA_CHE_1929-08-19_ORD_01_EV-SE_03_BI.
    pdf"
## [55] "PCIJ_A_22_FreeZonesUpperSavoyGex_FRA_CHE_1929-08-19_ORD_01_EV-SE_04_BI.
    pdf"
## [56] "PCIJ_A_22_FreeZonesUpperSavoyGex_FRA_CHE_1929-08-19_ORD_01_EV-SE_05_BI.
    pdf"
## [57] "PCIJ_A_23_OderCommission_GBR-CSK-DNK-FRA-DEU_POL_1929-08-15_ORD_01_EV
    _00_BI.pdf"
## [58] "PCIJ_A_23_OderCommission_GBR-CSK-DNK-FRA-DEU_POL_1929-08-20_ORD_01_EV
    _00_BI.pdf"
## [59] "PCIJ_A_23_OderCommission_GBR-CSK-DNK-FRA-DEU_POL_1929-08-15_ORD_02_TL
    _00_BI.pdf"
## [60] "PCIJ_A_24_FreeZonesUpperSavoyGex-SecondPhase_FRA_CHE_1930-12-06_ORD_01_
    SE_01_BI.pdf"
```

```
## [61] "PCIJ_A_24_FreeZonesUpperSavoyGex-SecondPhase_FRA_CHE_1930-12-06_ORD_01_
    SE_02_BI.pdf"
## [62] "PCIJ_AB_40_GermanMinoritySchools_DEU_POL_1931-05-15_ADV_01_NA_01_BI.pdf
    "
## [63] "PCIJ_AB_40_GermanMinoritySchools_DEU_POL_1931-05-15_ANX_01_NA_NA_BI.pdf
    "
## [64] "PCIJ_AB_40_GermanMinoritySchools_DEU_POL_1931-05-15_ANX_02_NA_NA_BI.pdf
    "
## [65] "PCIJ_AB_41_CustomsRegime_LNC_NA_1931-09-05_ADV_01_NA_01_BI.pdf"
## [66] "PCIJ_AB_41_CustomsRegime_LNC_NA_1931-09-05_ADV_01_NA_02_BI.pdf"
## [67] "PCIJ_AB_41_CustomsRegime_LNC_NA_1931-09-05_ANX_02_NA_NA_BI.pdf"
## [68] "PCIJ_AB_41_CustomsRegime_LNC_NA_1931-09-05_ANX_03_NA_NA_BI.pdf"
## [69] "PCIJ_AB_41_CustomsRegime_LNC_NA_1931-07-20_ORD_01_AJ_00_BI.pdf"
## [70] "PCIJ_AB_41_CustomsRegime_LNC_NA_1931-07-20_ORD_01_AJ_01_BI.pdf"
## [71] "PCIJ_AB_43_PolishWarVessels_LNC_NA_1931-12-11_ADV_01_NA_01_BI.pdf"
## [72] "PCIJ_AB_43_PolishWarVessels_LNC_NA_1931-12-11_ANX_01_NA_NA_BI.pdf"
## [73] "PCIJ_AB_44_TreatmentPolishNationals_LNC_NA_1932-02-04_ADV_01_NA_01_BI.
    pdf"
## [74] "PCIJ_AB_44_TreatmentPolishNationals_LNC_NA_1932-02-04_ADV_01_NA_02_BI.
    pdf"
## [75] "PCIJ_AB_44_TreatmentPolishNationals_LNC_NA_1932-02-04_ANX_01_NA_NA_BI.
    pdf"
## [76] "PCIJ_AB_45_GrecoBulgarianAgreement_LNC_NA_1932-03-08_ANX_01_NA_NA_BI.
    pdf"
## [77] "PCIJ_AB_46_FreeZonesUpperSavoyGex_FRA_CHE_1932-06-07_JUD_01_ME_01_BI.
    pdf"
## [78] "PCIJ_AB_46_FreeZonesUpperSavoyGex_FRA_CHE_1932-06-07_JUD_01_ME_02_BI.
    pdf"
## [79] "PCIJ_AB_46_FreeZonesUpperSavoyGex_FRA_CHE_1932-06-07_JUD_01_ME_03_BI.
    pdf"
## [80] "PCIJ_AB_46_FreeZonesUpperSavoyGex_FRA_CHE_1931-08-06_ORD_01_TL-DH_00_BI
    .pdf"
## [81] "PCIJ_AB_46_FreeZonesUpperSavoyGex_FRA_CHE_1932-06-07_ANX_01_NA_NA_BI.
    pdf"
## [82] "PCIJ_AB_47_InterpretationMemelStatute_GBR-FRA-ITA-JPN_LTU_1932-06-24_
    JUD_01_PO_00_BI.pdf"
## [83] "PCIJ_AB_47_InterpretationMemelStatute_GBR-FRA-ITA-JPN_LTU_1932-06-24_
    JUD_01_PO_01_BI.pdf"
## [84] "PCIJ_AB_47_InterpretationMemelStatute_GBR-FRA-ITA-JPN_LTU_1932-06-24_
    ORD_01_TL_00_BI.pdf"
## [85] "PCIJ_AB_49_InterpretationMemelStatute_GBR-FRA-ITA-JPN_LTU_1932-08-11_
    JUD_01_ME_01_BI.pdf"
## [86] "PCIJ_AB_49_InterpretationMemelStatute_GBR-FRA-ITA-JPN_LTU_1932-08-11_
    JUD_01_ME_02_BI.pdf"
## [87] "PCIJ_AB_49_InterpretationMemelStatute_GBR-FRA-ITA-JPN_LTU_1932-08-11_
    ANX_01_NA_NA_BI.pdf"
## [88] "PCIJ_AB_50_WomenNightWork_LNC_NA_1932-11-15_ADV_01_NA_01_BI.pdf"
## [89] "PCIJ_AB_50_WomenNightWork_LNC_NA_1932-11-15_ANX_01_NA_NA_BI.pdf"
## [90] "PCIJ_AB_53_EasternGreenland_DNK_NOR_1933-04-05_JUD_01_ME_01_BI.pdf"
## [91] "PCIJ_AB_53_EasternGreenland_DNK_NOR_1933-04-05_JUD_01_ME_02_BI.pdf"
## [92] "PCIJ_AB_53_EasternGreenland_DNK_NOR_1933-04-05_JUD_01_ME_03_BI.pdf"
## [93] "PCIJ_AB_58_PolishAgrarianReform_DEU_POL_1933-07-29_ORD_01_IM_01_BI.pdf"
## [94] "PCIJ_AB_58_PolishAgrarianReform_DEU_POL_1933-07-29_ORD_01_IM_02_BI.pdf"
## [95] "PCIJ_AB_58_PolishAgrarianReform_DEU_POL_1933-07-29_ORD_01_IM_03_BI.pdf"
## [96] "PCIJ_AB_58_PolishAgrarianReform_DEU_POL_1933-07-29_ANX_01_NA_NA_BI.pdf"
## [97] "PCIJ_AB_61_Pazmany_CSK_HUN_1933-12-15_JUD_01_ME_01_BI.pdf"
```

```
##  [98] "PCIJ_AB_61_Pazmany_CSK_HUN_1933-12-15_ANX_01_NA_NA_BI.pdf"
##  [99] "PCIJ_AB_62_Lighthouses_FRA_GRC_1934-03-17_JUD_01_ME_01_BI.pdf"
## [100] "PCIJ_AB_62_Lighthouses_FRA_GRC_1934-03-17_JUD_01_ME_02_BI.pdf"
## [101] "PCIJ_AB_62_Lighthouses_FRA_GRC_1934-03-17_ANX_01_NA_NA_BI.pdf"
## [102] "PCIJ_AB_63_OscarChinn_GBR_BEL_1934-12-12_JUD_01_ME_01_BI.pdf"
## [103] "PCIJ_AB_63_OscarChinn_GBR_BEL_1934-12-12_JUD_01_ME_02_BI.pdf"
## [104] "PCIJ_AB_63_OscarChinn_GBR_BEL_1934-12-12_JUD_01_ME_03_BI.pdf"
## [105] "PCIJ_AB_63_OscarChinn_GBR_BEL_1934-12-12_JUD_01_ME_04_BI.pdf"
## [106] "PCIJ_AB_63_OscarChinn_GBR_BEL_1934-12-12_JUD_01_ME_05_BI.pdf"
## [107] "PCIJ_AB_63_OscarChinn_GBR_BEL_1934-12-12_ANX_01_NA_NA_BI.pdf"
## [108] "PCIJ_AB_64_MinoritySchoolsAlbania_LNC_NA_1935-04-06_ADV_01_NA_01_BI.pdf
    "
## [109] "PCIJ_AB_64_MinoritySchoolsAlbania_LNC_NA_1935-04-06_ANX_01_NA_NA_BI.pdf
    "
## [110] "PCIJ_AB_65_DanzigLegislativeDecrees_LNC_NA_1935-12-04_ADV_01_NA_01_BI.
    pdf"
## [111] "PCIJ_AB_65_DanzigLegislativeDecrees_LNC_NA_1935-12-04_ADV_01_NA_02_BI.
    pdf"
## [112] "PCIJ_AB_65_DanzigLegislativeDecrees_LNC_NA_1935-12-04_ADV_01_NA_03_BI.
    pdf"
## [113] "PCIJ_AB_65_DanzigLegislativeDecrees_LNC_NA_1935-10-31_ORD_01_AJ_00_BI.
    pdf"
## [114] "PCIJ_AB_65_DanzigLegislativeDecrees_LNC_NA_1935-12-04_ANX_01_NA_NA_BI.
    pdf"
## [115] "PCIJ_AB_68_PajzsCsakyEsterhazy_HUN_YUG_1936-12-16_JUD_01_ME_01_BI.pdf"
## [116] "PCIJ_AB_68_PajzsCsakyEsterhazy_HUN_YUG_1936-12-16_JUD_01_ME_02_BI.pdf"
## [117] "PCIJ_AB_68_PajzsCsakyEsterhazy_HUN_YUG_1936-12-16_JUD_01_ME_03_BI.pdf"
## [118] "PCIJ_AB_68_PajzsCsakyEsterhazy_HUN_YUG_1936-12-16_JUD_01_ME_04_BI.pdf"
## [119] "PCIJ_AB_68_PajzsCsakyEsterhazy_HUN_YUG_1936-12-16_JUD_01_ME_05_BI.pdf"
## [120] "PCIJ_AB_68_PajzsCsakyEsterhazy_HUN_YUG_1936-12-16_ANX_01_NA_NA_BI.pdf"
## [121] "PCIJ_AB_70_Meuse_NLD_BEL_1937-06-28_JUD_01_ME_01_BI.pdf"
## [122] "PCIJ_AB_70_Meuse_NLD_BEL_1937-06-28_JUD_01_ME_02_BI.pdf"
## [123] "PCIJ_AB_70_Meuse_NLD_BEL_1937-06-28_JUD_01_ME_03_BI.pdf"
## [124] "PCIJ_AB_70_Meuse_NLD_BEL_1937-06-28_JUD_01_ME_04_BI.pdf"
## [125] "PCIJ_AB_70_Meuse_NLD_BEL_1937-06-28_JUD_01_ME_05_BI.pdf"
## [126] "PCIJ_AB_70_Meuse_NLD_BEL_1937-06-28_ANX_02_NA_NA_BI.pdf"
## [127] "PCIJ_AB_71_LighthousesCreteSamos_FRA_GRC_1937-10-08_JUD_01_ME_01_BI.pdf
    "
## [128] "PCIJ_AB_71_LighthousesCreteSamos_FRA_GRC_1937-10-08_JUD_01_ME_02_BI.pdf
    "
## [129] "PCIJ_AB_71_LighthousesCreteSamos_FRA_GRC_1937-10-08_JUD_01_ME_03_BI.pdf
    "
## [130] "PCIJ_AB_71_LighthousesCreteSamos_FRA_GRC_1937-10-08_JUD_01_ME_04_BI.pdf
    "
## [131] "PCIJ_AB_71_LighthousesCreteSamos_FRA_GRC_1937-10-08_ANX_01_NA_NA_BI.pdf
    "
## [132] "PCIJ_AB_72_Borchgrave_BEL_ESP_1937-11-06_ORD_01_TL_00_BI.pdf"
## [133] "PCIJ_AB_72_Borchgrave_BEL_ESP_1937-11-06_ANX_01_NA_NA_BI.pdf"
## [134] "PCIJ_AB_74_PhosphatesMarocco_ITA_FRA_1938-06-14_JUD_01_PO_01_BI.pdf"
## [135] "PCIJ_AB_74_PhosphatesMarocco_ITA_FRA_1938-06-14_JUD_01_PO_02_BI.pdf"
## [136] "PCIJ_AB_74_PhosphatesMarocco_ITA_FRA_1938-06-14_ANX_01_NA_NA_BI.pdf"
## [137] "PCIJ_AB_76_PanevezysSaldutiskisRailway_EST_LTU_1939-02-28_JUD_01_ME_01_
    BI.pdf"
## [138] "PCIJ_AB_76_PanevezysSaldutiskisRailway_EST_LTU_1939-02-28_JUD_01_ME_02_
    BI.pdf"
## [139] "PCIJ_AB_76_PanevezysSaldutiskisRailway_EST_LTU_1939-02-28_JUD_01_ME_03_
```

```
    BI.pdf"
## [140] "PCIJ_AB_76_PanevezysSaldutiskisRailway_EST_LTU_1939-02-28_JUD_01_ME_04_
    BI.pdf"
## [141] "PCIJ_AB_76_PanevezysSaldutiskisRailway_EST_LTU_1939-02-28_ANX_01_NA_NA_
    BI.pdf"
## [142] "PCIJ_AB_77_ElectricityCompanySofiaBulgaria_BEL_BGR_1939-04-04_JUD_01_PO
    _01_BI.pdf"
## [143] "PCIJ_AB_77_ElectricityCompanySofiaBulgaria_BEL_BGR_1939-04-04_JUD_01_PO
    _02_BI.pdf"
## [144] "PCIJ_AB_77_ElectricityCompanySofiaBulgaria_BEL_BGR_1939-04-04_JUD_01_PO
    _03_BI.pdf"
## [145] "PCIJ_AB_77_ElectricityCompanySofiaBulgaria_BEL_BGR_1939-04-04_JUD_01_PO
    _04_BI.pdf"
## [146] "PCIJ_AB_77_ElectricityCompanySofiaBulgaria_BEL_BGR_1939-04-04_JUD_01_PO
    _05_BI.pdf"
## [147] "PCIJ_AB_77_ElectricityCompanySofiaBulgaria_BEL_BGR_1939-04-04_JUD_01_PO
    _06_BI.pdf"
## [148] "PCIJ_AB_77_ElectricityCompanySofiaBulgaria_BEL_BGR_1939-04-04_JUD_01_PO
    _07_BI.pdf"
## [149] "PCIJ_AB_77_ElectricityCompanySofiaBulgaria_BEL_BGR_1939-04-04_ORD_01_TL
    _00_BI.pdf"
## [150] "PCIJ_AB_77_ElectricityCompanySofiaBulgaria_BEL_BGR_1939-04-04_ANX_01_NA
    _NA_BI.pdf"
## [151] "PCIJ_AB_78_SocieteCommercialeBelgique_BEL_GRC_1939-06-15_JUD_01_ME_01_
    BI.pdf"
## [152] "PCIJ_AB_78_SocieteCommercialeBelgique_BEL_GRC_1939-06-15_JUD_01_ME_02_
    BI.pdf"
## [153] "PCIJ_AB_78_SocieteCommercialeBelgique_BEL_GRC_1939-06-15_ANX_01_NA_NA_
    BI.pdf"
## [154] "PCIJ_B_01_WorkersDelegateILO_LNC_NA_1922-05-22_APP_01_NA_NA_BI.pdf"
## [155] "PCIJ_B_01_WorkersDelegateILO_LNC_NA_1922-07-31_ADV_01_NA_00_BI.pdf"
## [156] "PCIJ_B_02_ILOCompetencePersonsAgriculture_LNC_NA_1922-08-12_ADV_01_NA
    _00_BI.pdf"
## [157] "PCIJ_B_02_ILOCompetencePersonsAgriculture_LNC_NA_1922-05-22_APP_01_NA_
    NA_BI.pdf"
## [158] "PCIJ_B_03_ILOCompetenceMethodsAgriculture_LNC_NA_1922-08-12_ADV_01_NA
    _00_BI.pdf"
## [159] "PCIJ_B_03_ILOCompetenceMethodsAgriculture_LNC_NA_1922-07-18_APP_01_NA_
    NA_BI.pdf"
## [160] "PCIJ_B_04_NationalityDecrees_LNC_NA_1923-02-07_ADV_01_NA_00_BI.pdf"
## [161] "PCIJ_B_04_NationalityDecrees_LNC_NA_1922-11-06_APP_01_NA_NA_BI.pdf"
## [162] "PCIJ_B_05_EasternCarelia_LNC_NA_1923-04-27_APP_01_NA_NA_BI.pdf"
## [163] "PCIJ_B_05_EasternCarelia_LNC_NA_1923-07-23_ADV_01_NA_00_BI.pdf"
## [164] "PCIJ_B_07_AcquisitionPolishNationality_LNC_NA_1923-09-15_ADV_01_NA_01_
    BI.pdf"
## [165] "PCIJ_B_09_MonasterySaintNaoum_LNC_NA_1924-09-04_ADV_01_NA_00_BI.pdf"
## [166] "PCIJ_B_10_ExchangeGreekTurkishPopulations_LNC_NA_1925-02-21_ANX_01_NA_
    NA_BI.pdf"
## [167] "PCIJ_B_11_PostalServiceDanzig_LNC_NA_1925-05-16_ANX_01_NA_NA_BI.pdf"
## [168] "PCIJ_B_12_TreatyLausanne_LNC_NA_1925-11-21_ANX_01_NA_NA_BI.pdf"
## [169] "PCIJ_B_13_ILOCompetenceEmployer_LNC_NA_1926-07-23_ANX_01_NA_NA_BI.pdf"
## [170] "PCIJ_B_14_DanubeCommission_LNC_NA_1927-12-08_ADV_01_NA_01_BI.pdf"
## [171] "PCIJ_B_14_DanubeCommission_LNC_NA_1927-12-08_ADV_01_NA_02_BI.pdf"
## [172] "PCIJ_B_14_DanubeCommission_LNC_NA_1927-12-08_ADV_01_NA_03_BI.pdf"
## [173] "PCIJ_B_14_DanubeCommission_LNC_NA_1927-12-08_ANX_01_NA_NA_BI.pdf"
## [174] "PCIJ_B_15_DanzigCourts_LNC_NA_1928-03-03_ANX_01_NA_NA_BI.pdf"
```

```
## [175] "PCIJ_B_15_DanzigCourts_LNC_NA_1928-03-03_ANX_03_NA_NA_BI.pdf"
## [176] "PCIJ_B_16_GrecoTurkishAgreement_LNC_NA_1928-08-28_ANX_01_NA_NA_BI.pdf"
## [177] "PCIJ_B_17_GrecoBulgarianCommunities_LNC_NA_1930-07-31_ANX_01_NA_NA_BI.
    pdf"
## [178] "PCIJ_B_17_GrecoBulgarianCommunities_LNC_NA_1930-07-31_ANX_02_NA_NA_BI.
    pdf"
## [179] "PCIJ_B_18_DanzigILO_LNC_NA_1930-08-26_ADV_01_NA_01_BI.pdf"
## [180] "PCIJ_B_18_DanzigILO_LNC_NA_1930-08-26_ADV_01_NA_02_BI.pdf"
```

10.7.3 Execute Split

```
out <- foreach(file = even.english,
               .errorhandling = 'pass',
               .combine = c) %dopar% {

    out1 <- vector(mode = "list",
                   length = 2)

    temp1 <- seq(1, pdf_length(file), 1)

    even <- temp1[lapply(seq(1, max(temp1), 1), "%%", 2) == 0]
    even.name <- gsub("BI\\.pdf",
                      "EN\\.pdf",
                      file)
    out1[[1]] <- pdf_subset(file,
                            pages = even,
                            output = even.name)


    odd <- temp1[lapply(seq(1, max(temp1), 1), "%%", 2) != 0]
    odd.name <- gsub("BI\\.pdf",
                     "FR\\.pdf",
                     file)
    out1[[2]] <- pdf_subset(file,
                            pages = odd,
                            output = odd.name)

    return(out1)
}
```

10.7.4 Print Split Results

```
print(unlist(out))
```

```
##   [1] "PCIJ_A_01_Wimbledon_GBR-FRA-ITA-JPN_DEU_1923-05-22_APP_01_NA_NA_EN.pdf"
##   [2] "PCIJ_A_01_Wimbledon_GBR-FRA-ITA-JPN_DEU_1923-05-22_APP_01_NA_NA_FR.pdf"
##   [3] "PCIJ_A_01_Wimbledon_GBR-FRA-ITA-JPN_DEU_1923-06-28_JUD_01_IN_00_EN.pdf"
##   [4] "PCIJ_A_01_Wimbledon_GBR-FRA-ITA-JPN_DEU_1923-06-28_JUD_01_IN_00_FR.pdf"
##   [5] "PCIJ_A_01_Wimbledon_GBR-FRA-ITA-JPN_DEU_1923-08-17_JUD_01_ME_00_EN.pdf"
```

```
##   [6] "PCIJ_A_01_Wimbledon_GBR-FRA-ITA-JPN_DEU_1923-08-17_JUD_01_ME_00_FR.pdf"
##   [7] "PCIJ_A_01_Wimbledon_GBR-FRA-ITA-JPN_DEU_1923-08-17_JUD_01_ME_01_EN.pdf"
##   [8] "PCIJ_A_01_Wimbledon_GBR-FRA-ITA-JPN_DEU_1923-08-17_JUD_01_ME_01_FR.pdf"
##   [9] "PCIJ_A_01_Wimbledon_GBR-FRA-ITA-JPN_DEU_1923-08-17_JUD_01_ME_02_EN.pdf"
##  [10] "PCIJ_A_01_Wimbledon_GBR-FRA-ITA-JPN_DEU_1923-08-17_JUD_01_ME_02_FR.pdf"
##  [11] "PCIJ_A_02_MavrommatisPalestine_GRC_GBR_1924-08-30_JUD_01_PO_01_EN.pdf"
##  [12] "PCIJ_A_02_MavrommatisPalestine_GRC_GBR_1924-08-30_JUD_01_PO_01_FR.pdf"
##  [13] "PCIJ_A_02_MavrommatisPalestine_GRC_GBR_1924-08-30_JUD_01_PO_02_EN.pdf"
##  [14] "PCIJ_A_02_MavrommatisPalestine_GRC_GBR_1924-08-30_JUD_01_PO_02_FR.pdf"
##  [15] "PCIJ_A_02_MavrommatisPalestine_GRC_GBR_1924-08-30_JUD_01_PO_03_EN.pdf"
##  [16] "PCIJ_A_02_MavrommatisPalestine_GRC_GBR_1924-08-30_JUD_01_PO_03_FR.pdf"
##  [17] "PCIJ_A_02_MavrommatisPalestine_GRC_GBR_1924-08-30_JUD_01_PO_04_EN.pdf"
##  [18] "PCIJ_A_02_MavrommatisPalestine_GRC_GBR_1924-08-30_JUD_01_PO_04_FR.pdf"
##  [19] "PCIJ_A_02_MavrommatisPalestine_GRC_GBR_1924-08-30_JUD_01_PO_05_EN.pdf"
##  [20] "PCIJ_A_02_MavrommatisPalestine_GRC_GBR_1924-08-30_JUD_01_PO_05_FR.pdf"
##  [21] "PCIJ_A_06_GermanInterestsUpperSilesia_DEU_POL_1925-08-25_JUD_01_PO_01_
    EN.pdf"
##  [22] "PCIJ_A_06_GermanInterestsUpperSilesia_DEU_POL_1925-08-25_JUD_01_PO_01_
    FR.pdf"
##  [23] "PCIJ_A_06_GermanInterestsUpperSilesia_DEU_POL_1925-08-25_JUD_01_PO_02_
    EN.pdf"
##  [24] "PCIJ_A_06_GermanInterestsUpperSilesia_DEU_POL_1925-08-25_JUD_01_PO_02_
    FR.pdf"
##  [25] "PCIJ_A_07_GermanInterestsUpperSilesia_DEU_POL_1926-05-25_JUD_01_ME_01_
    EN.pdf"
##  [26] "PCIJ_A_07_GermanInterestsUpperSilesia_DEU_POL_1926-05-25_JUD_01_ME_01_
    FR.pdf"
##  [27] "PCIJ_A_07_GermanInterestsUpperSilesia_DEU_POL_1926-05-25_JUD_01_ME_02_
    EN.pdf"
##  [28] "PCIJ_A_07_GermanInterestsUpperSilesia_DEU_POL_1926-05-25_JUD_01_ME_02_
    FR.pdf"
##  [29] "PCIJ_A_08_SinoBelgianTreaty_BEL_CHN_1927-01-08_ORD_01_IM_00_EN.pdf"
##  [30] "PCIJ_A_08_SinoBelgianTreaty_BEL_CHN_1927-01-08_ORD_01_IM_00_FR.pdf"
##  [31] "PCIJ_A_08_SinoBelgianTreaty_BEL_CHN_1927-02-15_ORD_01_IM_00_EN.pdf"
##  [32] "PCIJ_A_08_SinoBelgianTreaty_BEL_CHN_1927-02-15_ORD_01_IM_00_FR.pdf"
##  [33] "PCIJ_A_08_SinoBelgianTreaty_BEL_CHN_1927-06-18_ORD_01_TL_00_EN.pdf"
##  [34] "PCIJ_A_08_SinoBelgianTreaty_BEL_CHN_1927-06-18_ORD_01_TL_00_FR.pdf"
##  [35] "PCIJ_A_09_ChorzowFactory_DEU_POL_1927-07-26_JUD_01_PO_01_EN.pdf"
##  [36] "PCIJ_A_09_ChorzowFactory_DEU_POL_1927-07-26_JUD_01_PO_01_FR.pdf"
##  [37] "PCIJ_A_10_Lotus_FRA_TUR_1927-09-07_JUD_01_ME_01_EN.pdf"
##  [38] "PCIJ_A_10_Lotus_FRA_TUR_1927-09-07_JUD_01_ME_01_FR.pdf"
##  [39] "PCIJ_A_10_Lotus_FRA_TUR_1927-09-07_JUD_01_ME_02_EN.pdf"
##  [40] "PCIJ_A_10_Lotus_FRA_TUR_1927-09-07_JUD_01_ME_02_FR.pdf"
##  [41] "PCIJ_A_10_Lotus_FRA_TUR_1927-09-07_JUD_01_ME_03_EN.pdf"
##  [42] "PCIJ_A_10_Lotus_FRA_TUR_1927-09-07_JUD_01_ME_03_FR.pdf"
##  [43] "PCIJ_A_10_Lotus_FRA_TUR_1927-09-07_JUD_01_ME_04_EN.pdf"
##  [44] "PCIJ_A_10_Lotus_FRA_TUR_1927-09-07_JUD_01_ME_04_FR.pdf"
##  [45] "PCIJ_A_10_Lotus_FRA_TUR_1927-09-07_JUD_01_ME_05_EN.pdf"
##  [46] "PCIJ_A_10_Lotus_FRA_TUR_1927-09-07_JUD_01_ME_05_FR.pdf"
##  [47] "PCIJ_A_10_Lotus_FRA_TUR_1927-09-07_JUD_01_ME_06_EN.pdf"
##  [48] "PCIJ_A_10_Lotus_FRA_TUR_1927-09-07_JUD_01_ME_06_FR.pdf"
##  [49] "PCIJ_A_10_Lotus_FRA_TUR_1927-09-07_ANX_01_NA_NA_EN.pdf"
##  [50] "PCIJ_A_10_Lotus_FRA_TUR_1927-09-07_ANX_01_NA_NA_FR.pdf"
##  [51] "PCIJ_A_11_MavrommatisJerusalem-Readaptation_GRC_GBR_1927-10-10_JUD_01_
    PO_01_EN.pdf"
##  [52] "PCIJ_A_11_MavrommatisJerusalem-Readaptation_GRC_GBR_1927-10-10_JUD_01_
```

```
    PO_01_FR.pdf"
##  [53] "PCIJ_A_11_MavrommatisJerusalem-Readaptation_GRC_GBR_1927-10-10_JUD_01_
    PO_02_EN.pdf"
##  [54] "PCIJ_A_11_MavrommatisJerusalem-Readaptation_GRC_GBR_1927-10-10_JUD_01_
    PO_02_FR.pdf"
##  [55] "PCIJ_A_11_MavrommatisJerusalem-Readaptation_GRC_GBR_1927-10-10_JUD_01_
    PO_03_EN.pdf"
##  [56] "PCIJ_A_11_MavrommatisJerusalem-Readaptation_GRC_GBR_1927-10-10_JUD_01_
    PO_03_FR.pdf"
##  [57] "PCIJ_A_13_ChorzowFactory-Interpretation_DEU_POL_1927-12-16_JUD_01_ME
    _01_EN.pdf"
##  [58] "PCIJ_A_13_ChorzowFactory-Interpretation_DEU_POL_1927-12-16_JUD_01_ME
    _01_FR.pdf"
##  [59] "PCIJ_A_15_MinoritySchoolsUpperSilesia_DEU_POL_1928-04-26_JUD_01_ME_01_
    EN.pdf"
##  [60] "PCIJ_A_15_MinoritySchoolsUpperSilesia_DEU_POL_1928-04-26_JUD_01_ME_01_
    FR.pdf"
##  [61] "PCIJ_A_15_MinoritySchoolsUpperSilesia_DEU_POL_1928-04-26_JUD_01_ME_02_
    EN.pdf"
##  [62] "PCIJ_A_15_MinoritySchoolsUpperSilesia_DEU_POL_1928-04-26_JUD_01_ME_02_
    FR.pdf"
##  [63] "PCIJ_A_15_MinoritySchoolsUpperSilesia_DEU_POL_1928-04-26_JUD_01_ME_03_
    EN.pdf"
##  [64] "PCIJ_A_15_MinoritySchoolsUpperSilesia_DEU_POL_1928-04-26_JUD_01_ME_03_
    FR.pdf"
##  [65] "PCIJ_A_15_MinoritySchoolsUpperSilesia_DEU_POL_1928-04-26_JUD_01_ME_04_
    EN.pdf"
##  [66] "PCIJ_A_15_MinoritySchoolsUpperSilesia_DEU_POL_1928-04-26_JUD_01_ME_04_
    FR.pdf"
##  [67] "PCIJ_A_15_MinoritySchoolsUpperSilesia_DEU_POL_1928-04-26_ANX_01_NA_NA_
    EN.pdf"
##  [68] "PCIJ_A_15_MinoritySchoolsUpperSilesia_DEU_POL_1928-04-26_ANX_01_NA_NA_
    FR.pdf"
##  [69] "PCIJ_A_15_MinoritySchoolsUpperSilesia_DEU_POL_1928-04-26_ANX_02_NA_NA_
    EN.pdf"
##  [70] "PCIJ_A_15_MinoritySchoolsUpperSilesia_DEU_POL_1928-04-26_ANX_02_NA_NA_
    FR.pdf"
##  [71] "PCIJ_A_17_ChorzowFactory-Indemnities_DEU_POL_1928-09-13_JUD_01_ME_01_EN
    .pdf"
##  [72] "PCIJ_A_17_ChorzowFactory-Indemnities_DEU_POL_1928-09-13_JUD_01_ME_01_FR
    .pdf"
##  [73] "PCIJ_A_17_ChorzowFactory-Indemnities_DEU_POL_1928-09-13_JUD_01_ME_02_EN
    .pdf"
##  [74] "PCIJ_A_17_ChorzowFactory-Indemnities_DEU_POL_1928-09-13_JUD_01_ME_02_FR
    .pdf"
##  [75] "PCIJ_A_17_ChorzowFactory-Indemnities_DEU_POL_1928-09-13_JUD_01_ME_03_EN
    .pdf"
##  [76] "PCIJ_A_17_ChorzowFactory-Indemnities_DEU_POL_1928-09-13_JUD_01_ME_03_FR
    .pdf"
##  [77] "PCIJ_A_17_ChorzowFactory-Indemnities_DEU_POL_1928-09-13_JUD_01_ME_04_EN
    .pdf"
##  [78] "PCIJ_A_17_ChorzowFactory-Indemnities_DEU_POL_1928-09-13_JUD_01_ME_04_FR
    .pdf"
##  [79] "PCIJ_A_17_ChorzowFactory-Indemnities_DEU_POL_1928-09-13_ORD_01_EX_00_EN
    .pdf"
##  [80] "PCIJ_A_17_ChorzowFactory-Indemnities_DEU_POL_1928-09-13_ORD_01_EX_00_FR
```

```
    .pdf"
##  [81] "PCIJ_A_17_ChorzowFactory-Indemnities_DEU_POL_1928-09-13_ANX_01_NA_NA_EN
    .pdf"
##  [82] "PCIJ_A_17_ChorzowFactory-Indemnities_DEU_POL_1928-09-13_ANX_01_NA_NA_FR
    .pdf"
##  [83] "PCIJ_A_19_ChorzowFactory-Indemnities_DEU_POL_1929-05-25_ORD_01_EX_00_EN
    .pdf"
##  [84] "PCIJ_A_19_ChorzowFactory-Indemnities_DEU_POL_1929-05-25_ORD_01_EX_00_FR
    .pdf"
##  [85] "PCIJ_A_19_ChorzowFactory-Indemnities_DEU_POL_1928-12-15_ORD_01_DI_00_EN
    .pdf"
##  [86] "PCIJ_A_19_ChorzowFactory-Indemnities_DEU_POL_1928-12-15_ORD_01_DI_00_FR
    .pdf"
##  [87] "PCIJ_A_20_SerbianLoans_FRA_YUG_1929-07-12_JUD_01_ME_01_EN.pdf"
##  [88] "PCIJ_A_20_SerbianLoans_FRA_YUG_1929-07-12_JUD_01_ME_01_FR.pdf"
##  [89] "PCIJ_A_20_SerbianLoans_FRA_YUG_1929-07-12_JUD_01_ME_02_EN.pdf"
##  [90] "PCIJ_A_20_SerbianLoans_FRA_YUG_1929-07-12_JUD_01_ME_02_FR.pdf"
##  [91] "PCIJ_A_20_SerbianLoans_FRA_YUG_1929-07-12_JUD_01_ME_03_EN.pdf"
##  [92] "PCIJ_A_20_SerbianLoans_FRA_YUG_1929-07-12_JUD_01_ME_03_FR.pdf"
##  [93] "PCIJ_A_20_SerbianLoans_FRA_YUG_1929-07-12_ANX_01_NA_NA_EN.pdf"
##  [94] "PCIJ_A_20_SerbianLoans_FRA_YUG_1929-07-12_ANX_01_NA_NA_FR.pdf"
##  [95] "PCIJ_A_21_BrazilianLoans_FRA_BRA_1929-07-12_JUD_01_ME_00_EN.pdf"
##  [96] "PCIJ_A_21_BrazilianLoans_FRA_BRA_1929-07-12_JUD_01_ME_00_FR.pdf"
##  [97] "PCIJ_A_21_BrazilianLoans_FRA_BRA_1929-07-12_JUD_01_ME_01_EN.pdf"
##  [98] "PCIJ_A_21_BrazilianLoans_FRA_BRA_1929-07-12_JUD_01_ME_01_FR.pdf"
##  [99] "PCIJ_A_21_BrazilianLoans_FRA_BRA_1929-07-12_JUD_01_ME_02_EN.pdf"
## [100] "PCIJ_A_21_BrazilianLoans_FRA_BRA_1929-07-12_JUD_01_ME_02_FR.pdf"
## [101] "PCIJ_A_21_BrazilianLoans_FRA_BRA_1929-07-12_ANX_01_NA_NA_EN.pdf"
## [102] "PCIJ_A_21_BrazilianLoans_FRA_BRA_1929-07-12_ANX_01_NA_NA_FR.pdf"
## [103] "PCIJ_A_22_FreeZonesUpperSavoyGex_FRA_CHE_1929-08-19_ORD_01_EV-SE_01_EN.
    pdf"
## [104] "PCIJ_A_22_FreeZonesUpperSavoyGex_FRA_CHE_1929-08-19_ORD_01_EV-SE_01_FR.
    pdf"
## [105] "PCIJ_A_22_FreeZonesUpperSavoyGex_FRA_CHE_1929-08-19_ORD_01_EV-SE_02_EN.
    pdf"
## [106] "PCIJ_A_22_FreeZonesUpperSavoyGex_FRA_CHE_1929-08-19_ORD_01_EV-SE_02_FR.
    pdf"
## [107] "PCIJ_A_22_FreeZonesUpperSavoyGex_FRA_CHE_1929-08-19_ORD_01_EV-SE_03_EN.
    pdf"
## [108] "PCIJ_A_22_FreeZonesUpperSavoyGex_FRA_CHE_1929-08-19_ORD_01_EV-SE_03_FR.
    pdf"
## [109] "PCIJ_A_22_FreeZonesUpperSavoyGex_FRA_CHE_1929-08-19_ORD_01_EV-SE_04_EN.
    pdf"
## [110] "PCIJ_A_22_FreeZonesUpperSavoyGex_FRA_CHE_1929-08-19_ORD_01_EV-SE_04_FR.
    pdf"
## [111] "PCIJ_A_22_FreeZonesUpperSavoyGex_FRA_CHE_1929-08-19_ORD_01_EV-SE_05_EN.
    pdf"
## [112] "PCIJ_A_22_FreeZonesUpperSavoyGex_FRA_CHE_1929-08-19_ORD_01_EV-SE_05_FR.
    pdf"
## [113] "PCIJ_A_23_OderCommission_GBR-CSK-DNK-FRA-DEU_POL_1929-08-15_ORD_01_EV
    _00_EN.pdf"
## [114] "PCIJ_A_23_OderCommission_GBR-CSK-DNK-FRA-DEU_POL_1929-08-15_ORD_01_EV
    _00_FR.pdf"
## [115] "PCIJ_A_23_OderCommission_GBR-CSK-DNK-FRA-DEU_POL_1929-08-20_ORD_01_EV
    _00_EN.pdf"
## [116] "PCIJ_A_23_OderCommission_GBR-CSK-DNK-FRA-DEU_POL_1929-08-20_ORD_01_EV
```

```
    _00_FR.pdf"
## [117] "PCIJ_A_23_OderCommission_GBR-CSK-DNK-FRA-DEU_POL_1929-08-15_ORD_02_TL
    _00_EN.pdf"
## [118] "PCIJ_A_23_OderCommission_GBR-CSK-DNK-FRA-DEU_POL_1929-08-15_ORD_02_TL
    _00_FR.pdf"
## [119] "PCIJ_A_24_FreeZonesUpperSavoyGex-SecondPhase_FRA_CHE_1930-12-06_ORD_01_
    SE_01_EN.pdf"
## [120] "PCIJ_A_24_FreeZonesUpperSavoyGex-SecondPhase_FRA_CHE_1930-12-06_ORD_01_
    SE_01_FR.pdf"
## [121] "PCIJ_A_24_FreeZonesUpperSavoyGex-SecondPhase_FRA_CHE_1930-12-06_ORD_01_
    SE_02_EN.pdf"
## [122] "PCIJ_A_24_FreeZonesUpperSavoyGex-SecondPhase_FRA_CHE_1930-12-06_ORD_01_
    SE_02_FR.pdf"
## [123] "PCIJ_AB_40_GermanMinoritySchools_DEU_POL_1931-05-15_ADV_01_NA_01_EN.pdf
    "
## [124] "PCIJ_AB_40_GermanMinoritySchools_DEU_POL_1931-05-15_ADV_01_NA_01_FR.pdf
    "
## [125] "PCIJ_AB_40_GermanMinoritySchools_DEU_POL_1931-05-15_ANX_01_NA_NA_EN.pdf
    "
## [126] "PCIJ_AB_40_GermanMinoritySchools_DEU_POL_1931-05-15_ANX_01_NA_NA_FR.pdf
    "
## [127] "PCIJ_AB_40_GermanMinoritySchools_DEU_POL_1931-05-15_ANX_02_NA_NA_EN.pdf
    "
## [128] "PCIJ_AB_40_GermanMinoritySchools_DEU_POL_1931-05-15_ANX_02_NA_NA_FR.pdf
    "
## [129] "PCIJ_AB_41_CustomsRegime_LNC_NA_1931-09-05_ADV_01_NA_01_EN.pdf"
## [130] "PCIJ_AB_41_CustomsRegime_LNC_NA_1931-09-05_ADV_01_NA_01_FR.pdf"
## [131] "PCIJ_AB_41_CustomsRegime_LNC_NA_1931-09-05_ADV_01_NA_02_EN.pdf"
## [132] "PCIJ_AB_41_CustomsRegime_LNC_NA_1931-09-05_ADV_01_NA_02_FR.pdf"
## [133] "PCIJ_AB_41_CustomsRegime_LNC_NA_1931-09-05_ANX_02_NA_NA_EN.pdf"
## [134] "PCIJ_AB_41_CustomsRegime_LNC_NA_1931-09-05_ANX_02_NA_NA_FR.pdf"
## [135] "PCIJ_AB_41_CustomsRegime_LNC_NA_1931-09-05_ANX_03_NA_NA_EN.pdf"
## [136] "PCIJ_AB_41_CustomsRegime_LNC_NA_1931-09-05_ANX_03_NA_NA_FR.pdf"
## [137] "PCIJ_AB_41_CustomsRegime_LNC_NA_1931-07-20_ORD_01_AJ_00_EN.pdf"
## [138] "PCIJ_AB_41_CustomsRegime_LNC_NA_1931-07-20_ORD_01_AJ_00_FR.pdf"
## [139] "PCIJ_AB_41_CustomsRegime_LNC_NA_1931-07-20_ORD_01_AJ_01_EN.pdf"
## [140] "PCIJ_AB_41_CustomsRegime_LNC_NA_1931-07-20_ORD_01_AJ_01_FR.pdf"
## [141] "PCIJ_AB_43_PolishWarVessels_LNC_NA_1931-12-11_ADV_01_NA_01_EN.pdf"
## [142] "PCIJ_AB_43_PolishWarVessels_LNC_NA_1931-12-11_ADV_01_NA_01_FR.pdf"
## [143] "PCIJ_AB_43_PolishWarVessels_LNC_NA_1931-12-11_ANX_01_NA_NA_EN.pdf"
## [144] "PCIJ_AB_43_PolishWarVessels_LNC_NA_1931-12-11_ANX_01_NA_NA_FR.pdf"
## [145] "PCIJ_AB_44_TreatmentPolishNationals_LNC_NA_1932-02-04_ADV_01_NA_01_EN.
    pdf"
## [146] "PCIJ_AB_44_TreatmentPolishNationals_LNC_NA_1932-02-04_ADV_01_NA_01_FR.
    pdf"
## [147] "PCIJ_AB_44_TreatmentPolishNationals_LNC_NA_1932-02-04_ADV_01_NA_02_EN.
    pdf"
## [148] "PCIJ_AB_44_TreatmentPolishNationals_LNC_NA_1932-02-04_ADV_01_NA_02_FR.
    pdf"
## [149] "PCIJ_AB_44_TreatmentPolishNationals_LNC_NA_1932-02-04_ANX_01_NA_NA_EN.
    pdf"
## [150] "PCIJ_AB_44_TreatmentPolishNationals_LNC_NA_1932-02-04_ANX_01_NA_NA_FR.
    pdf"
## [151] "PCIJ_AB_45_GrecoBulgarianAgreement_LNC_NA_1932-03-08_ANX_01_NA_NA_EN.
    pdf"
## [152] "PCIJ_AB_45_GrecoBulgarianAgreement_LNC_NA_1932-03-08_ANX_01_NA_NA_FR.
```

```
    pdf"
## [153] "PCIJ_AB_46_FreeZonesUpperSavoyGex_FRA_CHE_1932-06-07_JUD_01_ME_01_EN.
    pdf"
## [154] "PCIJ_AB_46_FreeZonesUpperSavoyGex_FRA_CHE_1932-06-07_JUD_01_ME_01_FR.
    pdf"
## [155] "PCIJ_AB_46_FreeZonesUpperSavoyGex_FRA_CHE_1932-06-07_JUD_01_ME_02_EN.
    pdf"
## [156] "PCIJ_AB_46_FreeZonesUpperSavoyGex_FRA_CHE_1932-06-07_JUD_01_ME_02_FR.
    pdf"
## [157] "PCIJ_AB_46_FreeZonesUpperSavoyGex_FRA_CHE_1932-06-07_JUD_01_ME_03_EN.
    pdf"
## [158] "PCIJ_AB_46_FreeZonesUpperSavoyGex_FRA_CHE_1932-06-07_JUD_01_ME_03_FR.
    pdf"
## [159] "PCIJ_AB_46_FreeZonesUpperSavoyGex_FRA_CHE_1931-08-06_ORD_01_TL-DH_00_EN
    .pdf"
## [160] "PCIJ_AB_46_FreeZonesUpperSavoyGex_FRA_CHE_1931-08-06_ORD_01_TL-DH_00_FR
    .pdf"
## [161] "PCIJ_AB_46_FreeZonesUpperSavoyGex_FRA_CHE_1932-06-07_ANX_01_NA_NA_EN.
    pdf"
## [162] "PCIJ_AB_46_FreeZonesUpperSavoyGex_FRA_CHE_1932-06-07_ANX_01_NA_NA_FR.
    pdf"
## [163] "PCIJ_AB_47_InterpretationMemelStatute_GBR-FRA-ITA-JPN_LTU_1932-06-24_
    JUD_01_PO_00_EN.pdf"
## [164] "PCIJ_AB_47_InterpretationMemelStatute_GBR-FRA-ITA-JPN_LTU_1932-06-24_
    JUD_01_PO_00_FR.pdf"
## [165] "PCIJ_AB_47_InterpretationMemelStatute_GBR-FRA-ITA-JPN_LTU_1932-06-24_
    JUD_01_PO_01_EN.pdf"
## [166] "PCIJ_AB_47_InterpretationMemelStatute_GBR-FRA-ITA-JPN_LTU_1932-06-24_
    JUD_01_PO_01_FR.pdf"
## [167] "PCIJ_AB_47_InterpretationMemelStatute_GBR-FRA-ITA-JPN_LTU_1932-06-24_
    ORD_01_TL_00_EN.pdf"
## [168] "PCIJ_AB_47_InterpretationMemelStatute_GBR-FRA-ITA-JPN_LTU_1932-06-24_
    ORD_01_TL_00_FR.pdf"
## [169] "PCIJ_AB_49_InterpretationMemelStatute_GBR-FRA-ITA-JPN_LTU_1932-08-11_
    JUD_01_ME_01_EN.pdf"
## [170] "PCIJ_AB_49_InterpretationMemelStatute_GBR-FRA-ITA-JPN_LTU_1932-08-11_
    JUD_01_ME_01_FR.pdf"
## [171] "PCIJ_AB_49_InterpretationMemelStatute_GBR-FRA-ITA-JPN_LTU_1932-08-11_
    JUD_01_ME_02_EN.pdf"
## [172] "PCIJ_AB_49_InterpretationMemelStatute_GBR-FRA-ITA-JPN_LTU_1932-08-11_
    JUD_01_ME_02_FR.pdf"
## [173] "PCIJ_AB_49_InterpretationMemelStatute_GBR-FRA-ITA-JPN_LTU_1932-08-11_
    ANX_01_NA_NA_EN.pdf"
## [174] "PCIJ_AB_49_InterpretationMemelStatute_GBR-FRA-ITA-JPN_LTU_1932-08-11_
    ANX_01_NA_NA_FR.pdf"
## [175] "PCIJ_AB_50_WomenNightWork_LNC_NA_1932-11-15_ADV_01_NA_01_EN.pdf"
## [176] "PCIJ_AB_50_WomenNightWork_LNC_NA_1932-11-15_ADV_01_NA_01_FR.pdf"
## [177] "PCIJ_AB_50_WomenNightWork_LNC_NA_1932-11-15_ANX_01_NA_NA_EN.pdf"
## [178] "PCIJ_AB_50_WomenNightWork_LNC_NA_1932-11-15_ANX_01_NA_NA_FR.pdf"
## [179] "PCIJ_AB_53_EasternGreenland_DNK_NOR_1933-04-05_JUD_01_ME_01_EN.pdf"
## [180] "PCIJ_AB_53_EasternGreenland_DNK_NOR_1933-04-05_JUD_01_ME_01_FR.pdf"
## [181] "PCIJ_AB_53_EasternGreenland_DNK_NOR_1933-04-05_JUD_01_ME_02_EN.pdf"
## [182] "PCIJ_AB_53_EasternGreenland_DNK_NOR_1933-04-05_JUD_01_ME_02_FR.pdf"
## [183] "PCIJ_AB_53_EasternGreenland_DNK_NOR_1933-04-05_JUD_01_ME_03_EN.pdf"
## [184] "PCIJ_AB_53_EasternGreenland_DNK_NOR_1933-04-05_JUD_01_ME_03_FR.pdf"
## [185] "PCIJ_AB_58_PolishAgrarianReform_DEU_POL_1933-07-29_ORD_01_IM_01_EN.pdf"
```

```
## [186] "PCIJ_AB_58_PolishAgrarianReform_DEU_POL_1933-07-29_ORD_01_IM_01_FR.pdf"
## [187] "PCIJ_AB_58_PolishAgrarianReform_DEU_POL_1933-07-29_ORD_01_IM_02_EN.pdf"
## [188] "PCIJ_AB_58_PolishAgrarianReform_DEU_POL_1933-07-29_ORD_01_IM_02_FR.pdf"
## [189] "PCIJ_AB_58_PolishAgrarianReform_DEU_POL_1933-07-29_ORD_01_IM_03_EN.pdf"
## [190] "PCIJ_AB_58_PolishAgrarianReform_DEU_POL_1933-07-29_ORD_01_IM_03_FR.pdf"
## [191] "PCIJ_AB_58_PolishAgrarianReform_DEU_POL_1933-07-29_ANX_01_NA_NA_EN.pdf"
## [192] "PCIJ_AB_58_PolishAgrarianReform_DEU_POL_1933-07-29_ANX_01_NA_NA_FR.pdf"
## [193] "PCIJ_AB_61_Pazmany_CSK_HUN_1933-12-15_JUD_01_ME_01_EN.pdf"
## [194] "PCIJ_AB_61_Pazmany_CSK_HUN_1933-12-15_JUD_01_ME_01_FR.pdf"
## [195] "PCIJ_AB_61_Pazmany_CSK_HUN_1933-12-15_ANX_01_NA_NA_EN.pdf"
## [196] "PCIJ_AB_61_Pazmany_CSK_HUN_1933-12-15_ANX_01_NA_NA_FR.pdf"
## [197] "PCIJ_AB_62_Lighthouses_FRA_GRC_1934-03-17_JUD_01_ME_01_EN.pdf"
## [198] "PCIJ_AB_62_Lighthouses_FRA_GRC_1934-03-17_JUD_01_ME_01_FR.pdf"
## [199] "PCIJ_AB_62_Lighthouses_FRA_GRC_1934-03-17_JUD_01_ME_02_EN.pdf"
## [200] "PCIJ_AB_62_Lighthouses_FRA_GRC_1934-03-17_JUD_01_ME_02_FR.pdf"
## [201] "PCIJ_AB_62_Lighthouses_FRA_GRC_1934-03-17_ANX_01_NA_NA_EN.pdf"
## [202] "PCIJ_AB_62_Lighthouses_FRA_GRC_1934-03-17_ANX_01_NA_NA_FR.pdf"
## [203] "PCIJ_AB_63_OscarChinn_GBR_BEL_1934-12-12_JUD_01_ME_01_EN.pdf"
## [204] "PCIJ_AB_63_OscarChinn_GBR_BEL_1934-12-12_JUD_01_ME_01_FR.pdf"
## [205] "PCIJ_AB_63_OscarChinn_GBR_BEL_1934-12-12_JUD_01_ME_02_EN.pdf"
## [206] "PCIJ_AB_63_OscarChinn_GBR_BEL_1934-12-12_JUD_01_ME_02_FR.pdf"
## [207] "PCIJ_AB_63_OscarChinn_GBR_BEL_1934-12-12_JUD_01_ME_03_EN.pdf"
## [208] "PCIJ_AB_63_OscarChinn_GBR_BEL_1934-12-12_JUD_01_ME_03_FR.pdf"
## [209] "PCIJ_AB_63_OscarChinn_GBR_BEL_1934-12-12_JUD_01_ME_04_EN.pdf"
## [210] "PCIJ_AB_63_OscarChinn_GBR_BEL_1934-12-12_JUD_01_ME_04_FR.pdf"
## [211] "PCIJ_AB_63_OscarChinn_GBR_BEL_1934-12-12_JUD_01_ME_05_EN.pdf"
## [212] "PCIJ_AB_63_OscarChinn_GBR_BEL_1934-12-12_JUD_01_ME_05_FR.pdf"
## [213] "PCIJ_AB_63_OscarChinn_GBR_BEL_1934-12-12_ANX_01_NA_NA_EN.pdf"
## [214] "PCIJ_AB_63_OscarChinn_GBR_BEL_1934-12-12_ANX_01_NA_NA_FR.pdf"
## [215] "PCIJ_AB_64_MinoritySchoolsAlbania_LNC_NA_1935-04-06_ADV_01_NA_01_EN.pdf
    "
## [216] "PCIJ_AB_64_MinoritySchoolsAlbania_LNC_NA_1935-04-06_ADV_01_NA_01_FR.pdf
    "
## [217] "PCIJ_AB_64_MinoritySchoolsAlbania_LNC_NA_1935-04-06_ANX_01_NA_NA_EN.pdf
    "
## [218] "PCIJ_AB_64_MinoritySchoolsAlbania_LNC_NA_1935-04-06_ANX_01_NA_NA_FR.pdf
    "
## [219] "PCIJ_AB_65_DanzigLegislativeDecrees_LNC_NA_1935-12-04_ADV_01_NA_01_EN.
    pdf"
## [220] "PCIJ_AB_65_DanzigLegislativeDecrees_LNC_NA_1935-12-04_ADV_01_NA_01_FR.
    pdf"
## [221] "PCIJ_AB_65_DanzigLegislativeDecrees_LNC_NA_1935-12-04_ADV_01_NA_02_EN.
    pdf"
## [222] "PCIJ_AB_65_DanzigLegislativeDecrees_LNC_NA_1935-12-04_ADV_01_NA_02_FR.
    pdf"
## [223] "PCIJ_AB_65_DanzigLegislativeDecrees_LNC_NA_1935-12-04_ADV_01_NA_03_EN.
    pdf"
## [224] "PCIJ_AB_65_DanzigLegislativeDecrees_LNC_NA_1935-12-04_ADV_01_NA_03_FR.
    pdf"
## [225] "PCIJ_AB_65_DanzigLegislativeDecrees_LNC_NA_1935-10-31_ORD_01_AJ_00_EN.
    pdf"
## [226] "PCIJ_AB_65_DanzigLegislativeDecrees_LNC_NA_1935-10-31_ORD_01_AJ_00_FR.
    pdf"
## [227] "PCIJ_AB_65_DanzigLegislativeDecrees_LNC_NA_1935-12-04_ANX_01_NA_NA_EN.
    pdf"
## [228] "PCIJ_AB_65_DanzigLegislativeDecrees_LNC_NA_1935-12-04_ANX_01_NA_NA_FR.
```

```
    pdf"
## [229] "PCIJ_AB_68_PajzsCsakyEsterhazy_HUN_YUG_1936-12-16_JUD_01_ME_01_EN.pdf"
## [230] "PCIJ_AB_68_PajzsCsakyEsterhazy_HUN_YUG_1936-12-16_JUD_01_ME_01_FR.pdf"
## [231] "PCIJ_AB_68_PajzsCsakyEsterhazy_HUN_YUG_1936-12-16_JUD_01_ME_02_EN.pdf"
## [232] "PCIJ_AB_68_PajzsCsakyEsterhazy_HUN_YUG_1936-12-16_JUD_01_ME_02_FR.pdf"
## [233] "PCIJ_AB_68_PajzsCsakyEsterhazy_HUN_YUG_1936-12-16_JUD_01_ME_03_EN.pdf"
## [234] "PCIJ_AB_68_PajzsCsakyEsterhazy_HUN_YUG_1936-12-16_JUD_01_ME_03_FR.pdf"
## [235] "PCIJ_AB_68_PajzsCsakyEsterhazy_HUN_YUG_1936-12-16_JUD_01_ME_04_EN.pdf"
## [236] "PCIJ_AB_68_PajzsCsakyEsterhazy_HUN_YUG_1936-12-16_JUD_01_ME_04_FR.pdf"
## [237] "PCIJ_AB_68_PajzsCsakyEsterhazy_HUN_YUG_1936-12-16_JUD_01_ME_05_EN.pdf"
## [238] "PCIJ_AB_68_PajzsCsakyEsterhazy_HUN_YUG_1936-12-16_JUD_01_ME_05_FR.pdf"
## [239] "PCIJ_AB_68_PajzsCsakyEsterhazy_HUN_YUG_1936-12-16_ANX_01_NA_NA_EN.pdf"
## [240] "PCIJ_AB_68_PajzsCsakyEsterhazy_HUN_YUG_1936-12-16_ANX_01_NA_NA_FR.pdf"
## [241] "PCIJ_AB_70_Meuse_NLD_BEL_1937-06-28_JUD_01_ME_01_EN.pdf"
## [242] "PCIJ_AB_70_Meuse_NLD_BEL_1937-06-28_JUD_01_ME_01_FR.pdf"
## [243] "PCIJ_AB_70_Meuse_NLD_BEL_1937-06-28_JUD_01_ME_02_EN.pdf"
## [244] "PCIJ_AB_70_Meuse_NLD_BEL_1937-06-28_JUD_01_ME_02_FR.pdf"
## [245] "PCIJ_AB_70_Meuse_NLD_BEL_1937-06-28_JUD_01_ME_03_EN.pdf"
## [246] "PCIJ_AB_70_Meuse_NLD_BEL_1937-06-28_JUD_01_ME_03_FR.pdf"
## [247] "PCIJ_AB_70_Meuse_NLD_BEL_1937-06-28_JUD_01_ME_04_EN.pdf"
## [248] "PCIJ_AB_70_Meuse_NLD_BEL_1937-06-28_JUD_01_ME_04_FR.pdf"
## [249] "PCIJ_AB_70_Meuse_NLD_BEL_1937-06-28_JUD_01_ME_05_EN.pdf"
## [250] "PCIJ_AB_70_Meuse_NLD_BEL_1937-06-28_JUD_01_ME_05_FR.pdf"
## [251] "PCIJ_AB_70_Meuse_NLD_BEL_1937-06-28_ANX_02_NA_NA_EN.pdf"
## [252] "PCIJ_AB_70_Meuse_NLD_BEL_1937-06-28_ANX_02_NA_NA_FR.pdf"
## [253] "PCIJ_AB_71_LighthousesCreteSamos_FRA_GRC_1937-10-08_JUD_01_ME_01_EN.pdf
    "
## [254] "PCIJ_AB_71_LighthousesCreteSamos_FRA_GRC_1937-10-08_JUD_01_ME_01_FR.pdf
    "
## [255] "PCIJ_AB_71_LighthousesCreteSamos_FRA_GRC_1937-10-08_JUD_01_ME_02_EN.pdf
    "
## [256] "PCIJ_AB_71_LighthousesCreteSamos_FRA_GRC_1937-10-08_JUD_01_ME_02_FR.pdf
    "
## [257] "PCIJ_AB_71_LighthousesCreteSamos_FRA_GRC_1937-10-08_JUD_01_ME_03_EN.pdf
    "
## [258] "PCIJ_AB_71_LighthousesCreteSamos_FRA_GRC_1937-10-08_JUD_01_ME_03_FR.pdf
    "
## [259] "PCIJ_AB_71_LighthousesCreteSamos_FRA_GRC_1937-10-08_JUD_01_ME_04_EN.pdf
    "
## [260] "PCIJ_AB_71_LighthousesCreteSamos_FRA_GRC_1937-10-08_JUD_01_ME_04_FR.pdf
    "
## [261] "PCIJ_AB_71_LighthousesCreteSamos_FRA_GRC_1937-10-08_ANX_01_NA_NA_EN.pdf
    "
## [262] "PCIJ_AB_71_LighthousesCreteSamos_FRA_GRC_1937-10-08_ANX_01_NA_NA_FR.pdf
    "
## [263] "PCIJ_AB_72_Borchgrave_BEL_ESP_1937-11-06_ORD_01_TL_00_EN.pdf"
## [264] "PCIJ_AB_72_Borchgrave_BEL_ESP_1937-11-06_ORD_01_TL_00_FR.pdf"
## [265] "PCIJ_AB_72_Borchgrave_BEL_ESP_1937-11-06_ANX_01_NA_NA_EN.pdf"
## [266] "PCIJ_AB_72_Borchgrave_BEL_ESP_1937-11-06_ANX_01_NA_NA_FR.pdf"
## [267] "PCIJ_AB_74_PhosphatesMarocco_ITA_FRA_1938-06-14_JUD_01_PO_01_EN.pdf"
## [268] "PCIJ_AB_74_PhosphatesMarocco_ITA_FRA_1938-06-14_JUD_01_PO_01_FR.pdf"
## [269] "PCIJ_AB_74_PhosphatesMarocco_ITA_FRA_1938-06-14_JUD_01_PO_02_EN.pdf"
## [270] "PCIJ_AB_74_PhosphatesMarocco_ITA_FRA_1938-06-14_JUD_01_PO_02_FR.pdf"
## [271] "PCIJ_AB_74_PhosphatesMarocco_ITA_FRA_1938-06-14_ANX_01_NA_NA_EN.pdf"
## [272] "PCIJ_AB_74_PhosphatesMarocco_ITA_FRA_1938-06-14_ANX_01_NA_NA_FR.pdf"
## [273] "PCIJ_AB_76_PanevezysSaldutiskisRailway_EST_LTU_1939-02-28_JUD_01_ME_01_
```

```
    EN.pdf"
## [274] "PCIJ_AB_76_PanevezysSaldutiskisRailway_EST_LTU_1939-02-28_JUD_01_ME_01_
    FR.pdf"
## [275] "PCIJ_AB_76_PanevezysSaldutiskisRailway_EST_LTU_1939-02-28_JUD_01_ME_02_
    EN.pdf"
## [276] "PCIJ_AB_76_PanevezysSaldutiskisRailway_EST_LTU_1939-02-28_JUD_01_ME_02_
    FR.pdf"
## [277] "PCIJ_AB_76_PanevezysSaldutiskisRailway_EST_LTU_1939-02-28_JUD_01_ME_03_
    EN.pdf"
## [278] "PCIJ_AB_76_PanevezysSaldutiskisRailway_EST_LTU_1939-02-28_JUD_01_ME_03_
    FR.pdf"
## [279] "PCIJ_AB_76_PanevezysSaldutiskisRailway_EST_LTU_1939-02-28_JUD_01_ME_04_
    EN.pdf"
## [280] "PCIJ_AB_76_PanevezysSaldutiskisRailway_EST_LTU_1939-02-28_JUD_01_ME_04_
    FR.pdf"
## [281] "PCIJ_AB_76_PanevezysSaldutiskisRailway_EST_LTU_1939-02-28_ANX_01_NA_NA_
    EN.pdf"
## [282] "PCIJ_AB_76_PanevezysSaldutiskisRailway_EST_LTU_1939-02-28_ANX_01_NA_NA_
    FR.pdf"
## [283] "PCIJ_AB_77_ElectricityCompanySofiaBulgaria_BEL_BGR_1939-04-04_JUD_01_PO
    _01_EN.pdf"
## [284] "PCIJ_AB_77_ElectricityCompanySofiaBulgaria_BEL_BGR_1939-04-04_JUD_01_PO
    _01_FR.pdf"
## [285] "PCIJ_AB_77_ElectricityCompanySofiaBulgaria_BEL_BGR_1939-04-04_JUD_01_PO
    _02_EN.pdf"
## [286] "PCIJ_AB_77_ElectricityCompanySofiaBulgaria_BEL_BGR_1939-04-04_JUD_01_PO
    _02_FR.pdf"
## [287] "PCIJ_AB_77_ElectricityCompanySofiaBulgaria_BEL_BGR_1939-04-04_JUD_01_PO
    _03_EN.pdf"
## [288] "PCIJ_AB_77_ElectricityCompanySofiaBulgaria_BEL_BGR_1939-04-04_JUD_01_PO
    _03_FR.pdf"
## [289] "PCIJ_AB_77_ElectricityCompanySofiaBulgaria_BEL_BGR_1939-04-04_JUD_01_PO
    _04_EN.pdf"
## [290] "PCIJ_AB_77_ElectricityCompanySofiaBulgaria_BEL_BGR_1939-04-04_JUD_01_PO
    _04_FR.pdf"
## [291] "PCIJ_AB_77_ElectricityCompanySofiaBulgaria_BEL_BGR_1939-04-04_JUD_01_PO
    _05_EN.pdf"
## [292] "PCIJ_AB_77_ElectricityCompanySofiaBulgaria_BEL_BGR_1939-04-04_JUD_01_PO
    _05_FR.pdf"
## [293] "PCIJ_AB_77_ElectricityCompanySofiaBulgaria_BEL_BGR_1939-04-04_JUD_01_PO
    _06_EN.pdf"
## [294] "PCIJ_AB_77_ElectricityCompanySofiaBulgaria_BEL_BGR_1939-04-04_JUD_01_PO
    _06_FR.pdf"
## [295] "PCIJ_AB_77_ElectricityCompanySofiaBulgaria_BEL_BGR_1939-04-04_JUD_01_PO
    _07_EN.pdf"
## [296] "PCIJ_AB_77_ElectricityCompanySofiaBulgaria_BEL_BGR_1939-04-04_JUD_01_PO
    _07_FR.pdf"
## [297] "PCIJ_AB_77_ElectricityCompanySofiaBulgaria_BEL_BGR_1939-04-04_ORD_01_TL
    _00_EN.pdf"
## [298] "PCIJ_AB_77_ElectricityCompanySofiaBulgaria_BEL_BGR_1939-04-04_ORD_01_TL
    _00_FR.pdf"
## [299] "PCIJ_AB_77_ElectricityCompanySofiaBulgaria_BEL_BGR_1939-04-04_ANX_01_NA
    _NA_EN.pdf"
## [300] "PCIJ_AB_77_ElectricityCompanySofiaBulgaria_BEL_BGR_1939-04-04_ANX_01_NA
    _NA_FR.pdf"
## [301] "PCIJ_AB_78_SocieteCommercialeBelgique_BEL_GRC_1939-06-15_JUD_01_ME_01_
```

```
    EN.pdf"
## [302] "PCIJ_AB_78_SocieteCommercialeBelgique_BEL_GRC_1939-06-15_JUD_01_ME_01_
    FR.pdf"
## [303] "PCIJ_AB_78_SocieteCommercialeBelgique_BEL_GRC_1939-06-15_JUD_01_ME_02_
    EN.pdf"
## [304] "PCIJ_AB_78_SocieteCommercialeBelgique_BEL_GRC_1939-06-15_JUD_01_ME_02_
    FR.pdf"
## [305] "PCIJ_AB_78_SocieteCommercialeBelgique_BEL_GRC_1939-06-15_ANX_01_NA_NA_
    EN.pdf"
## [306] "PCIJ_AB_78_SocieteCommercialeBelgique_BEL_GRC_1939-06-15_ANX_01_NA_NA_
    FR.pdf"
## [307] "PCIJ_B_01_WorkersDelegateILO_LNC_NA_1922-05-22_APP_01_NA_NA_EN.pdf"
## [308] "PCIJ_B_01_WorkersDelegateILO_LNC_NA_1922-05-22_APP_01_NA_NA_FR.pdf"
## [309] "PCIJ_B_01_WorkersDelegateILO_LNC_NA_1922-07-31_ADV_01_NA_00_EN.pdf"
## [310] "PCIJ_B_01_WorkersDelegateILO_LNC_NA_1922-07-31_ADV_01_NA_00_FR.pdf"
## [311] "PCIJ_B_02_ILOCompetencePersonsAgriculture_LNC_NA_1922-08-12_ADV_01_NA
    _00_EN.pdf"
## [312] "PCIJ_B_02_ILOCompetencePersonsAgriculture_LNC_NA_1922-08-12_ADV_01_NA
    _00_FR.pdf"
## [313] "PCIJ_B_02_ILOCompetencePersonsAgriculture_LNC_NA_1922-05-22_APP_01_NA_
    NA_EN.pdf"
## [314] "PCIJ_B_02_ILOCompetencePersonsAgriculture_LNC_NA_1922-05-22_APP_01_NA_
    NA_FR.pdf"
## [315] "PCIJ_B_03_ILOCompetenceMethodsAgriculture_LNC_NA_1922-08-12_ADV_01_NA
    _00_EN.pdf"
## [316] "PCIJ_B_03_ILOCompetenceMethodsAgriculture_LNC_NA_1922-08-12_ADV_01_NA
    _00_FR.pdf"
## [317] "PCIJ_B_03_ILOCompetenceMethodsAgriculture_LNC_NA_1922-07-18_APP_01_NA_
    NA_EN.pdf"
## [318] "PCIJ_B_03_ILOCompetenceMethodsAgriculture_LNC_NA_1922-07-18_APP_01_NA_
    NA_FR.pdf"
## [319] "PCIJ_B_04_NationalityDecrees_LNC_NA_1923-02-07_ADV_01_NA_00_EN.pdf"
## [320] "PCIJ_B_04_NationalityDecrees_LNC_NA_1923-02-07_ADV_01_NA_00_FR.pdf"
## [321] "PCIJ_B_04_NationalityDecrees_LNC_NA_1922-11-06_APP_01_NA_NA_EN.pdf"
## [322] "PCIJ_B_04_NationalityDecrees_LNC_NA_1922-11-06_APP_01_NA_NA_FR.pdf"
## [323] "PCIJ_B_05_EasternCarelia_LNC_NA_1923-04-27_APP_01_NA_NA_EN.pdf"
## [324] "PCIJ_B_05_EasternCarelia_LNC_NA_1923-04-27_APP_01_NA_NA_FR.pdf"
## [325] "PCIJ_B_05_EasternCarelia_LNC_NA_1923-07-23_ADV_01_NA_00_EN.pdf"
## [326] "PCIJ_B_05_EasternCarelia_LNC_NA_1923-07-23_ADV_01_NA_00_FR.pdf"
## [327] "PCIJ_B_07_AcquisitionPolishNationality_LNC_NA_1923-09-15_ADV_01_NA_01_
    EN.pdf"
## [328] "PCIJ_B_07_AcquisitionPolishNationality_LNC_NA_1923-09-15_ADV_01_NA_01_
    FR.pdf"
## [329] "PCIJ_B_09_MonasterySaintNaoum_LNC_NA_1924-09-04_ADV_01_NA_00_EN.pdf"
## [330] "PCIJ_B_09_MonasterySaintNaoum_LNC_NA_1924-09-04_ADV_01_NA_00_FR.pdf"
## [331] "PCIJ_B_10_ExchangeGreekTurkishPopulations_LNC_NA_1925-02-21_ANX_01_NA_
    NA_EN.pdf"
## [332] "PCIJ_B_10_ExchangeGreekTurkishPopulations_LNC_NA_1925-02-21_ANX_01_NA_
    NA_FR.pdf"
## [333] "PCIJ_B_11_PostalServiceDanzig_LNC_NA_1925-05-16_ANX_01_NA_NA_EN.pdf"
## [334] "PCIJ_B_11_PostalServiceDanzig_LNC_NA_1925-05-16_ANX_01_NA_NA_FR.pdf"
## [335] "PCIJ_B_12_TreatyLausanne_LNC_NA_1925-11-21_ANX_01_NA_NA_EN.pdf"
## [336] "PCIJ_B_12_TreatyLausanne_LNC_NA_1925-11-21_ANX_01_NA_NA_FR.pdf"
## [337] "PCIJ_B_13_ILOCompetenceEmployer_LNC_NA_1926-07-23_ANX_01_NA_NA_EN.pdf"
## [338] "PCIJ_B_13_ILOCompetenceEmployer_LNC_NA_1926-07-23_ANX_01_NA_NA_FR.pdf"
## [339] "PCIJ_B_14_DanubeCommission_LNC_NA_1927-12-08_ADV_01_NA_01_EN.pdf"
```

```
## [340] "PCIJ_B_14_DanubeCommission_LNC_NA_1927-12-08_ADV_01_NA_01_FR.pdf"
## [341] "PCIJ_B_14_DanubeCommission_LNC_NA_1927-12-08_ADV_01_NA_02_EN.pdf"
## [342] "PCIJ_B_14_DanubeCommission_LNC_NA_1927-12-08_ADV_01_NA_02_FR.pdf"
## [343] "PCIJ_B_14_DanubeCommission_LNC_NA_1927-12-08_ADV_01_NA_03_EN.pdf"
## [344] "PCIJ_B_14_DanubeCommission_LNC_NA_1927-12-08_ADV_01_NA_03_FR.pdf"
## [345] "PCIJ_B_14_DanubeCommission_LNC_NA_1927-12-08_ANX_01_NA_NA_EN.pdf"
## [346] "PCIJ_B_14_DanubeCommission_LNC_NA_1927-12-08_ANX_01_NA_NA_FR.pdf"
## [347] "PCIJ_B_15_DanzigCourts_LNC_NA_1928-03-03_ANX_01_NA_NA_EN.pdf"
## [348] "PCIJ_B_15_DanzigCourts_LNC_NA_1928-03-03_ANX_01_NA_NA_FR.pdf"
## [349] "PCIJ_B_15_DanzigCourts_LNC_NA_1928-03-03_ANX_03_NA_NA_EN.pdf"
## [350] "PCIJ_B_15_DanzigCourts_LNC_NA_1928-03-03_ANX_03_NA_NA_FR.pdf"
## [351] "PCIJ_B_16_GrecoTurkishAgreement_LNC_NA_1928-08-28_ANX_01_NA_NA_EN.pdf"
## [352] "PCIJ_B_16_GrecoTurkishAgreement_LNC_NA_1928-08-28_ANX_01_NA_NA_FR.pdf"
## [353] "PCIJ_B_17_GrecoBulgarianCommunities_LNC_NA_1930-07-31_ANX_01_NA_NA_EN.
    pdf"
## [354] "PCIJ_B_17_GrecoBulgarianCommunities_LNC_NA_1930-07-31_ANX_01_NA_NA_FR.
    pdf"
## [355] "PCIJ_B_17_GrecoBulgarianCommunities_LNC_NA_1930-07-31_ANX_02_NA_NA_EN.
    pdf"
## [356] "PCIJ_B_17_GrecoBulgarianCommunities_LNC_NA_1930-07-31_ANX_02_NA_NA_FR.
    pdf"
## [357] "PCIJ_B_18_DanzigILO_LNC_NA_1930-08-26_ADV_01_NA_01_EN.pdf"
## [358] "PCIJ_B_18_DanzigILO_LNC_NA_1930-08-26_ADV_01_NA_01_FR.pdf"
## [359] "PCIJ_B_18_DanzigILO_LNC_NA_1930-08-26_ADV_01_NA_02_EN.pdf"
## [360] "PCIJ_B_18_DanzigILO_LNC_NA_1930-08-26_ADV_01_NA_02_FR.pdf"
```

10.8 Shutdown Fork Cluster

```
stopCluster(cl)
```

10.9 Clean up Multilingual Originals

```
files.pdf.bi <- list.files(pattern = "BI\\.pdf$")

length(files.pdf.bi)
```

```
## [1] 260
```

```
file_move(files.pdf.bi,
          "MULT_PDF_ORIGINAL_FULL")
```

10.10 Copy English and French Originals

```
file_copy("PCIJ_A_03_Neuilly_BGR_GRC_1924-09-12_ANX_01_NA_NA_EN.pdf",
          "MULT_PDF_ORIGINAL_FULL")

file_copy("PCIJ_A_07_GermanInterestsUpperSilesia_DEU_POL_1926-05-25_ANX_01_NA_NA_
    FR.pdf",
          "MULT_PDF_ORIGINAL_FULL")

file_copy("PCIJ_AB_70_Meuse_NLD_BEL_1937-06-28_ANX_01_NA_NA_FR.pdf",
          "MULT_PDF_ORIGINAL_FULL")
```

11 Detect Missing Counterparts for each Language Variant

```
files.de <- list.files(pattern = "DE\\.pdf")
files.en <- list.files(pattern = "EN\\.pdf")
files.fr <- list.files(pattern = "FR\\.pdf")
```

11.1 Difference between French and English File Lists

```
abs(length(files.en) - length(files.fr))
```

```
## [1] 2
```

11.2 Show Missing French Documents

```
files.fr.temp <- gsub("FR\\.pdf",
                      "EN\\.pdf",
                      files.fr)

frenchmissing <- setdiff(files.en,
                         files.fr.temp)

frenchmissing <- gsub("EN\\.pdf",
                      "FR\\.pdf",
                      frenchmissing)

print(frenchmissing)
```

```
## [1] "PCIJ_A_03_Neuilly_BGR_GRC_1924-09-12_ANX_01_NA_NA_FR.pdf"
```

11.3 Show Missing English Documents

```
files.en.temp <- gsub("EN\\.pdf",
                      "FR\\.pdf",
                      files.en)

englishmissing <- setdiff(files.fr,
                          files.en.temp)

englishmissing <- gsub("FR\\.pdf",
                       "EN\\.pdf",
                       englishmissing)

print(englishmissing)
```

```
## [1] "PCIJ_A_07_GermanInterestsUpperSilesia_DEU_POL_1926-05-25_ANX_01_NA_NA_EN.
    pdf"
## [2] "PCIJ_AB_70_Meuse_NLD_BEL_1937-06-28_ANX_01_NA_NA_EN.pdf"
## [3] "PCIJ_B_15_DanzigCourts_LNC_NA_1928-03-03_ANX_02_NA_NA_EN.pdf"
```

11.4 Show German Documents

```
print(files.de)
```

```
## [1] "PCIJ_AB_41_CustomsRegime_LNC_NA_1931-09-05_ANX_01_NA_NA_DE.pdf"
## [2] "PCIJ_B_15_DanzigCourts_LNC_NA_1928-03-03_ANX_02_NA_NA_DE.pdf"
```

11.5 Clean up German Originals

Note: Strictly speaking one of the German documents (the Danzig Courts file) is not a true original, as it was split from a bilingual file. However the quality of the document (scan and OCR) is original, so it is stored with the other originals to avoid creating another variant for a single document.

```
file_move(files.de,
          "MULT_PDF_ORIGINAL_FULL")
```

12 Text Extraction Module

12.1 Define Set of Files to Process

```
files.pdf <- list.files(pattern = "\\.pdf$",
                        ignore.case = TRUE)
```

12.2 Number of Files to Process

```
length(files.pdf)
```

```
## [1] 520
```

12.3 Show Function: f.dopar.pagenums

```
print(f.dopar.pagenums)
```

function(x, sum = FALSE, threads = detectCores()){

```
print(paste("Parallel processing using", threads, "threads."))

cl <- makeForkCluster(threads)
registerDoParallel(cl)

pagenums <- foreach(filename = x,
                    .combine = 'c',
                    .errorhandling = 'remove',
                    .inorder = TRUE) %dopar% {
                        pdf_length(filename)
                    }
stopCluster(cl)

if (sum == TRUE){
    sum.out <- sum(pagenums)
    print(paste("Total number of pages:", sum.out))
    return(sum.out)
}else{
    return(pagenums)
}
```

}

12.4 Count Pages

```
f.dopar.pagenums(files.pdf,
                 sum = TRUE,
                 threads = fullCores)
```

```
## [1] "Parallel processing using 16 threads."
## [1] "Total number of pages: 6252"
```

```
## [1] 6252
```

12.5 Show Function: f.dopar.pdfextract

```
print(f.dopar.pdfextract)
```

function(x, threads = detectCores()){

```
begin.extract <- Sys.time()

print(paste("Parallel processing using", threads, "threads. Begin at", begin.
    extract))


cl <- makeForkCluster(threads)
registerDoParallel(cl)

newnames <- gsub("\\.pdf",
                 "\\.txt",
                 x)

result <- foreach(i = seq_along(x),
                  .errorhandling = 'pass') %dopar% {

                      ## Extract text layer from PDF
                      pdf.extracted <- pdf_text(x[i])

                      ## Write TXT to Disk
                      write.table(pdf.extracted,
                                  newnames[i],
                                  quote = FALSE,
                                  row.names = FALSE,
                                  col.names = FALSE)
               }
stopCluster(cl)

end.extract <- Sys.time()
```

```
duration.extract <- end.extract - begin.extract

print(paste0("Processed ",
              length(result),
              " files. Runtime was ",
              round(duration.extract,
                    digits = 2),
              " ",
              attributes(duration.extract)$units,
              ". Ended at ",
             end.extract, "."))

return(result)
```

}

12.6 Extract Text

```
result <- f.dopar.pdfextract(files.pdf,
                             threads = fullCores)
```

```
## [1] "Parallel processing using 16 threads. Begin at 2022-09-06 19:06:44"
## [1] "Processed 520 files. Runtime was 1 secs. Ended at 2022-09-06 19:06:45."
```

12.7 Move Extracted TXT Files

```
txt.extracted.en <- list.files(pattern = "EN\\.txt")
txt.extracted.fr <- list.files(pattern = "FR\\.txt")

file_move(txt.extracted.en,
          "EN_TXT_EXTRACTED_FULL")

file_move(txt.extracted.fr,
          "FR_TXT_EXTRACTED_FULL")
```

13 Tesseract OCR Module

13.1 Show Function: f.dopar.pdfocr

```
print(f.dopar.pdfocr)
```

function(x, dpi = 300, lang = "eng," output = "pdf txt," jobs = round(detectCores() / 4)){

```
begin.ocr <- Sys.time()

print(paste("Parallel processing running", jobs, "jobs. Begin at", begin.ocr))

cl <- makeForkCluster(jobs)
registerDoParallel(cl)


result <- foreach(file = x,
                  .combine = 'c') %dopar% {

                      name.tiff <- gsub("\\.pdf",
                                        "\\.tiff",
                                        file)

                      name.out <- gsub("\\.pdf",
                                       "_TESSERACT",
                                       file)

                      system2("convert",
                              paste("-density",
                                    dpi,
                                    "-depth 8 -compress LZW -strip -background
    white -alpha off",
                                    file,
                                    name.tiff))

                      system2("tesseract",
                              paste(name.tiff,
                                    name.out,
                                    "-l",
                                    lang,
                                    output))

                      unlink(name.tiff)
                  }

stopCluster(cl)

end.ocr <- Sys.time()
duration.ocr <- end.ocr - begin.ocr

print(paste0("Processed ",
```

```
              length(result),
              " files. Runtime was ",
              round(duration.ocr,
                    digits = 2),
              " ",
              attributes(duration.ocr)$units,
              ". Ended at ",
             end.ocr, "."))

return(result)
```

}

13.2 English

13.2.1 Set of English Documents to Process

```
files.ocr.en <- list.files(pattern = "EN\\.pdf")
```

13.2.2 Number of English Documents to Process

```
length(files.ocr.en)
```

```
## [1] 259
```

13.2.3 Number of English Pages to Process

```
f.dopar.pagenums(files.ocr.en,
                 sum = TRUE,
                 threads = fullCores)
```

```
## [1] "Parallel processing using 16 threads."
## [1] "Total number of pages: 3154"
```

```
## [1] 3154
```

13.2.4 Run OCR on English Documents

Note: Training data is set to include both English and French. Lengthy quotations in a non-dominant language are common in international law. Order in language setting matters and for English documents "eng" is set as the primary training data.

```
result <- f.dopar.pdfocr(files.ocr.en,
                         dpi = ocr.dpi,
                         lang = "eng+fra",
                         output = "pdf txt",
                         jobs = ocrCores)
```

```
## [1] "Parallel processing running 5 jobs. Begin at 2022-09-06 19:06:45"
## [1] "Processed 259 files. Runtime was 25.17 mins. Ended at 2022-09-06
    19:31:55."
```

13.3 French

13.3.1 Set of French Documents to Process

```
files.ocr.fr <- list.files(pattern = "FR\\.pdf")
```

13.3.2 Number of French Documents to Process

```
length(files.ocr.fr)
```

```
## [1] 261
```

13.3.3 Number of French Pages to Process

```
f.dopar.pagenums(files.ocr.fr,
                 sum = TRUE,
                 threads = fullCores)
```

```
## [1] "Parallel processing using 16 threads."
## [1] "Total number of pages: 3098"
```

```
## [1] 3098
```

13.3.4 Run OCR on French Documents

Note: Training data is set to include both French and English. Lengthy quotations in a non-dominant language are common in international law. Order in language setting matters and for French documents "fra" is set as the primary training data.

```
result <- f.dopar.pdfocr(files.ocr.fr,
                         dpi = ocr.dpi,
                         lang = "fra+eng",
                         output = "pdf txt",
                         jobs = ocrCores)
```

```
## [1] "Parallel processing running 5 jobs. Begin at 2022-09-06 19:31:55"
## [1] "Processed 261 files. Runtime was 28.48 mins. Ended at 2022-09-06
    20:00:24."
```

13.4 Rename Files

```
files.pdf <- list.files(pattern = "\\.pdf$")

files.pdf.enhanced <- gsub("_TESSERACT.pdf",
                           "_ENHANCED.pdf",
                           files.pdf)

file.rename(files.pdf,
            files.pdf.enhanced)
```

```
files.txt <- list.files(pattern = "\\.txt$")

files.txt.new <- gsub("_TESSERACT.txt",
                      ".txt",
                      files.txt)

file.rename(files.txt,
            files.txt.new)
```

13.5 Move TXT files

```
files.ocr.txt.en <- list.files(pattern = "EN\\.txt")
files.ocr.txt.fr <- list.files(pattern = "FR\\.txt")

file_move(files.ocr.txt.en,
          "EN_TXT_TESSERACT_FULL")

file_move(files.ocr.txt.fr,
          "FR_TXT_TESSERACT_FULL")
```

13.6 Move PDF files

```
files.ocr.pdf.enhanced.en <- list.files(pattern = "EN_ENHANCED\\.pdf")
files.ocr.pdf.enhanced.fr <- list.files(pattern = "FR_ENHANCED\\.pdf")

files.ocr.pdf.original.en <- list.files(pattern = "EN\\.pdf")
files.ocr.pdf.original.fr <- list.files(pattern = "FR\\.pdf")


file_move(files.ocr.pdf.enhanced.en,
          "EN_PDF_ENHANCED_FULL")

file_move(files.ocr.pdf.enhanced.fr,
          "FR_PDF_ENHANCED_FULL")

file_move(files.ocr.pdf.original.en,
          "EN_PDF_ORIGINALSPLIT_FULL")

file_move(files.ocr.pdf.original.fr,
          "FR_PDF_ORIGINALSPLIT_FULL")
```

14 Create Majority-Only Variant

```
majonly.en <- list.files("EN_PDF_ENHANCED_FULL",
                         pattern ="(JUD|ADV|ORD|DEC)_[0-9]{2}_[A-Z-]+_00_EN_
    ENHANCED\\.pdf",
                         full.names = TRUE)

majonly.fr <- list.files("FR_PDF_ENHANCED_FULL",
                         pattern ="(JUD|ADV|ORD|DEC)_[0-9]{2}_[A-Z-]+_00_FR_
    ENHANCED\\.pdf",
                         full.names = TRUE)

file_copy(majonly.en,
          "EN_PDF_ENHANCED_MajorityOpinions")

file_copy(majonly.fr,
          "FR_PDF_ENHANCED_MajorityOpinions")
```

15 Read in TXT Files

15.1 Define Variable Names

```
names.variables <- c("court",
                     "series",
                     "seriesno",
                     "shortname",
                     "applicant",
                     "respondent",
                     "date",
                     "doctype",
                     "collision",
                     "stage",
                     "opinion",
                     "language")
```

15.2 TESSERACT Variants

```
data.tesseract.en <- readtext("EN_TXT_TESSERACT_FULL/*.txt",
                        docvarsfrom = "filenames",
                        docvarnames = names.variables,
                        dvsep = "_",
                        encoding = "UTF-8")


data.tesseract.fr <- readtext("FR_TXT_TESSERACT_FULL/*.txt",
                        docvarsfrom = "filenames",
                        docvarnames = names.variables,
                        dvsep = "_",
                        encoding = "UTF-8")
```

15.3 EXTRACTED Variants

```
data.extracted.en <- readtext("EN_TXT_EXTRACTED_FULL/*.txt",
                              docvarsfrom = "filenames",
                              docvarnames = names.variables,
                              dvsep = "_",
                              encoding = "UTF-8")


data.extracted.fr <- readtext("FR_TXT_EXTRACTED_FULL/*.txt",
                              docvarsfrom = "filenames",
                              docvarnames = names.variables,
                              dvsep = "_",
                              encoding = "UTF-8")
```

15.4 Convert to Data Table

```
setDT(data.tesseract.en)
setDT(data.tesseract.fr)
setDT(data.extracted.en)
setDT(data.extracted.fr)
```

16 Clean Texts

16.1 Remove Hyphenation across Linebreaks

Hyphenation across linebreaks is a serious issue for longer texts. Such hyphenated words are often not recognized as a single token by standard tokenization. The result is two unique and non-expressive tokens instead of a single, expressive token. This section removes these hyphenations.

16.1.1 Show Function: f.hyphen.remove

```
print(f.hyphen.remove)
```

```
## function(text){
##     ## Examples: Ham-\nburg, Mei-\n   nungsäußerung
##     text.out <- gsub("([a-zöäüß])-[[:blank:]]*\n[[:blank:]]*([a-zöäüß])",
##                      "\\1\\2",
##                      text)
##     ## Examples: SARS-CoV-\n2
##     text.out <- gsub("([a-zA-ZöäüÖÄÜß])-[[:blank:]]*\n[[:blank:]]*([A-Z0-9ÖÄÜß
    ])",
##                      "\\1-\\2",
##                      text.out)
##     ## Example: hat-    2\nte, Unsterb-    6\nliche
##     text.out <- gsub("([a-zöäüß])-[[:blank:]]*[0-9]+[[:blank:]]*\n[[:blank:]]*
    ([a-zöäüß])",
##                      "\\1\\2",
##                      text.out)
##
##     ## Example: hat-  \n  2 te, Unsterb-  \n  6 liche
##     text.out <- gsub("([a-zöäüß])-[[:space:]]*[0-9]+[[:blank:]]*([a-zöäüß])",
##                      "\\1\\2",
##                      text.out)
##
##     return(text.out)
## }
```

16.1.2 Execute Function

```
data.tesseract.en[, text := lapply(.(text), f.hyphen.remove)]
data.tesseract.fr[, text := lapply(.(text), f.hyphen.remove)]

data.extracted.en[, text := lapply(.(text), f.hyphen.remove)]
data.extracted.fr[, text := lapply(.(text), f.hyphen.remove)]
```

16.2 Replace Special Characters

This section replaces special characters with their closest equivalents in the Latin alphabet, as some R functions have difficulties processing the originals. These characters usually occur due to OCR mistakes.

16.2.1 Show Function: f.special.replace

```
print(f.special.replace)
```

```
## function(text){
##     text.out <- gsub("ff",
##                      "ff",
##                      text)
##
##     text.out <- gsub("fi",
##                      "fi",
##                      text.out)
##
##
##     text.out <- gsub("fl",
##                      "fl",
##                      text.out)
##
##     return(text.out)
## }
```

16.2.2 Execute Function

```
data.tesseract.en[, text := lapply(.(text), f.special.replace)]
data.tesseract.fr[, text := lapply(.(text), f.special.replace)]

data.extracted.en[, text := lapply(.(text), f.special.replace)]
data.extracted.fr[, text := lapply(.(text), f.special.replace)]
```

17 OCR Quality Control Module

This module measures the quality of the new Tesseract-generated OCR text against the OCR text provided by the ICJ, which was extracted from the original documents.

17.1 Create Corpora

```
corpus.en.b <- corpus(data.tesseract.en)
corpus.en.e <- corpus(data.extracted.en)

corpus.fr.b <- corpus(data.tesseract.fr)
corpus.fr.e <- corpus(data.extracted.fr)
```

17.2 Show Function: f.token.processor

```
print(f.token.processor)
```

```
## function(corpus){
##     tokens <- tokens(corpus,
##                      remove_numbers = TRUE,
##                      remove_punct = TRUE,
##                      remove_symbols = TRUE,
##                      remove_separators = TRUE)
##     tokens <- tokens_tolower(tokens)
##     tokens <- tokens_remove(tokens,
##                             pattern = c(stopwords("english"),
##                                         stopwords("french")))
##     return(tokens)
##     }
```

17.3 Tokenize

```
quanteda_options(tokens_locale = "en") # Set Locale for Tokenization

tokens.en.b <- f.token.processor(corpus.en.b)
tokens.en.e <- f.token.processor(corpus.en.e)


quanteda_options(tokens_locale = "fr") # Set Locale for Tokenization

tokens.fr.b <- f.token.processor(corpus.fr.b)
tokens.fr.e <- f.token.processor(corpus.fr.e)
```

17.4 Create Document-Feature-Matrices

```
dfm.en.b <- dfm(tokens.en.b)
dfm.en.e <- dfm(tokens.en.e)

dfm.fr.b <- dfm(tokens.fr.b)
dfm.fr.e <- dfm(tokens.fr.e)
```

17.5 Number of Features TESSERACT

17.5.1 English

```
nfeat(dfm.en.b)
```

```
## [1] 20428
```

17.5.2 French

```
nfeat(dfm.fr.b)
```

```
## [1] 27029
```

17.6 Number of Features EXTRACTED

17.6.1 English

```
nfeat(dfm.en.e)
```

```
## [1] 30228
```

17.6.2 French

```
nfeat(dfm.fr.e)
```

```
## [1] 30689
```

17.7 Features Reduction

Note: This is the number of features which have been saved by using advanced OCR in comparison to the OCR used by the ICJ.

17.7.1 English

Absolute Reduction

```
nfeat(dfm.en.e)- nfeat(dfm.en.b)
```

```
## [1] 9800
```

Relative Reduction in Percent

```
(1 - (nfeat(dfm.en.b) / nfeat(dfm.en.e))) * 100
```

```
## [1] 32.42027
```

17.7.2 French

Absolute Reduction

```
nfeat(dfm.fr.e)- nfeat(dfm.fr.b)
```

```
## [1] 3660
```

Relative Reduction in Percent

```
(1 - (nfeat(dfm.fr.b) / nfeat(dfm.fr.e))) * 100
```

```
## [1] 11.9261
```

18 Language Purity Module

This module automatically analyzes the n-gram patterns of each document with **textcat** to detect the most likely language. Only English and French are considered. This is to ensure maximum monolinguality of documents, which is an advantage in Natural Language Processing.

18.1 Limit Detection to English and French

```
lang.profiles <- TC_byte_profiles[names(TC_byte_profiles) %in% c("english",
                                                                 "french")]
```

18.2 Automatic Language Detection

```
data.tesseract.en$textcat <- textcat(data.tesseract.en$text,
                                     p = lang.profiles)

data.tesseract.fr$textcat <- textcat(data.tesseract.fr$text,
                                     p = lang.profiles)
```

18.3 Detected Languages

18.3.1 Should only read 'english'

```
unique(data.tesseract.en$textcat)
```

```
## [1] "english"
```

18.3.2 Should only read 'french'

```
unique(data.tesseract.fr$textcat)
```

```
## [1] "french"
```

18.4 Show Mismatches

Print names of files which failed to match the language specified in metadata.

```
langtest.fail.en <- data.tesseract.en[textcat != "english", .(doc_id, textcat)]
print(langtest.fail.en)
```

```
## Empty data.table (0 rows and 2 cols): doc_id,textcat
```

```
langtest.fail.fr <- data.tesseract.fr[textcat != "french", .(doc_id, textcat)]
print(langtest.fail.fr)
```

```
## Empty data.table (0 rows and 2 cols): doc_id,textcat
```

18.5 Final Note: Human Review of Mismatches

All documents flagged by textcat were reviewed and appropriate remedies devised. Documents falsely flagged as "catalan" were correctly labelled when possible languages were limited to English and French. Some documents received customized split instructions after being flagged.

19 Add and Delete Variables

19.1 Delete Textcat Classifications

```
data.tesseract.en$textcat <- NULL
data.tesseract.fr$textcat <- NULL
```

19.2 Add Variable "year"

```
data.tesseract.en$year <- year(data.tesseract.en$date)
data.tesseract.fr$year <- year(data.tesseract.fr$date)
```

19.3 Add Variable "minority"

"0" indicates a majority opinion, "1" a minority opinion.

```
data.tesseract.en$minority <- (data.tesseract.en$opinion != 0) * 1
data.tesseract.fr$minority <- (data.tesseract.fr$opinion != 0) * 1
```

19.4 Add Variable "fullname"

19.4.1 Read Hand Coded Data

```
casenames <- fread("data/CD-PCIJ_Source_Filenames-FullNames-SplitInstructions.csv
    ",
                   header = TRUE)
```

19.4.2 Create Variable

```
data.tesseract.en$fullname <- casenames$casename[match(data.tesseract.en$doc_id,
                                                       gsub("(BI|FR|EN)\\.pdf",
                                                            "EN\\.txt",
                                                            casenames$newname))]

data.tesseract.fr$fullname <- casenames$casename[match(data.tesseract.fr$doc_id,
                                                       gsub("(BI|FR|EN)\\.pdf",
                                                            "FR\\.txt",
                                                            casenames$newname))]
```

19.5 Add Variable "caseno"

The "caseno" variable is constructed by joining the "series" and "seriesno" variables, e.g. "AB" and "41" become "AB41." This is intended to be used as a unique case identifier similar to the "caseno" variable in the CD-PCIJ, though there are limitations to this approach, as quite a few cases span multiple combined numbers.

```
data.tesseract.en$caseno <- paste0(data.tesseract.en$series,
                                   data.tesseract.en$seriesno)

data.tesseract.fr$caseno <- paste0(data.tesseract.fr$series,
                                   data.tesseract.fr$seriesno)
```

19.6 Add Variable "applicant_region"

19.6.1 Read Hand Coded Data

```
countrycodes <- fread("data/CD-PCIJ_Source_CountryCodes.csv")
```

19.6.2 Merge Regions for English Version

```
applicant_region <- data.tesseract.en$applicant

applicant_region <- gsub("LNC",
                         "NA",
                         applicant_region)

applicant_region <- gsub("-",
                         "|",
                         applicant_region)


applicant_region <- mgsub(applicant_region,
                          countrycodes$ISO3,
                          countrycodes$region)

data.tesseract.en$applicant_region <- applicant_region
```

19.6.3 Merge Regions for French Version

```
applicant_region <- data.tesseract.fr$applicant

applicant_region <- gsub("LNC",
                         "NA",
                         applicant_region)

applicant_region <- gsub("-",
```

```
                          "|",
                          applicant_region)


applicant_region <- mgsub(applicant_region,
                          countrycodes$ISO3,
                          countrycodes$region)

data.tesseract.fr$applicant_region <- applicant_region
```

19.7 Add Variable "respondent_region"

19.7.1 Read Hand Coded Data

```
countrycodes <- fread("data/CD-PCIJ_Source_CountryCodes.csv")
```

19.7.2 Merge Regions for English Version

```
respondent_region <- data.tesseract.en$respondent

respondent_region <- gsub("-",
                          "|",
                          respondent_region)

respondent_region <- mgsub(respondent_region,
                           countrycodes$ISO3,
                           countrycodes$region)

data.tesseract.en$respondent_region <- respondent_region
```

19.7.3 Merge Regions for French Version

```
respondent_region <- data.tesseract.fr$respondent

respondent_region <- gsub("-",
                          "|",
                          respondent_region)

respondent_region <- mgsub(respondent_region,
                           countrycodes$ISO3,
                           countrycodes$region)

data.tesseract.fr$respondent_region <- respondent_region
```

19.8 Add Variable "applicant_subregion"

19.8.1 Read Hand Coded Data

```
countrycodes <- fread("data/CD-PCIJ_Source_CountryCodes.csv")
```

19.8.2 Merge Subregions for English Version

```
applicant_subregion <- data.tesseract.en$applicant

applicant_subregion <- gsub("LNC",
                            "NA",
                            applicant_subregion)

applicant_subregion <- gsub("-",
                            "|",
                            applicant_subregion)


applicant_subregion <- mgsub(applicant_subregion,
                             countrycodes$ISO3,
                             countrycodes$subregion)

data.tesseract.en$applicant_subregion <- applicant_subregion
```

19.8.3 Merge Subregions for French Version

```
applicant_subregion <- data.tesseract.fr$applicant

applicant_subregion <- gsub("LNC",
                            "NA",
                            applicant_subregion)

applicant_subregion <- gsub("-",
                            "|",
                            applicant_subregion)


applicant_subregion <- mgsub(applicant_subregion,
                             countrycodes$ISO3,
                             countrycodes$subregion)

data.tesseract.fr$applicant_subregion <- applicant_subregion
```

19.9 Add Variable "respondent_subregion"

19.9.1 Read Hand Coded Data

```
countrycodes <- fread("data/CD-PCIJ_Source_CountryCodes.csv")
```

19.9.2 Merge Subregions for English Version

```
respondent_subregion <- data.tesseract.en$respondent

respondent_subregion <- gsub("-",
                          "|",
                          respondent_subregion)

respondent_subregion <- mgsub(respondent_subregion,
                           countrycodes$ISO3,
                           countrycodes$subregion)

data.tesseract.en$respondent_subregion <- respondent_subregion
```

19.9.3 Merge Subregions for French Version

```
respondent_subregion <- data.tesseract.fr$respondent

respondent_subregion <- gsub("-",
                             "|",
                             respondent_subregion)

respondent_subregion <- mgsub(respondent_subregion,
                              countrycodes$ISO3,
                              countrycodes$subregion)

data.tesseract.fr$respondent_subregion <- respondent_subregion
```

19.10 Add Variable "doi_concept"

```
data.tesseract.en$doi_concept <- rep(doi.concept,
                                     data.tesseract.en[,.N])

data.tesseract.fr$doi_concept <- rep(doi.concept,
                                     data.tesseract.fr[,.N])
```

19.11 Add Variable "doi_version"

```
data.tesseract.en$doi_version <- rep(doi.version,
                                     data.tesseract.en[,.N])

data.tesseract.fr$doi_version <- rep(doi.version,
                                     data.tesseract.fr[,.N])
```

19.12 Add Variable "version"

```
data.tesseract.en$version <- as.character(rep(version,
                                              data.tesseract.en[,.N]))

data.tesseract.fr$version <- as.character(rep(version,
                                              data.tesseract.fr[,.N]))
```

19.13 Add Variable "license"

```
data.tesseract.en$license <- as.character(rep(license,
                                              data.tesseract.en[,.N]))

data.tesseract.fr$license <- as.character(rep(license,
                                              data.tesseract.fr[,.N]))
```

20 Frequency Tables

Frequency tables are a very useful tool for checking the plausibility of categorical variables and detecting anomalies in the data. This section will calculate frequency tables for all variables of interest.

20.1 Show Function: f.fast.freqtable

```
print(f.fast.freqtable)
```

function(x, varlist = names(x), sumrow = TRUE, output.list = TRUE, output.kable = FALSE, output.csv = FALSE, outputdir = "./," prefix = "", align ="r"){

```
## Begin List
freqtable.list <- vector("list", length(varlist))

## Calculate Frequency Table
for (i in seq_along(varlist)){

    varname <- varlist[i]

    freqtable <- x[, .N, keyby=c(paste0(varname))]

    freqtable[, c("exactpercent",
                  "roundedpercent",
                  "cumulpercent") := {
                      exactpercent  <-  N/sum(N)*100
                      roundedpercent <- round(exactpercent, 2)
                      cumulpercent <- round(cumsum(exactpercent), 2)
                      list(exactpercent,
                           roundedpercent,
                           cumulpercent)}]

    ## Calculate Summary Row
    if (sumrow == TRUE){
        colsums <-  cbind("Total",
                          freqtable[, lapply(.SD, function(x){round(sum(x))}),
                                    .SDcols = c("N",
                                                "exactpercent",
                                                "roundedpercent")
                                    ], round(max(freqtable$cumulpercent)))

        colnames(colsums)[c(1,5)] <- c(varname, "cumulpercent")
        freqtable <- rbind(freqtable, colsums)
    }

    ## Add Frequency Table to List
    freqtable.list[[i]] <- freqtable

    ## Write CSV
    if (output.csv == TRUE){
```

```
        fwrite(freqtable,
               paste0(outputdir,
                      prefix,
                      varname,
                      ".csv"),
               na = "NA")

    }

    ## Output Kable
    if (output.kable == TRUE){

        cat("\n-----------------------------------------------\n")
        cat(paste0("Frequency Table for Variable:   ", varname, "\n"))
        cat("-----------------------------------------------\n")
        cat(paste0("\n ",
                   x[, .N, keyby=c(paste0(varname))][,.N],
                   " unique value(s) detected.\n\n"))


        print(kable(freqtable,
                    format = "latex",
                    align = align,
                    booktabs = TRUE,
                    longtable = TRUE) %>% kable_styling(latex_options = "repeat_
    header"))
    }
}

## Return List of Frequency Tables
if (output.list == TRUE){
    return(freqtable.list)
}
```

}

20.2 English Corpus

20.2.1 Variables to Ignore

```
print(freq.var.ignore)
```

```
## [1] "date"   "doc_id" "text"
```

20.2.2 Variables to Analyze

```
varlist <- names(data.tesseract.en)

varlist <- setdiff(varlist,
```

```
                    freq.var.ignore)

print(varlist)
```

```
##  [1] "court"                "series"               "seriesno"
##  [4] "shortname"            "applicant"            "respondent"
##  [7] "doctype"              "collision"            "stage"
## [10] "opinion"              "language"             "year"
## [13] "minority"             "fullname"             "caseno"
## [16] "applicant_region"     "respondent_region"    "applicant_subregion"
## [19] "respondent_subregion" "doi_concept"          "doi_version"
## [22] "version"              "license"
```

20.2.3 Construct Frequency Tables

```
prefix <- paste0(datashort,
                 "_EN_01_FrequencyTable_var-")
```

```
f.fast.freqtable(data.tesseract.en,
                 varlist = varlist,
                 sumrow = TRUE,
                 output.list = FALSE,
                 output.kable = TRUE,
                 output.csv = TRUE,
                 outputdir = outputdir,
                 prefix = prefix,
                 align = c("p{5cm}",
                           rep("r", 4)))
```

Frequency Table for Variable: court

1 unique value(s) detected.

court	N	exactpercent	roundedpercent	cumulpercent
PCIJ	259	100	100	100
Total	259	100	100	100

Frequency Table for Variable: series

3 unique value(s) detected.

series	N	exactpercent	roundedpercent	cumulpercent
A	87	33.59073	33.59	33.59
AB	133	51.35135	51.35	84.94
B	39	15.05792	15.06	100.00
Total	259	100.00000	100.00	100.00

Frequency Table for Variable: seriesno

65 unique value(s) detected.

seriesno	N	exactpercent	roundedpercent	cumulpercent
1	8	3.0888031	3.09	3.09
2	8	3.0888031	3.09	6.18
3	4	1.5444015	1.54	7.72
4	3	1.1583012	1.16	8.88
5	3	1.1583012	1.16	10.04
6	4	1.5444015	1.54	11.58
7	7	2.7027027	2.70	14.29
8	5	1.9305019	1.93	16.22
9	3	1.1583012	1.16	17.37
10	10	3.8610039	3.86	21.24
11	6	2.3166023	2.32	23.55
12	4	1.5444015	1.54	25.10
13	4	1.5444015	1.54	26.64
14	6	2.3166023	2.32	28.96
15	10	3.8610039	3.86	32.82
16	3	1.1583012	1.16	33.98
17	10	3.8610039	3.86	37.84
18	4	1.5444015	1.54	39.38
19	2	0.7722008	0.77	40.15
20	5	1.9305019	1.93	42.08

(continued)

seriesno	N	exactpercent	roundedpercent	cumulpercent
21	4	1.5444015	1.54	43.63
22	6	2.3166023	2.32	45.95
23	4	1.5444015	1.54	47.49
24	3	1.1583012	1.16	48.65
40	4	1.5444015	1.54	50.19
41	7	2.7027027	2.70	52.90
42	1	0.3861004	0.39	53.28
43	3	1.1583012	1.16	54.44
44	4	1.5444015	1.54	55.98
45	2	0.7722008	0.77	56.76
46	6	2.3166023	2.32	59.07
47	3	1.1583012	1.16	60.23
48	2	0.7722008	0.77	61.00
49	4	1.5444015	1.54	62.55
50	3	1.1583012	1.16	63.71
51	1	0.3861004	0.39	64.09
52	1	0.3861004	0.39	64.48
53	4	1.5444015	1.54	66.02
54	1	0.3861004	0.39	66.41
55	1	0.3861004	0.39	66.80
56	1	0.3861004	0.39	67.18
57	1	0.3861004	0.39	67.57
58	5	1.9305019	1.93	69.50
59	1	0.3861004	0.39	69.88
60	1	0.3861004	0.39	70.27
61	3	1.1583012	1.16	71.43
62	4	1.5444015	1.54	72.97
63	7	2.7027027	2.70	75.68
64	3	1.1583012	1.16	76.83

(continued)

seriesno	N	exactpercent	roundedpercent	cumulpercent
65	6	2.3166023	2.32	79.15
66	1	0.3861004	0.39	79.54
67	1	0.3861004	0.39	79.92
68	7	2.7027027	2.70	82.63
69	1	0.3861004	0.39	83.01
70	7	2.7027027	2.70	85.71
71	6	2.3166023	2.32	88.03
72	3	1.1583012	1.16	89.19
73	1	0.3861004	0.39	89.58
74	4	1.5444015	1.54	91.12
75	1	0.3861004	0.39	91.51
76	6	2.3166023	2.32	93.82
77	10	3.8610039	3.86	97.68
78	4	1.5444015	1.54	99.23
79	1	0.3861004	0.39	99.61
80	1	0.3861004	0.39	100.00
Total	259	100.0000000	100.00	100.00

Frequency Table for Variable: shortname

63 unique value(s) detected.

shortname	N	exactpercent	roundedpercent	cumulpercent
AcquisitionPolishNationality	2	0.7722008	0.77	0.77
Borchgrave	4	1.5444015	1.54	2.32
BrazilianLoans	4	1.5444015	1.54	3.86
Castellorizo	1	0.3861004	0.39	4.25
ChorzowFactory	2	0.7722008	0.77	5.02
ChorzowFactory-Indemnities	11	4.2471042	4.25	9.27

(continued)

shortname	N	exactpercent	roundedpercent	cumulpercent
ChorzowFactory-Interpretation	2	0.7722008	0.77	10.04
CustomsRegime	7	2.7027027	2.70	12.74
DanubeCommission	5	1.9305019	1.93	14.67
DanzigCourts	3	1.1583012	1.16	15.83
DanzigILO	3	1.1583012	1.16	16.99
DanzigLegislativeDecrees	6	2.3166023	2.32	19.31
EasternCarelia	2	0.7722008	0.77	20.08
EasternGreenland	7	2.7027027	2.70	22.78
ElectricityCompanySofiaBulgaria	12	4.6332046	4.63	27.41
ExchangeGreekTurkishPopulations	2	0.7722008	0.77	28.19
FreeZonesUpperSavoyGex	12	4.6332046	4.63	32.82
FreeZonesUpperSavoyGex-SecondPhase	3	1.1583012	1.16	33.98
GermanInterestsUpperSilesia	8	3.0888031	3.09	37.07
GermanMinoritySchools	4	1.5444015	1.54	38.61
GermanSettlers	1	0.3861004	0.39	39.00
GrecoBulgarianAgreement	2	0.7722008	0.77	39.77
GrecoBulgarianCommunities	3	1.1583012	1.16	40.93
GrecoTurkishAgreement	2	0.7722008	0.77	41.70
HungaroCzechoslovakMixedTribunal	1	0.3861004	0.39	42.08
ILOCompetenceEmployer	2	0.7722008	0.77	42.86
ILOCompetenceMethodsAgriculture	2	0.7722008	0.77	43.63
ILOCompetencePersonsAgriculture	2	0.7722008	0.77	44.40
InterpretationMemelStatute	7	2.7027027	2.70	47.10
InterpretationNo3	1	0.3861004	0.39	47.49
Jaworzina	1	0.3861004	0.39	47.88
Lighthouses	4	1.5444015	1.54	49.42
LighthousesCreteSamos	6	2.3166023	2.32	51.74
Losinger	2	0.7722008	0.77	52.51

(continued)

shortname	N	exactpercent	roundedpercent	cumulpercent
Lotus	8	3.0888031	3.09	55.60
MavrommatisJerusalem	1	0.3861004	0.39	55.98
MavrommatisJerusalem-Readaptation	4	1.5444015	1.54	57.53
MavrommatisPalestine	6	2.3166023	2.32	59.85
Meuse	7	2.7027027	2.70	62.55
MinoritySchoolsAlbania	3	1.1583012	1.16	63.71
MinoritySchoolsUpperSilesia	7	2.7027027	2.70	66.41
MonasterySaintNaoum	1	0.3861004	0.39	66.80
NationalityDecrees	2	0.7722008	0.77	67.57
Neuilly	2	0.7722008	0.77	68.34
OderCommission	4	1.5444015	1.54	69.88
OscarChinn	7	2.7027027	2.70	72.59
PajzsCsakyEsterhazy	8	3.0888031	3.09	75.68
PanevezysSaldutiskisRailway	7	2.7027027	2.70	78.38
Pazmany	3	1.1583012	1.16	79.54
PhosphatesMarocco	4	1.5444015	1.54	81.08
PolishAgrarianReform	6	2.3166023	2.32	83.40
PolishWarVessels	3	1.1583012	1.16	84.56
PostalServiceDanzig	2	0.7722008	0.77	85.33
PrinceVonPless	4	1.5444015	1.54	86.87
RailwayTraffic	1	0.3861004	0.39	87.26
SerbianLoans	5	1.9305019	1.93	89.19
SinoBelgianTreaty	7	2.7027027	2.70	91.89
SocieteCommercialeBelgique	4	1.5444015	1.54	93.44
TreatmentPolishNationals	4	1.5444015	1.54	94.98
TreatyLausanne	2	0.7722008	0.77	95.75
Wimbledon	6	2.3166023	2.32	98.07
WomenNightWork	3	1.1583012	1.16	99.23

(continued)

shortname	N	exactpercent	roundedpercent	cumulpercent
WorkersDelegateILO	2	0.7722008	0.77	100.00
Total	259	100.0000000	100.00	100.00

Frequency Table for Variable: applicant

17 unique value(s) detected.

applicant	N	exactpercent	roundedpercent	cumulpercent
BEL	27	10.4247104	10.42	10.42
BGR	3	1.1583012	1.16	11.58
CHE	2	0.7722008	0.77	12.36
CSK	4	1.5444015	1.54	13.90
DEU	44	16.9884170	16.99	30.89
DNK	7	2.7027027	2.70	33.59
EST	7	2.7027027	2.70	36.29
FRA	42	16.2162162	16.22	52.51
GBR	7	2.7027027	2.70	55.21
GBR-CSK-DNK-FRA-DEU	4	1.5444015	1.54	56.76
GBR-FRA-ITA-JPN	13	5.0193050	5.02	61.78
GRC	11	4.2471042	4.25	66.02
HUN	8	3.0888031	3.09	69.11
ITA	4	1.5444015	1.54	70.66
LNC	68	26.2548263	26.25	96.91
NLD	7	2.7027027	2.70	99.61
TUR	1	0.3861004	0.39	100.00
Total	259	100.0000000	100.00	100.00

Frequency Table for Variable: respondent

18 unique value(s) detected.

respondent	N	exactpercent	roundedpercent	cumulpercent
NA	68	26.2548263	26.25	26.25
BEL	14	5.4054054	5.41	31.66
BGR	12	4.6332046	4.63	36.29
BRA	4	1.5444015	1.54	37.84
CHE	15	5.7915058	5.79	43.63
CHN	7	2.7027027	2.70	46.33
DEU	6	2.3166023	2.32	48.65
ESP	4	1.5444015	1.54	50.19
FRA	4	1.5444015	1.54	51.74
GBR	11	4.2471042	4.25	55.98
GRC	17	6.5637066	6.56	62.55
HUN	4	1.5444015	1.54	64.09
ITA	1	0.3861004	0.39	64.48
LTU	14	5.4054054	5.41	69.88
NOR	7	2.7027027	2.70	72.59
POL	48	18.5328185	18.53	91.12
TUR	8	3.0888031	3.09	94.21
YUG	15	5.7915058	5.79	100.00
Total	259	100.0000000	100.00	100.00

Frequency Table for Variable: doctype

6 unique value(s) detected.

doctype	N	exactpercent	roundedpercent	cumulpercent
ADV	44	16.9884170	16.99	16.99
ANX	41	15.8301158	15.83	32.82
APP	9	3.4749035	3.47	36.29
DEC	1	0.3861004	0.39	36.68

(continued)

doctype	N	exactpercent	roundedpercent	cumulpercent
JUD	113	43.6293436	43.63	80.31
ORD	51	19.6911197	19.69	100.00
Total	259	100.0000000	100.00	100.00

Frequency Table for Variable: collision

3 unique value(s) detected.

collision	N	exactpercent	roundedpercent	cumulpercent
1	251	96.9111969	96.91	96.91
2	6	2.3166023	2.32	99.23
3	2	0.7722008	0.77	100.00
Total	259	100.0000000	100.00	100.00

Frequency Table for Variable: stage

17 unique value(s) detected.

stage	N	exactpercent	roundedpercent	cumulpercent
NA	94	36.2934363	36.29	36.29
AJ	3	1.1583012	1.16	37.45
DH	2	0.7722008	0.77	38.22
DI	9	3.4749035	3.47	41.70
EV	2	0.7722008	0.77	42.47
EV-SE	6	2.3166023	2.32	44.79
EX	2	0.7722008	0.77	45.56
IM	10	3.8610039	3.86	49.42
IN	1	0.3861004	0.39	49.81
JO	1	0.3861004	0.39	50.19

(continued)

stage	N	exactpercent	roundedpercent	cumulpercent
JO-TL	1	0.3861004	0.39	50.58
ME	83	32.0463320	32.05	82.63
PO	33	12.7413127	12.74	95.37
PR	1	0.3861004	0.39	95.75
SE	3	1.1583012	1.16	96.91
TL	7	2.7027027	2.70	99.61
TL-DH	1	0.3861004	0.39	100.00
Total	259	100.0000000	100.00	100.00

Frequency Table for Variable: opinion

9 unique value(s) detected.

opinion	N	exactpercent	roundedpercent	cumulpercent
NA	50	19.3050193	19.31	19.31
0	100	38.6100386	38.61	57.92
1	40	15.4440154	15.44	73.36
2	30	11.5830116	11.58	84.94
3	18	6.9498069	6.95	91.89
4	11	4.2471042	4.25	96.14
5	7	2.7027027	2.70	98.84
6	2	0.7722008	0.77	99.61
7	1	0.3861004	0.39	100.00
Total	259	100.0000000	100.00	100.00

Frequency Table for Variable: language

1 unique value(s) detected.

language	N	exactpercent	roundedpercent	cumulpercent
EN	259	100	100	100
Total	259	100	100	100

Frequency Table for Variable: year

19 unique value(s) detected.

year	N	exactpercent	roundedpercent	cumulpercent
1922	7	2.7027027	2.70	2.70
1923	13	5.0193050	5.02	7.72
1924	9	3.4749035	3.47	11.20
1925	11	4.2471042	4.25	15.44
1926	8	3.0888031	3.09	18.53
1927	26	10.0386100	10.04	28.57
1928	22	8.4942085	8.49	37.07
1929	21	8.1081081	8.11	45.17
1930	9	3.4749035	3.47	48.65
1931	16	6.1776062	6.18	54.83
1932	23	8.8803089	8.88	63.71
1933	20	7.7220077	7.72	71.43
1934	11	4.2471042	4.25	75.68
1935	9	3.4749035	3.47	79.15
1936	10	3.8610039	3.86	83.01
1937	16	6.1776062	6.18	89.19
1938	6	2.3166023	2.32	91.51
1939	21	8.1081081	8.11	99.61
1940	1	0.3861004	0.39	100.00
Total	259	100.0000000	100.00	100.00

Frequency Table for Variable: minority

3 unique value(s) detected.

minority	N	exactpercent	roundedpercent	cumulpercent
NA	50	19.30502	19.31	19.31
0	100	38.61004	38.61	57.92
1	109	42.08494	42.08	100.00
Total	259	100.00000	100.00	100.00

Frequency Table for Variable: fullname

78 unique value(s) detected.

fullname	N	exactpercent	roundedpercent	cumulpercent
Access to German Minority Schools in Upper Silesia	4	1.5444015	1.54	1.54
Access to, or Anchorage in, the port of Danzig, of Polish War Vessels	3	1.1583012	1.16	2.70
Acquisition of Polish Nationality	2	0.7722008	0.77	3.47
Administration of the Prince von Pless (Discontinuance)	1	0.3861004	0.39	3.86
Administration of the Prince von Pless (Interim Measures of Protection)	1	0.3861004	0.39	4.25
Administration of the Prince von Pless (Preliminary Objections)	1	0.3861004	0.39	4.63
Administration of the Prince von Pless (Prorogation)	1	0.3861004	0.39	5.02
Appeal from a Judgment of the Hungaro-Czecoslovak Mixed Arbitral Tribunal (The Peter Pázmány University v. The State of Czechoslovakia)	3	1.1583012	1.16	6.18

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Appeals from Certain Judgments of the Hungaro-Czechoslovak Mixed Arbitral Tribunal	1	0.3861004	0.39	6.56
Borchgrave (Discontinuance)	1	0.3861004	0.39	6.95
Borchgrave (Preliminary Objections)	3	1.1583012	1.16	8.11
Certain German Interests in Polish Upper Silesia (Merits)	5	1.9305019	1.93	10.04
Certain German Interests in Polish Upper Silesia (Preliminary Objections)	3	1.1583012	1.16	11.20
Competence of the ILO in regard to International Regulation of the Conditions of the Labour of Persons Employed in Agriculture	2	0.7722008	0.77	11.97
Competence of the ILO to Examine Proposal for the Organization and Development of the Methods of Agricultural Production	2	0.7722008	0.77	12.74
Competence of the ILO to Regulate, Incidentally, the Personal Work of the Employer	2	0.7722008	0.77	13.51
Consistency of Certain Danzig Legislative Decrees with the Constitution of the Free City	6	2.3166023	2.32	15.83
Customs Regime between Germany and Austria (Protocol of March 19th, 1931)	7	2.7027027	2.70	18.53
Delimitation of the Territorial Waters between the Island of Castellorizo and the Coasts of Anatolia	1	0.3861004	0.39	18.92
Denunciation of the Treaty of November 2nd, 1865, between China and Belgium	7	2.7027027	2.70	21.62

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Designation of the Workers' Delegate for the Netherlands at the Third Session of the International Labour Conference	2	0.7722008	0.77	22.39
Diversion of Water from the Meuse	7	2.7027027	2.70	25.10
Electricity Company of Sofia and Bulgaria (Date of Hearing)	1	0.3861004	0.39	25.48
Electricity Company of Sofia and Bulgaria (Interim Measures of Protection)	1	0.3861004	0.39	25.87
Electricity Company of Sofia and Bulgaria (Preliminary Objections)	10	3.8610039	3.86	29.73
Exchange of Greek and Turkish Populations (Lausanne Convention VI, January 30th, 1923, Article 2)	2	0.7722008	0.77	30.50
Factory at Chorzów (Claim for Indemnity) (Jurisdiction)	2	0.7722008	0.77	31.27
Factory at Chorzów (Claim for Indemnity) (Merits)	7	2.7027027	2.70	33.98
Factory at Chorzów (Indemnities)	4	1.5444015	1.54	35.52
Free City of Danzig and International Labour Organization	3	1.1583012	1.16	36.68
Free Zones of Upper Savoy and the District of Gex	12	4.6332046	4.63	41.31
Free Zones of Upper Savoy and the District of Gex (Second Phase)	3	1.1583012	1.16	42.47
Greco-Bulgarian Communities	3	1.1583012	1.16	43.63
Interpretation of Article 3, Paragraph 2, of the Treaty of Lausanne (Frontier between Turkey and Iraq)	2	0.7722008	0.77	44.40

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Interpretation of Judgment No.3	1	0.3861004	0.39	44.79
Interpretation of Judgments Nos. 7 and 8 (The Chorzów Factory)	2	0.7722008	0.77	45.56
Interpretation of the Convention of 1919 concerning Employment of Women during the Night	3	1.1583012	1.16	46.72
Interpretation of the Greco-Bulgarian Agreement of December 9th, 1927 (Caphandaris-Molloff Agreement)	2	0.7722008	0.77	47.49
Interpretation of the Greco-Turkish Agreement of December 1st, 1926 (Final Protocol, Article IV)	2	0.7722008	0.77	48.26
Interpretation of the Statute of the Memel Territory (Merits)	4	1.5444015	1.54	49.81
Interpretation of the Statute of the Memel Territory (Preliminary Objection)	3	1.1583012	1.16	50.97
Jaworzina (Polish-Czechoslovakian Frontier)	1	0.3861004	0.39	51.35
Jurisdiction of the Courts of Danzig	2	0.7722008	0.77	52.12
Jurisdiction of the Courts of Danzig (Pecuniary Claims of Danzig Railway Officials who have passed into the Polish Service, against the Polish Railways Administration)	1	0.3861004	0.39	52.51
Jurisdiction of the European Commission of the Danube between Galatz and Braila	5	1.9305019	1.93	54.44
Legal Status of Eastern Greenland	4	1.5444015	1.54	55.98

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Legal Status of the South-Eastern Territory of Greenland	3	1.1583012	1.16	57.14
Lighthouses Case between France and Greece	4	1.5444015	1.54	58.69
Lighthouses in Crete and Samos	6	2.3166023	2.32	61.00
Losinger & Co Case (Discontinuance)	1	0.3861004	0.39	61.39
Losinger & Co Case (Preliminary Objection)	1	0.3861004	0.39	61.78
Mavrommatis Jerusalem Concessions	1	0.3861004	0.39	62.16
Mavrommatis Palestine Concessions	6	2.3166023	2.32	64.48
Minority Schools in Albania	3	1.1583012	1.16	65.64
Monastery of Saint-Naoum (Albanian Frontier)	1	0.3861004	0.39	66.02
Nationality Decrees Issued in Tunis and Morocco	2	0.7722008	0.77	66.80
Oscar Chinn	7	2.7027027	2.70	69.50
Pajzs, Czáky, Esterházy (Merits)	7	2.7027027	2.70	72.20
Pajzs, Czáky, Esterházy (Preliminary Objection)	1	0.3861004	0.39	72.59
Panevezys-Saldutiskis Railway (Merits)	6	2.3166023	2.32	74.90
Panevezys-Saldutiskis Railway (Preliminary Objections)	1	0.3861004	0.39	75.29
Payment in Gold of the Brazilian Federal Loans Issued in France	4	1.5444015	1.54	76.83
Payment of Various Serbian Loans Issued in France	5	1.9305019	1.93	78.76
Phosphates in Morocco	4	1.5444015	1.54	80.31

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Polish Agrarian Reform and the German Minority (Discontinuance)	1	0.3861004	0.39	80.69
Polish Agrarian Reform and the German Minority (Interim Measures of Protection)	5	1.9305019	1.93	82.63
Polish Postal Service in Danzig	2	0.7722008	0.77	83.40
Railway Traffic between Lithuania and Poland (Railway Sector Landwarów-Kaisiadorys)	1	0.3861004	0.39	83.78
Readaptation of the Mavrommatis Jerusalem Concessions (Jurisdiction)	4	1.5444015	1.54	85.33
Rights of Minorities in Upper Silesia (Minority Schools)	7	2.7027027	2.70	88.03
S.S. Lotus	8	3.0888031	3.09	91.12
S.S. Wimbledon	6	2.3166023	2.32	93.44
Settlers of German Origin in the Territory Ceded by Germany to Poland	1	0.3861004	0.39	93.82
Société Commerciale de Belgique	4	1.5444015	1.54	95.37
Status of Eastern Carelia	2	0.7722008	0.77	96.14
Territorial Jurisdiction of the International Commission of the River Oder	4	1.5444015	1.54	97.68
Treatment of Polish Nationals and Other Persons of Polish Origin or Speech in the Danzig Territory	4	1.5444015	1.54	99.23
Treaty of Neuilly, Article 179, Annex, Paragraph 4 (Interpretation)	2	0.7722008	0.77	100.00
Total	259	100.0000000	100.00	100.00

Frequency Table for Variable: caseno

83 unique value(s) detected.

caseno	N	exactpercent	roundedpercent	cumulpercent
A1	6	2.3166023	2.32	2.32
A10	8	3.0888031	3.09	5.41
A11	4	1.5444015	1.54	6.95
A12	2	0.7722008	0.77	7.72
A13	2	0.7722008	0.77	8.49
A14	1	0.3861004	0.39	8.88
A15	7	2.7027027	2.70	11.58
A16	1	0.3861004	0.39	11.97
A17	7	2.7027027	2.70	14.67
A18	1	0.3861004	0.39	15.06
A19	2	0.7722008	0.77	15.83
A2	6	2.3166023	2.32	18.15
A20	5	1.9305019	1.93	20.08
A21	4	1.5444015	1.54	21.62
A22	6	2.3166023	2.32	23.94
A23	4	1.5444015	1.54	25.48
A24	3	1.1583012	1.16	26.64
A3	2	0.7722008	0.77	27.41
A4	1	0.3861004	0.39	27.80
A5	1	0.3861004	0.39	28.19
A6	3	1.1583012	1.16	29.34
A7	5	1.9305019	1.93	31.27
A8	4	1.5444015	1.54	32.82
A9	2	0.7722008	0.77	33.59
AB40	4	1.5444015	1.54	35.14
AB41	7	2.7027027	2.70	37.84

(continued)

caseno	N	exactpercent	roundedpercent	cumulpercent
AB42	1	0.3861004	0.39	38.22
AB43	3	1.1583012	1.16	39.38
AB44	4	1.5444015	1.54	40.93
AB45	2	0.7722008	0.77	41.70
AB46	6	2.3166023	2.32	44.02
AB47	3	1.1583012	1.16	45.17
AB48	2	0.7722008	0.77	45.95
AB49	4	1.5444015	1.54	47.49
AB50	3	1.1583012	1.16	48.65
AB51	1	0.3861004	0.39	49.03
AB52	1	0.3861004	0.39	49.42
AB53	4	1.5444015	1.54	50.97
AB54	1	0.3861004	0.39	51.35
AB55	1	0.3861004	0.39	51.74
AB56	1	0.3861004	0.39	52.12
AB57	1	0.3861004	0.39	52.51
AB58	5	1.9305019	1.93	54.44
AB59	1	0.3861004	0.39	54.83
AB60	1	0.3861004	0.39	55.21
AB61	3	1.1583012	1.16	56.37
AB62	4	1.5444015	1.54	57.92
AB63	7	2.7027027	2.70	60.62
AB64	3	1.1583012	1.16	61.78
AB65	6	2.3166023	2.32	64.09
AB66	1	0.3861004	0.39	64.48
AB67	1	0.3861004	0.39	64.86
AB68	7	2.7027027	2.70	67.57
AB69	1	0.3861004	0.39	67.95
AB70	7	2.7027027	2.70	70.66

(continued)

caseno	N	exactpercent	roundedpercent	cumulpercent
AB71	6	2.3166023	2.32	72.97
AB72	3	1.1583012	1.16	74.13
AB73	1	0.3861004	0.39	74.52
AB74	4	1.5444015	1.54	76.06
AB75	1	0.3861004	0.39	76.45
AB76	6	2.3166023	2.32	78.76
AB77	10	3.8610039	3.86	82.63
AB78	4	1.5444015	1.54	84.17
AB79	1	0.3861004	0.39	84.56
AB80	1	0.3861004	0.39	84.94
B1	2	0.7722008	0.77	85.71
B10	2	0.7722008	0.77	86.49
B11	2	0.7722008	0.77	87.26
B12	2	0.7722008	0.77	88.03
B13	2	0.7722008	0.77	88.80
B14	5	1.9305019	1.93	90.73
B15	3	1.1583012	1.16	91.89
B16	2	0.7722008	0.77	92.66
B17	3	1.1583012	1.16	93.82
B18	3	1.1583012	1.16	94.98
B2	2	0.7722008	0.77	95.75
B3	2	0.7722008	0.77	96.53
B4	2	0.7722008	0.77	97.30
B5	2	0.7722008	0.77	98.07
B6	1	0.3861004	0.39	98.46
B7	2	0.7722008	0.77	99.23
B8	1	0.3861004	0.39	99.61
B9	1	0.3861004	0.39	100.00
Total	259	100.0000000	100.00	100.00

Frequency Table for Variable: applicant_region

5 unique value(s) detected.

applicant_region	N	exactpercent	roundedpercent	cumulpercent
Asia	1	0.3861004	0.39	0.39
Europe	173	66.7953668	66.80	67.18
Europe|Europe|Europe|Asia	13	5.0193050	5.02	72.20
Europe|Europe|Europe|Europe|Europe	4	1.5444015	1.54	73.75
NA	68	26.2548263	26.25	100.00
Total	259	100.0000000	100.00	100.00

Frequency Table for Variable: respondent_region

4 unique value(s) detected.

respondent_region	N	exactpercent	roundedpercent	cumulpercent
NA	68	26.254826	26.25	26.25
Americas	4	1.544401	1.54	27.80
Asia	15	5.791506	5.79	33.59
Europe	172	66.409266	66.41	100.00
Total	259	100.000000	100.00	100.00

Frequency Table for Variable: applicant_subregion

8 unique value(s) detected.

applicant_subregion	N	exactpercent	roundedpercent	cumulpercent
Eastern Europe	15	5.7915058	5.79	5.79
NA	68	26.2548263	26.25	32.05
Northern Europe	21	8.1081081	8.11	40.15

(continued)

applicant_subregion	N	exactpercent	roundedpercent	cumulpercent
Northern Europe|Eastern Europe|Northern Europe|Western Europe|Western Europe	4	1.5444015	1.54	41.70
Northern Europe|Western Europe|Southern Europe|Eastern Asia	13	5.0193050	5.02	46.72
Southern Europe	15	5.7915058	5.79	52.51
Western Asia	1	0.3861004	0.39	52.90
Western Europe	122	47.1042471	47.10	100.00
Total	259	100.0000000	100.00	100.00

Frequency Table for Variable: respondent_subregion

8 unique value(s) detected.

respondent_subregion	N	exactpercent	roundedpercent	cumulpercent
NA	68	26.254826	26.25	26.25
Eastern Asia	7	2.702703	2.70	28.96
Eastern Europe	64	24.710425	24.71	53.67
Latin America and the Caribbean	4	1.544401	1.54	55.21
Northern Europe	32	12.355212	12.36	67.57
Southern Europe	37	14.285714	14.29	81.85
Western Asia	8	3.088803	3.09	84.94
Western Europe	39	15.057915	15.06	100.00
Total	259	100.000000	100.00	100.00

Frequency Table for Variable: doi_concept

1 unique value(s) detected.

doi_concept	N	exactpercent	roundedpercent	cumulpercent
10.5281/zenodo.3840479	259	100	100	100
Total	259	100	100	100

Frequency Table for Variable: doi_version

1 unique value(s) detected.

doi_version	N	exactpercent	roundedpercent	cumulpercent
10.5281/zenodo.7051934	259	100	100	100
Total	259	100	100	100

Frequency Table for Variable: version

1 unique value(s) detected.

version	N	exactpercent	roundedpercent	cumulpercent
1.1.0	259	100	100	100
Total	259	100	100	100

Frequency Table for Variable: license

1 unique value(s) detected.

license	N	exactpercent	roundedpercent	cumulpercent
Creative Commons Zero 1.0 Universal	259	100	100	100
Total	259	100	100	100

20.3 French Corpus

20.3.1 Variables to Ignore

```
print(freq.var.ignore)
```

```
## [1] "date"   "doc_id" "text"
```

20.3.2 Variables to Analyze

```
varlist <- names(data.tesseract.en)

varlist <- setdiff(varlist,
                   freq.var.ignore)

print(varlist)
```

```
##  [1] "court"                "series"               "seriesno"
##  [4] "shortname"            "applicant"            "respondent"
##  [7] "doctype"              "collision"            "stage"
## [10] "opinion"              "language"             "year"
## [13] "minority"             "fullname"             "caseno"
## [16] "applicant_region"     "respondent_region"    "applicant_subregion"
## [19] "respondent_subregion" "doi_concept"          "doi_version"
## [22] "version"              "license"
```

20.3.3 Construct Frequency Tables

```
prefix <- paste0(datashort,
                 "_FR_01_FrequencyTable_var-")
```

```
f.fast.freqtable(data.tesseract.fr,
                 varlist = varlist,
                 sumrow = TRUE,
                 output.list = FALSE,
                 output.kable = TRUE,
                 output.csv = TRUE,
                 outputdir = outputdir,
                 prefix = prefix,
                 align = c("p{5cm}",
                           rep("r", 4)))
```

Frequency Table for Variable: court

1 unique value(s) detected.

court	N	exactpercent	roundedpercent	cumulpercent
PCIJ	261	100	100	100
Total	261	100	100	100

Frequency Table for Variable: series

3 unique value(s) detected.

series	N	exactpercent	roundedpercent	cumulpercent
A	87	33.33333	33.33	33.33
AB	134	51.34100	51.34	84.67
B	40	15.32567	15.33	100.00
Total	261	100.00000	100.00	100.00

Frequency Table for Variable: seriesno

65 unique value(s) detected.

seriesno	N	exactpercent	roundedpercent	cumulpercent
1	8	3.0651341	3.07	3.07
2	8	3.0651341	3.07	6.13
3	3	1.1494253	1.15	7.28
4	3	1.1494253	1.15	8.43
5	3	1.1494253	1.15	9.58
6	4	1.5325670	1.53	11.11
7	8	3.0651341	3.07	14.18
8	5	1.9157088	1.92	16.09
9	3	1.1494253	1.15	17.24

(continued)

seriesno	N	exactpercent	roundedpercent	cumulpercent
10	10	3.8314176	3.83	21.07
11	6	2.2988506	2.30	23.37
12	4	1.5325670	1.53	24.90
13	4	1.5325670	1.53	26.44
14	6	2.2988506	2.30	28.74
15	11	4.2145594	4.21	32.95
16	3	1.1494253	1.15	34.10
17	10	3.8314176	3.83	37.93
18	4	1.5325670	1.53	39.46
19	2	0.7662835	0.77	40.23
20	5	1.9157088	1.92	42.15
21	4	1.5325670	1.53	43.68
22	6	2.2988506	2.30	45.98
23	4	1.5325670	1.53	47.51
24	3	1.1494253	1.15	48.66
40	4	1.5325670	1.53	50.19
41	7	2.6819923	2.68	52.87
42	1	0.3831418	0.38	53.26
43	3	1.1494253	1.15	54.41
44	4	1.5325670	1.53	55.94
45	2	0.7662835	0.77	56.70
46	6	2.2988506	2.30	59.00
47	3	1.1494253	1.15	60.15
48	2	0.7662835	0.77	60.92
49	4	1.5325670	1.53	62.45
50	3	1.1494253	1.15	63.60
51	1	0.3831418	0.38	63.98
52	1	0.3831418	0.38	64.37
53	4	1.5325670	1.53	65.90

(continued)

seriesno	N	exactpercent	roundedpercent	cumulpercent
54	1	0.3831418	0.38	66.28
55	1	0.3831418	0.38	66.67
56	1	0.3831418	0.38	67.05
57	1	0.3831418	0.38	67.43
58	5	1.9157088	1.92	69.35
59	1	0.3831418	0.38	69.73
60	1	0.3831418	0.38	70.11
61	3	1.1494253	1.15	71.26
62	4	1.5325670	1.53	72.80
63	7	2.6819923	2.68	75.48
64	3	1.1494253	1.15	76.63
65	6	2.2988506	2.30	78.93
66	1	0.3831418	0.38	79.31
67	1	0.3831418	0.38	79.69
68	7	2.6819923	2.68	82.38
69	1	0.3831418	0.38	82.76
70	8	3.0651341	3.07	85.82
71	6	2.2988506	2.30	88.12
72	3	1.1494253	1.15	89.27
73	1	0.3831418	0.38	89.66
74	4	1.5325670	1.53	91.19
75	1	0.3831418	0.38	91.57
76	6	2.2988506	2.30	93.87
77	10	3.8314176	3.83	97.70
78	4	1.5325670	1.53	99.23
79	1	0.3831418	0.38	99.62
80	1	0.3831418	0.38	100.00
Total	261	100.0000000	100.00	100.00

Frequency Table for Variable: shortname

63 unique value(s) detected.

shortname	N	exactpercent	roundedpercent	cumulpercent
AcquisitionPolishNationality	2	0.7662835	0.77	0.77
Borchgrave	4	1.5325670	1.53	2.30
BrazilianLoans	4	1.5325670	1.53	3.83
Castellorizo	1	0.3831418	0.38	4.21
ChorzowFactory	2	0.7662835	0.77	4.98
ChorzowFactory-Indemnities	11	4.2145594	4.21	9.20
ChorzowFactory-Interpretation	2	0.7662835	0.77	9.96
CustomsRegime	7	2.6819923	2.68	12.64
DanubeCommission	5	1.9157088	1.92	14.56
DanzigCourts	4	1.5325670	1.53	16.09
DanzigILO	3	1.1494253	1.15	17.24
DanzigLegislativeDecrees	6	2.2988506	2.30	19.54
EasternCarelia	2	0.7662835	0.77	20.31
EasternGreenland	7	2.6819923	2.68	22.99
ElectricityCompanySofiaBulgaria	12	4.5977011	4.60	27.59
ExchangeGreekTurkishPopulations	2	0.7662835	0.77	28.35
FreeZonesUpperSavoyGex	12	4.5977011	4.60	32.95
FreeZonesUpperSavoyGex-SecondPhase	3	1.1494253	1.15	34.10
GermanInterestsUpperSilesia	9	3.4482759	3.45	37.55
GermanMinoritySchools	4	1.5325670	1.53	39.08
GermanSettlers	1	0.3831418	0.38	39.46
GrecoBulgarianAgreement	2	0.7662835	0.77	40.23
GrecoBulgarianCommunities	3	1.1494253	1.15	41.38
GrecoTurkishAgreement	2	0.7662835	0.77	42.15
HungaroCzechoslovakMixedTribunal	1	0.3831418	0.38	42.53

(continued)

shortname	N	exactpercent	roundedpercent	cumulpercent
ILOCompetenceEmployer	2	0.7662835	0.77	43.30
ILOCompetenceMethodsAgriculture	2	0.7662835	0.77	44.06
ILOCompetencePersonsAgriculture	2	0.7662835	0.77	44.83
InterpretationMemelStatute	7	2.6819923	2.68	47.51
InterpretationNo3	1	0.3831418	0.38	47.89
Jaworzina	1	0.3831418	0.38	48.28
Lighthouses	4	1.5325670	1.53	49.81
LighthousesCreteSamos	6	2.2988506	2.30	52.11
Losinger	2	0.7662835	0.77	52.87
Lotus	8	3.0651341	3.07	55.94
MavrommatisJerusalem	1	0.3831418	0.38	56.32
MavrommatisJerusalem-Readaptation	4	1.5325670	1.53	57.85
MavrommatisPalestine	6	2.2988506	2.30	60.15
Meuse	8	3.0651341	3.07	63.22
MinoritySchoolsAlbania	3	1.1494253	1.15	64.37
MinoritySchoolsUpperSilesia	7	2.6819923	2.68	67.05
MonasterySaintNaoum	1	0.3831418	0.38	67.43
NationalityDecrees	2	0.7662835	0.77	68.20
Neuilly	1	0.3831418	0.38	68.58
OderCommission	4	1.5325670	1.53	70.11
OscarChinn	7	2.6819923	2.68	72.80
PajzsCsakyEsterhazy	8	3.0651341	3.07	75.86
PanevezysSaldutiskisRailway	7	2.6819923	2.68	78.54
Pazmany	3	1.1494253	1.15	79.69
PhosphatesMarocco	4	1.5325670	1.53	81.23
PolishAgrarianReform	6	2.2988506	2.30	83.52
PolishWarVessels	3	1.1494253	1.15	84.67
PostalServiceDanzig	2	0.7662835	0.77	85.44

(continued)

shortname	N	exactpercent	roundedpercent	cumulpercent
PrinceVonPless	4	1.5325670	1.53	86.97
RailwayTraffic	1	0.3831418	0.38	87.36
SerbianLoans	5	1.9157088	1.92	89.27
SinoBelgianTreaty	7	2.6819923	2.68	91.95
SocieteCommercialeBelgique	4	1.5325670	1.53	93.49
TreatmentPolishNationals	4	1.5325670	1.53	95.02
TreatyLausanne	2	0.7662835	0.77	95.79
Wimbledon	6	2.2988506	2.30	98.08
WomenNightWork	3	1.1494253	1.15	99.23
WorkersDelegateILO	2	0.7662835	0.77	100.00
Total	261	100.0000000	100.00	100.00

Frequency Table for Variable: applicant

17 unique value(s) detected.

applicant	N	exactpercent	roundedpercent	cumulpercent
BEL	27	10.3448276	10.34	10.34
BGR	2	0.7662835	0.77	11.11
CHE	2	0.7662835	0.77	11.88
CSK	4	1.5325670	1.53	13.41
DEU	45	17.2413793	17.24	30.65
DNK	7	2.6819923	2.68	33.33
EST	7	2.6819923	2.68	36.02
FRA	42	16.0919540	16.09	52.11
GBR	7	2.6819923	2.68	54.79
GBR-CSK-DNK-FRA-DEU	4	1.5325670	1.53	56.32
GBR-FRA-ITA-JPN	13	4.9808429	4.98	61.30
GRC	11	4.2145594	4.21	65.52

(continued)

applicant	N	exactpercent	roundedpercent	cumulpercent
HUN	8	3.0651341	3.07	68.58
ITA	4	1.5325670	1.53	70.11
LNC	69	26.4367816	26.44	96.55
NLD	8	3.0651341	3.07	99.62
TUR	1	0.3831418	0.38	100.00
Total	261	100.0000000	100.00	100.00

Frequency Table for Variable: respondent

18 unique value(s) detected.

respondent	N	exactpercent	roundedpercent	cumulpercent
NA	69	26.4367816	26.44	26.44
BEL	15	5.7471264	5.75	32.18
BGR	12	4.5977011	4.60	36.78
BRA	4	1.5325670	1.53	38.31
CHE	15	5.7471264	5.75	44.06
CHN	7	2.6819923	2.68	46.74
DEU	6	2.2988506	2.30	49.04
ESP	4	1.5325670	1.53	50.57
FRA	4	1.5325670	1.53	52.11
GBR	11	4.2145594	4.21	56.32
GRC	16	6.1302682	6.13	62.45
HUN	4	1.5325670	1.53	63.98
ITA	1	0.3831418	0.38	64.37
LTU	14	5.3639847	5.36	69.73
NOR	7	2.6819923	2.68	72.41
POL	49	18.7739464	18.77	91.19
TUR	8	3.0651341	3.07	94.25

(continued)

respondent	N	exactpercent	roundedpercent	cumulpercent
YUG	15	5.7471264	5.75	100.00
Total	261	100.0000000	100.00	100.00

Frequency Table for Variable: doctype

6 unique value(s) detected.

doctype	N	exactpercent	roundedpercent	cumulpercent
ADV	44	16.8582375	16.86	16.86
ANX	43	16.4750958	16.48	33.33
APP	9	3.4482759	3.45	36.78
DEC	1	0.3831418	0.38	37.16
JUD	113	43.2950192	43.30	80.46
ORD	51	19.5402299	19.54	100.00
Total	261	100.0000000	100.00	100.00

Frequency Table for Variable: collision

3 unique value(s) detected.

collision	N	exactpercent	roundedpercent	cumulpercent
1	252	96.5517241	96.55	96.55
2	7	2.6819923	2.68	99.23
3	2	0.7662835	0.77	100.00
Total	261	100.0000000	100.00	100.00

Frequency Table for Variable: stage

17 unique value(s) detected.

stage	N	exactpercent	roundedpercent	cumulpercent
NA	96	36.7816092	36.78	36.78
AJ	3	1.1494253	1.15	37.93
DH	2	0.7662835	0.77	38.70
DI	9	3.4482759	3.45	42.15
EV	2	0.7662835	0.77	42.91
EV-SE	6	2.2988506	2.30	45.21
EX	2	0.7662835	0.77	45.98
IM	10	3.8314176	3.83	49.81
IN	1	0.3831418	0.38	50.19
JO	1	0.3831418	0.38	50.57
JO-TL	1	0.3831418	0.38	50.96
ME	83	31.8007663	31.80	82.76
PO	33	12.6436782	12.64	95.40
PR	1	0.3831418	0.38	95.79
SE	3	1.1494253	1.15	96.93
TL	7	2.6819923	2.68	99.62
TL-DH	1	0.3831418	0.38	100.00
Total	261	100.0000000	100.00	100.00

Frequency Table for Variable: opinion

9 unique value(s) detected.

opinion	N	exactpercent	roundedpercent	cumulpercent
NA	52	19.9233716	19.92	19.92
0	100	38.3141762	38.31	58.24
1	40	15.3256705	15.33	73.56
2	30	11.4942529	11.49	85.06
3	18	6.8965517	6.90	91.95
4	11	4.2145594	4.21	96.17

(continued)

opinion	N	exactpercent	roundedpercent	cumulpercent
5	7	2.6819923	2.68	98.85
6	2	0.7662835	0.77	99.62
7	1	0.3831418	0.38	100.00
Total	261	100.0000000	100.00	100.00

Frequency Table for Variable: language

1 unique value(s) detected.

language	N	exactpercent	roundedpercent	cumulpercent
FR	261	100	100	100
Total	261	100	100	100

Frequency Table for Variable: year

19 unique value(s) detected.

year	N	exactpercent	roundedpercent	cumulpercent
1922	7	2.6819923	2.68	2.68
1923	13	4.9808429	4.98	7.66
1924	8	3.0651341	3.07	10.73
1925	11	4.2145594	4.21	14.94
1926	9	3.4482759	3.45	18.39
1927	26	9.9616858	9.96	28.35
1928	23	8.8122605	8.81	37.16
1929	21	8.0459770	8.05	45.21
1930	9	3.4482759	3.45	48.66
1931	16	6.1302682	6.13	54.79
1932	23	8.8122605	8.81	63.60

(continued)

year	N	exactpercent	roundedpercent	cumulpercent
1933	20	7.6628352	7.66	71.26
1934	11	4.2145594	4.21	75.48
1935	9	3.4482759	3.45	78.93
1936	10	3.8314176	3.83	82.76
1937	17	6.5134100	6.51	89.27
1938	6	2.2988506	2.30	91.57
1939	21	8.0459770	8.05	99.62
1940	1	0.3831418	0.38	100.00
Total	261	100.0000000	100.00	100.00

Frequency Table for Variable: minority

3 unique value(s) detected.

minority	N	exactpercent	roundedpercent	cumulpercent
NA	52	19.92337	19.92	19.92
0	100	38.31418	38.31	58.24
1	109	41.76245	41.76	100.00
Total	261	100.00000	100.00	100.00

Frequency Table for Variable: fullname

78 unique value(s) detected.

fullname	N	exactpercent	roundedpercent	cumulpercent
Access to German Minority Schools in Upper Silesia	4	1.5325670	1.53	1.53
Access to, or Anchorage in, the port of Danzig, of Polish War Vessels	3	1.1494253	1.15	2.68

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Acquisition of Polish Nationality	2	0.7662835	0.77	3.45
Administration of the Prince von Pless (Discontinuance)	1	0.3831418	0.38	3.83
Administration of the Prince von Pless (Interim Measures of Protection)	1	0.3831418	0.38	4.21
Administration of the Prince von Pless (Preliminary Objections)	1	0.3831418	0.38	4.60
Administration of the Prince von Pless (Prorogation)	1	0.3831418	0.38	4.98
Appeal from a Judgment of the Hungaro-Czecoslovak Mixed Arbitral Tribunal (The Peter Pázmány University v. The State of Czechoslovakia)	3	1.1494253	1.15	6.13
Appeals from Certain Judgments of the Hungaro-Czechoslovak Mixed Arbitral Tribunal	1	0.3831418	0.38	6.51
Borchgrave (Discontinuance)	1	0.3831418	0.38	6.90
Borchgrave (Preliminary Objections)	3	1.1494253	1.15	8.05
Certain German Interests in Polish Upper Silesia (Merits)	6	2.2988506	2.30	10.34
Certain German Interests in Polish Upper Silesia (Preliminary Objections)	3	1.1494253	1.15	11.49
Competence of the ILO in regard to International Regulation of the Conditions of the Labour of Persons Employed in Agriculture	2	0.7662835	0.77	12.26

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Competence of the ILO to Examine Proposal for the Organization and Development of the Methods of Agricultural Production	2	0.7662835	0.77	13.03
Competence of the ILO to Regulate, Incidentally, the Personal Work of the Employer	2	0.7662835	0.77	13.79
Consistency of Certain Danzig Legislative Decrees with the Constitution of the Free City	6	2.2988506	2.30	16.09
Customs Regime between Germany and Austria (Protocol of March 19th, 1931)	7	2.6819923	2.68	18.77
Delimitation of the Territorial Waters between the Island of Castellorizo and the Coasts of Anatolia	1	0.3831418	0.38	19.16
Denunciation of the Treaty of November 2nd, 1865, between China and Belgium	7	2.6819923	2.68	21.84
Designation of the Workers' Delegate for the Netherlands at the Third Session of the International Labour Conference	2	0.7662835	0.77	22.61
Diversion of Water from the Meuse	8	3.0651341	3.07	25.67
Electricity Company of Sofia and Bulgaria (Date of Hearing)	1	0.3831418	0.38	26.05
Electricity Company of Sofia and Bulgaria (Interim Measures of Protection)	1	0.3831418	0.38	26.44
Electricity Company of Sofia and Bulgaria (Preliminary Objections)	10	3.8314176	3.83	30.27

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Exchange of Greek and Turkish Populations (Lausanne Convention VI, January 30th, 1923, Article 2)	2	0.7662835	0.77	31.03
Factory at Chorzów (Claim for Indemnity) (Jurisdiction)	2	0.7662835	0.77	31.80
Factory at Chorzów (Claim for Indemnity) (Merits)	7	2.6819923	2.68	34.48
Factory at Chorzów (Indemnities)	4	1.5325670	1.53	36.02
Free City of Danzig and International Labour Organization	3	1.1494253	1.15	37.16
Free Zones of Upper Savoy and the District of Gex	12	4.5977011	4.60	41.76
Free Zones of Upper Savoy and the District of Gex (Second Phase)	3	1.1494253	1.15	42.91
Greco-Bulgarian Communities	3	1.1494253	1.15	44.06
Interpretation of Article 3, Paragraph 2, of the Treaty of Lausanne (Frontier between Turkey and Iraq)	2	0.7662835	0.77	44.83
Interpretation of Judgment No.3	1	0.3831418	0.38	45.21
Interpretation of Judgments Nos. 7 and 8 (The Chorzów Factory)	2	0.7662835	0.77	45.98
Interpretation of the Convention of 1919 concerning Employment of Women during the Night	3	1.1494253	1.15	47.13
Interpretation of the Greco-Bulgarian Agreement of December 9th, 1927 (Caphandaris-Molloff Agreement)	2	0.7662835	0.77	47.89

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Interpretation of the Greco-Turkish Agreement of December 1st, 1926 (Final Protocol, Article IV)	2	0.7662835	0.77	48.66
Interpretation of the Statute of the Memel Territory (Merits)	4	1.5325670	1.53	50.19
Interpretation of the Statute of the Memel Territory (Preliminary Objection)	3	1.1494253	1.15	51.34
Jaworzina (Polish-Czechoslovakian Frontier)	1	0.3831418	0.38	51.72
Jurisdiction of the Courts of Danzig	3	1.1494253	1.15	52.87
Jurisdiction of the Courts of Danzig (Pecuniary Claims of Danzig Railway Officials who have passed into the Polish Service, against the Polish Railways Administration)	1	0.3831418	0.38	53.26
Jurisdiction of the European Commission of the Danube between Galatz and Braila	5	1.9157088	1.92	55.17
Legal Status of Eastern Greenland	4	1.5325670	1.53	56.70
Legal Status of the South-Eastern Territory of Greenland	3	1.1494253	1.15	57.85
Lighthouses Case between France and Greece	4	1.5325670	1.53	59.39
Lighthouses in Crete and Samos	6	2.2988506	2.30	61.69
Losinger & Co Case (Discontinuance)	1	0.3831418	0.38	62.07
Losinger & Co Case (Preliminary Objection)	1	0.3831418	0.38	62.45
Mavrommatis Jerusalem Concessions	1	0.3831418	0.38	62.84

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Mavrommatis Palestine Concessions	6	2.2988506	2.30	65.13
Minority Schools in Albania	3	1.1494253	1.15	66.28
Monastery of Saint-Naoum (Albanian Frontier)	1	0.3831418	0.38	66.67
Nationality Decrees Issued in Tunis and Morocco	2	0.7662835	0.77	67.43
Oscar Chinn	7	2.6819923	2.68	70.11
Pajzs, Czáky, Esterházy (Merits)	7	2.6819923	2.68	72.80
Pajzs, Czáky, Esterházy (Preliminary Objection)	1	0.3831418	0.38	73.18
Panevezys-Saldutiskis Railway (Merits)	6	2.2988506	2.30	75.48
Panevezys-Saldutiskis Railway (Preliminary Objections)	1	0.3831418	0.38	75.86
Payment in Gold of the Brazilian Federal Loans Issued in France	4	1.5325670	1.53	77.39
Payment of Various Serbian Loans Issued in France	5	1.9157088	1.92	79.31
Phosphates in Morocco	4	1.5325670	1.53	80.84
Polish Agrarian Reform and the German Minority (Discontinuance)	1	0.3831418	0.38	81.23
Polish Agrarian Reform and the German Minority (Interim Measures of Protection)	5	1.9157088	1.92	83.14
Polish Postal Service in Danzig	2	0.7662835	0.77	83.91
Railway Traffic between Lithuania and Poland (Railway Sector Landwarów-Kaisiadorys)	1	0.3831418	0.38	84.29
Readaptation of the Mavrommatis Jerusalem Concessions (Jurisdiction)	4	1.5325670	1.53	85.82

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Rights of Minorities in Upper Silesia (Minority Schools)	7	2.6819923	2.68	88.51
S.S. Lotus	8	3.0651341	3.07	91.57
S.S. Wimbledon	6	2.2988506	2.30	93.87
Settlers of German Origin in the Territory Ceded by Germany to Poland	1	0.3831418	0.38	94.25
Société Commerciale de Belgique	4	1.5325670	1.53	95.79
Status of Eastern Carelia	2	0.7662835	0.77	96.55
Territorial Jurisdiction of the International Commission of the River Oder	4	1.5325670	1.53	98.08
Treatment of Polish Nationals and Other Persons of Polish Origin or Speech in the Danzig Territory	4	1.5325670	1.53	99.62
Treaty of Neuilly, Article 179, Annex, Paragraph 4 (Interpretation)	1	0.3831418	0.38	100.00
Total	261	100.0000000	100.00	100.00

Frequency Table for Variable: caseno

83 unique value(s) detected.

caseno	N	exactpercent	roundedpercent	cumulpercent
A1	6	2.2988506	2.30	2.30
A10	8	3.0651341	3.07	5.36
A11	4	1.5325670	1.53	6.90
A12	2	0.7662835	0.77	7.66
A13	2	0.7662835	0.77	8.43
A14	1	0.3831418	0.38	8.81

(continued)

caseno	N	exactpercent	roundedpercent	cumulpercent
A15	7	2.6819923	2.68	11.49
A16	1	0.3831418	0.38	11.88
A17	7	2.6819923	2.68	14.56
A18	1	0.3831418	0.38	14.94
A19	2	0.7662835	0.77	15.71
A2	6	2.2988506	2.30	18.01
A20	5	1.9157088	1.92	19.92
A21	4	1.5325670	1.53	21.46
A22	6	2.2988506	2.30	23.75
A23	4	1.5325670	1.53	25.29
A24	3	1.1494253	1.15	26.44
A3	1	0.3831418	0.38	26.82
A4	1	0.3831418	0.38	27.20
A5	1	0.3831418	0.38	27.59
A6	3	1.1494253	1.15	28.74
A7	6	2.2988506	2.30	31.03
A8	4	1.5325670	1.53	32.57
A9	2	0.7662835	0.77	33.33
AB40	4	1.5325670	1.53	34.87
AB41	7	2.6819923	2.68	37.55
AB42	1	0.3831418	0.38	37.93
AB43	3	1.1494253	1.15	39.08
AB44	4	1.5325670	1.53	40.61
AB45	2	0.7662835	0.77	41.38
AB46	6	2.2988506	2.30	43.68
AB47	3	1.1494253	1.15	44.83
AB48	2	0.7662835	0.77	45.59
AB49	4	1.5325670	1.53	47.13
AB50	3	1.1494253	1.15	48.28

(continued)

caseno	N	exactpercent	roundedpercent	cumulpercent
AB51	1	0.3831418	0.38	48.66
AB52	1	0.3831418	0.38	49.04
AB53	4	1.5325670	1.53	50.57
AB54	1	0.3831418	0.38	50.96
AB55	1	0.3831418	0.38	51.34
AB56	1	0.3831418	0.38	51.72
AB57	1	0.3831418	0.38	52.11
AB58	5	1.9157088	1.92	54.02
AB59	1	0.3831418	0.38	54.41
AB60	1	0.3831418	0.38	54.79
AB61	3	1.1494253	1.15	55.94
AB62	4	1.5325670	1.53	57.47
AB63	7	2.6819923	2.68	60.15
AB64	3	1.1494253	1.15	61.30
AB65	6	2.2988506	2.30	63.60
AB66	1	0.3831418	0.38	63.98
AB67	1	0.3831418	0.38	64.37
AB68	7	2.6819923	2.68	67.05
AB69	1	0.3831418	0.38	67.43
AB70	8	3.0651341	3.07	70.50
AB71	6	2.2988506	2.30	72.80
AB72	3	1.1494253	1.15	73.95
AB73	1	0.3831418	0.38	74.33
AB74	4	1.5325670	1.53	75.86
AB75	1	0.3831418	0.38	76.25
AB76	6	2.2988506	2.30	78.54
AB77	10	3.8314176	3.83	82.38
AB78	4	1.5325670	1.53	83.91
AB79	1	0.3831418	0.38	84.29

(continued)

caseno	N	exactpercent	roundedpercent	cumulpercent
AB80	1	0.3831418	0.38	84.67
B1	2	0.7662835	0.77	85.44
B10	2	0.7662835	0.77	86.21
B11	2	0.7662835	0.77	86.97
B12	2	0.7662835	0.77	87.74
B13	2	0.7662835	0.77	88.51
B14	5	1.9157088	1.92	90.42
B15	4	1.5325670	1.53	91.95
B16	2	0.7662835	0.77	92.72
B17	3	1.1494253	1.15	93.87
B18	3	1.1494253	1.15	95.02
B2	2	0.7662835	0.77	95.79
B3	2	0.7662835	0.77	96.55
B4	2	0.7662835	0.77	97.32
B5	2	0.7662835	0.77	98.08
B6	1	0.3831418	0.38	98.47
B7	2	0.7662835	0.77	99.23
B8	1	0.3831418	0.38	99.62
B9	1	0.3831418	0.38	100.00
Total	261	100.0000000	100.00	100.00

Frequency Table for Variable: applicant_region

5 unique value(s) detected.

applicant_region	N	exactpercent	roundedpercent	cumulpercent
Asia	1	0.3831418	0.38	0.38
Europe	174	66.6666667	66.67	67.05
Europe|Europe|Europe|Asia	13	4.9808429	4.98	72.03

(continued)

applicant_region	N	exactpercent	roundedpercent	cumulpercent
Europe|Europe|Europe|Europe|Europe	4	1.5325670	1.53	73.56
NA	69	26.4367816	26.44	100.00
Total	261	100.0000000	100.00	100.00

Frequency Table for Variable: respondent_region

4 unique value(s) detected.

respondent_region	N	exactpercent	roundedpercent	cumulpercent
NA	69	26.436782	26.44	26.44
Americas	4	1.532567	1.53	27.97
Asia	15	5.747126	5.75	33.72
Europe	173	66.283525	66.28	100.00
Total	261	100.000000	100.00	100.00

Frequency Table for Variable: applicant_subregion

8 unique value(s) detected.

applicant_subregion	N	exactpercent	roundedpercent	cumulpercent
Eastern Europe	14	5.3639847	5.36	5.36
NA	69	26.4367816	26.44	31.80
Northern Europe	21	8.0459770	8.05	39.85
Northern Europe|Eastern Europe|Northern Europe|Western Europe|Western Europe	4	1.5325670	1.53	41.38
Northern Europe|Western Europe|Southern Europe|Eastern Asia	13	4.9808429	4.98	46.36
Southern Europe	15	5.7471264	5.75	52.11

(continued)

applicant_subregion	N	exactpercent	roundedpercent	cumulpercent
Western Asia	1	0.3831418	0.38	52.49
Western Europe	124	47.5095785	47.51	100.00
Total	261	100.0000000	100.00	100.00

Frequency Table for Variable: respondent_subregion

8 unique value(s) detected.

respondent_subregion	N	exactpercent	roundedpercent	cumulpercent
NA	69	26.436782	26.44	26.44
Eastern Asia	7	2.681992	2.68	29.12
Eastern Europe	65	24.904215	24.90	54.02
Latin America and the Caribbean	4	1.532567	1.53	55.56
Northern Europe	32	12.260536	12.26	67.82
Southern Europe	36	13.793103	13.79	81.61
Western Asia	8	3.065134	3.07	84.67
Western Europe	40	15.325670	15.33	100.00
Total	261	100.000000	100.00	100.00

Frequency Table for Variable: doi_concept

1 unique value(s) detected.

doi_concept	N	exactpercent	roundedpercent	cumulpercent
10.5281/zenodo.3840479	261	100	100	100
Total	261	100	100	100

Frequency Table for Variable: doi_version

1 unique value(s) detected.

doi_version	N	exactpercent	roundedpercent	cumulpercent
10.5281/zenodo.7051934	261	100	100	100
Total	261	100	100	100

Frequency Table for Variable: version

1 unique value(s) detected.

version	N	exactpercent	roundedpercent	cumulpercent
1.1.0	261	100	100	100
Total	261	100	100	100

Frequency Table for Variable: license

1 unique value(s) detected.

license	N	exactpercent	roundedpercent	cumulpercent
Creative Commons Zero 1.0 Universal	261	100	100	100
Total	261	100	100	100

21 Visualize Frequency Tables

21.1 Load Tables

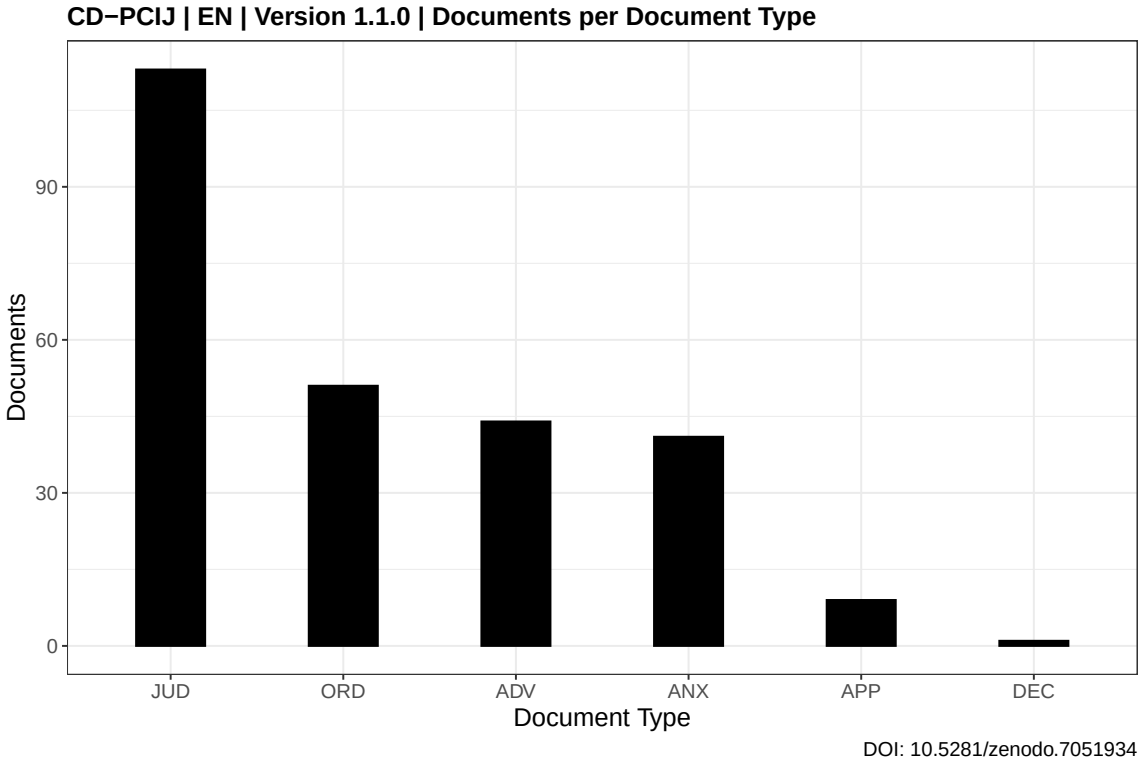
```
prefix.en <- paste0("ANALYSIS/",
                    datashort,
                    "_EN_01_FrequencyTable_var-")

prefix.fr <- paste0("ANALYSIS/",
                    datashort,
                    "_FR_01_FrequencyTable_var-")

table.en.doctype <- fread(paste0(prefix.en,
                                 "doctype.csv"))

table.en.opinion <- fread(paste0(prefix.en,
                                 "opinion.csv"))

table.en.year <- fread(paste0(prefix.en,
                              "year.csv"))

table.fr.doctype <- fread(paste0(prefix.fr,
                                 "doctype.csv"))

table.fr.opinion <- fread(paste0(prefix.fr,
                                 "opinion.csv"))

table.fr.year <- fread(paste0(prefix.fr,
                              "year.csv"))
```

21.2 Doctype

21.2.1 English

```
freqtable <- table.en.doctype[-.N]
```

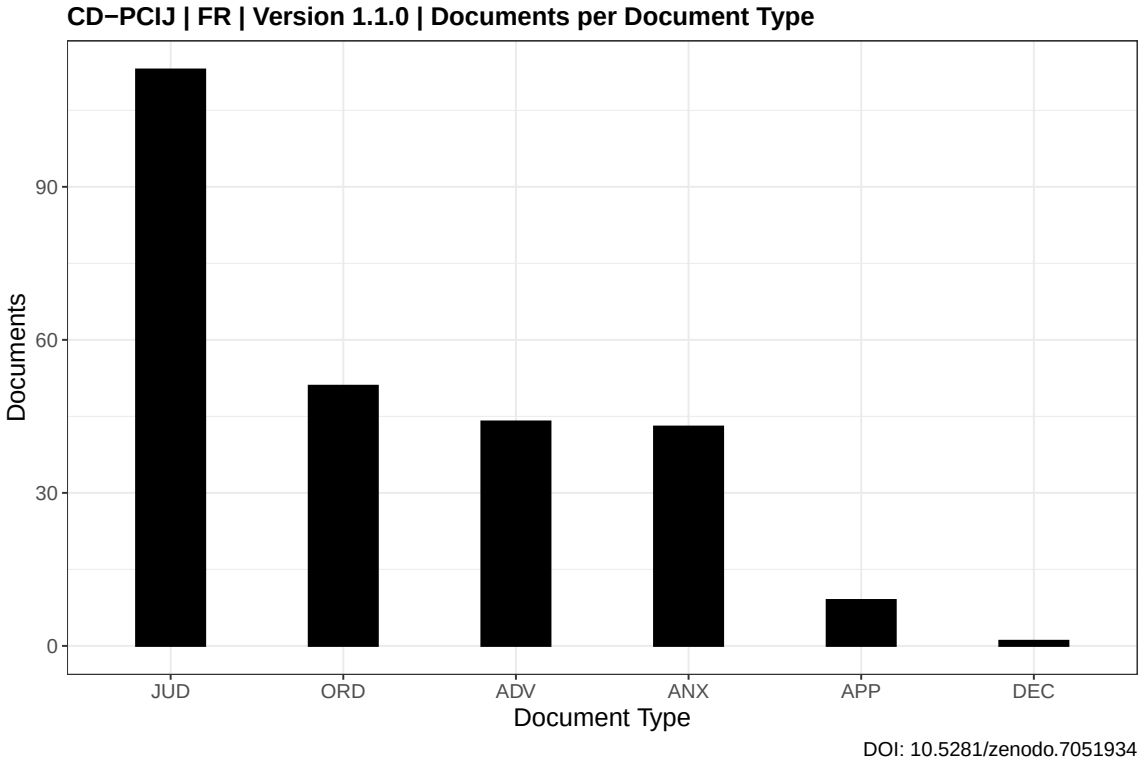
```
ggplot(data = freqtable) +
    geom_bar(aes(x = reorder(doctype,
                             -N),
                 y = N),
             stat = "identity",
             fill = "black",
             color = "black",
             width = 0.4) +
    theme_bw() +
    labs(
        title = paste(datashort,
                      "| EN | Version",
                      version,
                      "| Documents per Document Type"),
        caption = paste("DOI:",
                        doi.version),
        x = "Document Type",
        y = "Documents"
    )+
    theme(
        text = element_text(size = 14),
        plot.title = element_text(size = 14,
                                  face = "bold"),
        legend.position = "none",
        plot.margin = margin(10, 20, 10, 10)
    )
```


CD−PCIJ | EN | Version 1.1.0 | Documents per Document Type
Documents
90
60
30
0
JUD
ORD
ADV
ANX
APP
DEC
Document Type
DOI: 10.5281/zenodo.7051934

21.2.2 French

```
freqtable <- table.fr.doctype[-.N]
```

```
ggplot(data = freqtable) +
    geom_bar(aes(x = reorder(doctype,
                             -N),
                 y = N),
             stat = "identity",
             fill = "black",
             color = "black",
             width = 0.4) +
    theme_bw() +
    labs(
        title = paste(datashort,
                      "| FR | Version",
                      version,
                      "| Documents per Document Type"),
        caption = paste("DOI:",
                        doi.version),
        x = "Document Type",
        y = "Documents"
    )+
    theme(
        text = element_text(size = 14),
        plot.title = element_text(size = 14,
                                  face = "bold"),
        legend.position = "none",
        plot.margin = margin(10, 20, 10, 10)
    )
```

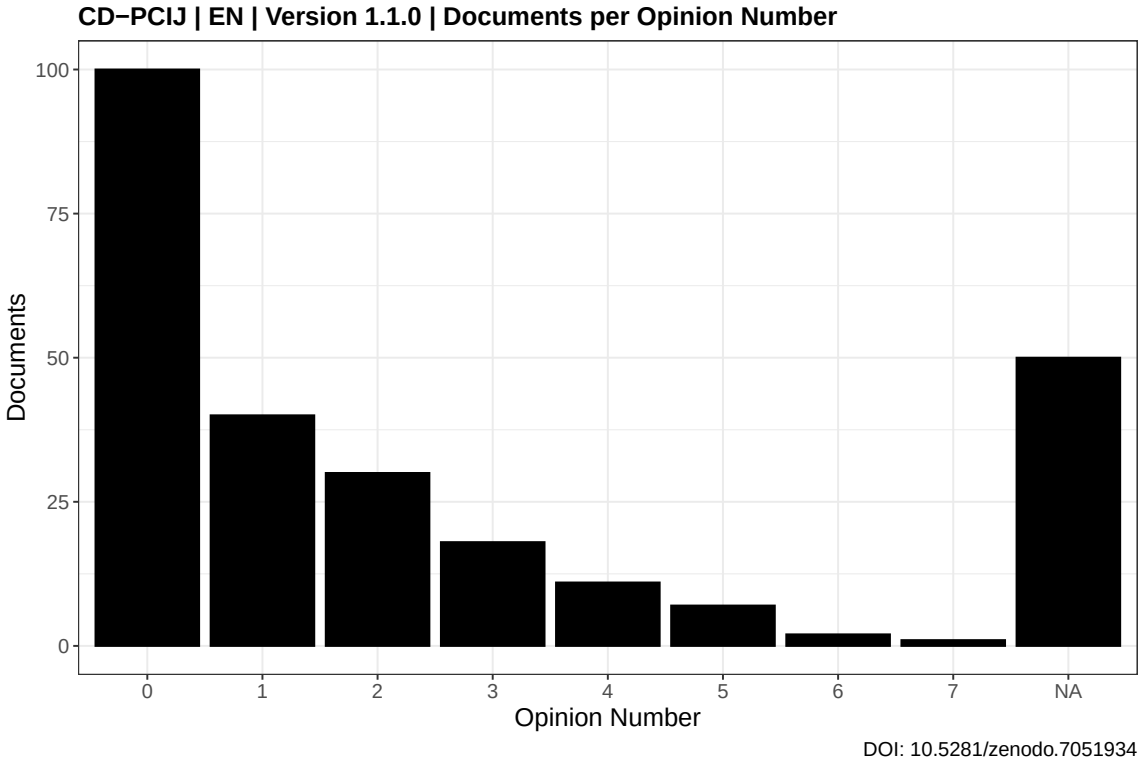

CD−PCIJ | FR | Version 1.1.0 | Documents per Document Type
Documents
90
60
30
0
JUD
ORD
ADV
ANX
APP
DEC
Document Type
DOI: 10.5281/zenodo.7051934

21.3 Opinion

21.3.1 English

```
freqtable <- table.en.opinion[-.N]
```

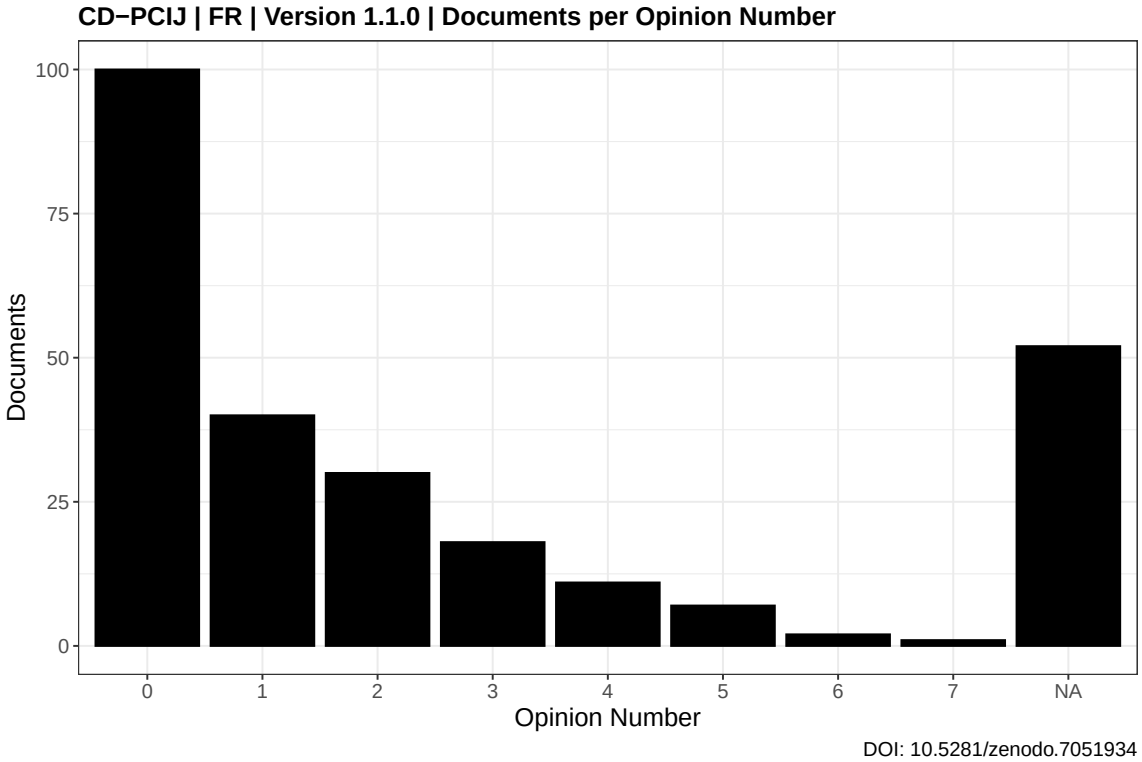
```
ggplot(data = freqtable) +
    geom_bar(aes(x = reorder(opinion,
                             -N),
                 y = N),
             stat = "identity",
             fill = "black",
             color = "black") +
    theme_bw() +
    labs(
        title = paste(datashort,
                      "| EN | Version",
                      version,
                      "| Documents per Opinion Number"),
        caption = paste("DOI:",
                        doi.version),
        x = "Opinion Number",
        y = "Documents"
    )+
    theme(
        text = element_text(size = 14),
        plot.title = element_text(size = 14,
                                  face = "bold"),
        legend.position = "none",
        plot.margin = margin(10, 20, 10, 10)
    )
```


CD−PCIJ | EN | Version 1.1.0 | Documents per Opinion Number
Documents
100
75
50
25
0
0
1
2
3
4
5
6
7
NA
Opinion Number
DOI: 10.5281/zenodo.7051934

21.3.2 French

```
freqtable <- table.fr.opinion[-.N]
```

```
ggplot(data = freqtable) +
    geom_bar(aes(x = reorder(opinion,
                             -N),
                 y = N),
             stat = "identity",
             fill = "black",
             color = "black") +
    theme_bw() +
    labs(
        title = paste(datashort,
                      "| FR | Version",
                      version,
                      "| Documents per Opinion Number"),
        caption = paste("DOI:",
                        doi.version),
        x = "Opinion Number",
        y = "Documents"
    )+
    theme(
        text = element_text(size = 14),
        plot.title = element_text(size = 14,
                                  face = "bold"),
        legend.position = "none",
        plot.margin = margin(10, 20, 10, 10)
    )
```

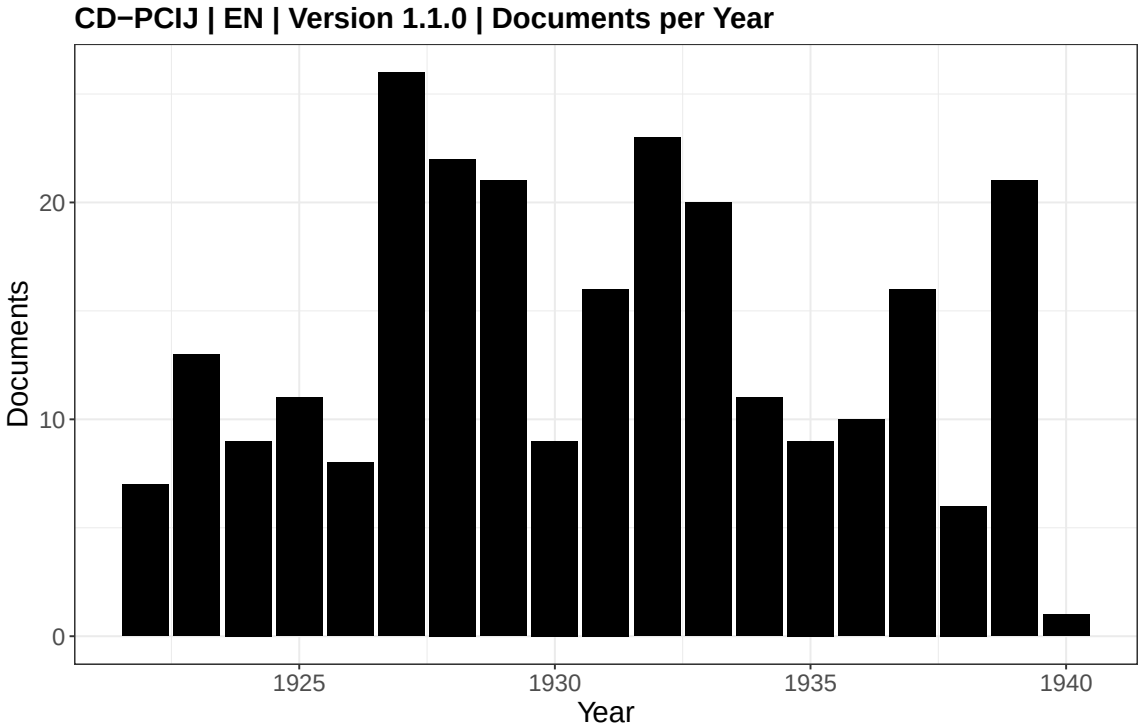

CD−PCIJ | FR | Version 1.1.0 | Documents per Opinion Number
Documents
100
75
50
25
0
0
1
2
3
4
5
6
7
NA
Opinion Number
DOI: 10.5281/zenodo.7051934

21.4 Year

21.4.1 English

```
freqtable <- table.en.year[-.N][,lapply(.SD, as.numeric)]
```

```
ggplot(data = freqtable) +
    geom_bar(aes(x = year,
                 y = N),
             stat = "identity",
             fill = "black") +
    theme_bw()+
    labs(
        title = paste(datashort,
                      "| EN | Version",
                      version,
                      "| Documents per Year"),
        caption = paste("DOI:",
                        doi.version),
        x = "Year",
        y = "Documents"
    )+
    theme(
        text = element_text(size = 16),
        plot.title = element_text(size = 16,
                                  face = "bold"),
        legend.position = "none",
        plot.margin = margin(10, 20, 10, 10)
    )
```

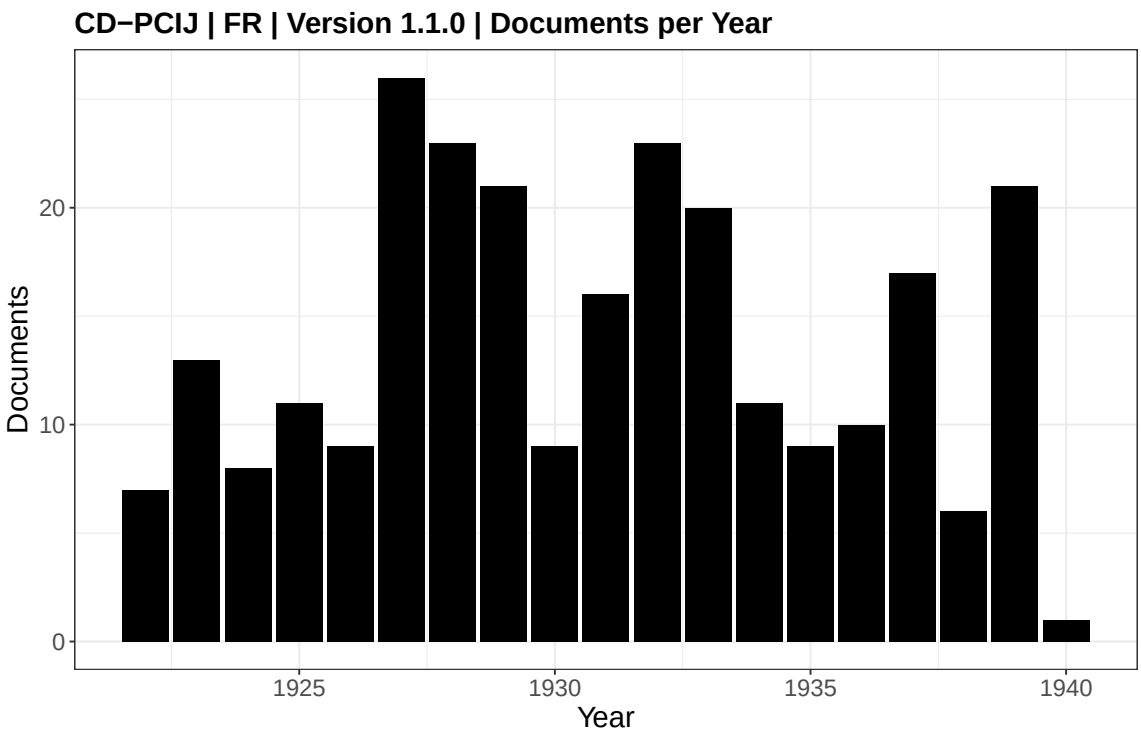


DOI: 10.5281/zenodo.7051934

21.4.2 French

```
freqtable <- table.fr.year[-.N][,lapply(.SD, as.numeric)]
```

```
ggplot(data = freqtable) +
    geom_bar(aes(x = year,
                 y = N),
             stat = "identity",
             fill = "black") +
    theme_bw() +
    labs(
        title = paste(datashort,
                      "| FR | Version",
                      version,
                      "| Documents per Year"),
        caption = paste("DOI:",
                        doi.version),
        x = "Year",
        y = "Documents"
    )+
    theme(
        text = element_text(size = 16),
        plot.title = element_text(size = 16,
                                  face = "bold"),
        legend.position = "none",
        plot.margin = margin(10, 20, 10, 10)
    )
```



DOI: 10.5281/zenodo.7051934

22 Summary Statistics

22.1 Linguistic Metrics

For the text of each document the number of characters, tokens, types and sentences will be calculated.

22.1.1 Show Function: f.lingsummarize.iterator

```
print(f.lingsummarize.iterator)
```

function(dt, threads = detectCores(), chunksize = 1){

```
begin.dopar <- Sys.time()

dt <- dt[,.(doc_id, text)]

nchars <- dt[, lapply(.(text), nchar)]

print(paste0("Parallel processing using ",
             threads,
             " threads. Begin at ",
             begin.dopar,
             ". Processing ",
             dt[,.N],
             " documents with a total length of ",
             sum(nchars),
             " characters."))


ord <- order(-nchars)
dt <- dt[ord]

cl <- makeForkCluster(threads)
registerDoParallel(cl)


itx <- iter(dt["nchars" > 0],
            by = "row",
            chunksize = chunksize)

result.list <- foreach(i = itx,
                       .errorhandling = 'pass') %dopar% {

                           corpus <- corpus(i)

                           tokens <- tokens(corpus,
                                            what = "word",
                                            remove_punct = FALSE,
                                            remove_symbols = FALSE,
                                            remove_numbers = FALSE,
                                            remove_url = FALSE,
```

```
                                                remove_separators = TRUE,
                                                split_hyphens = FALSE,
                                                include_docvars = FALSE,
                                                padding = FALSE
                                                )

                              ntokens <- unname(ntoken(tokens))
                              ntypes  <- unname(ntype(tokens))
                              nsentences <- unname(nsentence(corpus))

                              temp <- data.table(ntokens,
                                                 ntypes,
                                                 nsentences)

                              return(temp)
                          }

stopCluster(cl)

end.dopar <- Sys.time()
duration.dopar <- end.dopar - begin.dopar

result.dt <- rbindlist(result.list)

summary.corpus <- cbind(nchars[ord],
                        result.dt)

setnames(summary.corpus,
         "V1",
         "nchars")

if(dt["nchars" == 0, .N] > 0){

    dt.charnull <- dt["nchars" == 0]
    dt.charnull$text <- NULL
    dt.charnull$ntokens <- rep(0, dt.charnull[,.N])
    dt.charnull$ntypes <- rep(0, dt.charnull[,.N])
    dt.charnull$nsentences <- rep(0, dt.charnull[,.N])

    summary.corpus <- rbind(summary.corpus,
                            dt.charnull)
}


summary.corpus <- summary.corpus[order(ord)]


print(paste0("Runtime was ",
             round(duration.dopar,
                   digits = 2),
             " ",
             attributes(duration.dopar)$units,
             ". Ended at ",
             end.dopar, "."))

return(summary.corpus)
```

}

22.1.2 Calculate Linguistic Metrics

```
quanteda_options(tokens_locale = "en") # Set Locale for Tokenization

summary.corpus.en <- f.lingsummarize.iterator(data.tesseract.en,
                                              threads = fullCores,
                                              chunksize = 1)
```

```
## [1] "Parallel processing using 16 threads. Begin at 2022-09-06 20:00:55.
    Processing 259 documents with a total length of 6828188 characters."
```

```
## Warning in xtfrm.data.frame(x): cannot xtfrm data frames
```

```
## [1] "Runtime was 1.3 secs. Ended at 2022-09-06 20:00:57."
```

```
quanteda_options(tokens_locale = "fr") # Set Locale for Tokenization

summary.corpus.fr <- f.lingsummarize.iterator(data.tesseract.fr,
                                              threads = fullCores,
                                              chunksize = 1)
```

```
## [1] "Parallel processing using 16 threads. Begin at 2022-09-06 20:00:57.
    Processing 261 documents with a total length of 6890471 characters."
```

```
## Warning in xtfrm.data.frame(x): cannot xtfrm data frames
```

```
## [1] "Runtime was 1.38 secs. Ended at 2022-09-06 20:00:58."
```

22.1.3 Add Linguistic Metrics to Full Corpora

```
data.tesseract.en <- cbind(data.tesseract.en,
                           summary.corpus.en)

data.tesseract.fr <- cbind(data.tesseract.fr,
                           summary.corpus.fr)
```

22.1.4 Create Metadata-only Variants

```
meta.tesseract.en <- data.tesseract.en[, !"text"]
meta.tesseract.fr <- data.tesseract.fr[, !"text"]
```

22.1.5 Calculate Summaries: English

```
dt.summary.ling <- meta.tesseract.en[, lapply(.SD,
                                              function(x)unclass(summary(x))),
                                       .SDcols = c("nchars",
                                                   "ntokens",
                                                   "ntypes",
                                                   "nsentences")]


dt.sums.ling <- meta.tesseract.en[,
                                  lapply(.SD, sum),
                                  .SDcols = c("nchars",
                                              "ntokens",
                                              "ntypes",
                                              "nsentences")]

quanteda_options(tokens_locale = "en") # Set Locale for Tokenization

tokens.temp <- tokens(corpus(data.tesseract.en),
                      what = "word",
                      remove_punct = FALSE,
                      remove_symbols = FALSE,
                      remove_numbers = FALSE,
                      remove_url = FALSE,
                      remove_separators = TRUE,
                      split_hyphens = FALSE,
                      include_docvars = FALSE,
                      padding = FALSE
                      )

dt.sums.ling$ntypes <- nfeat(dfm(tokens.temp))




dt.stats.ling <- rbind(dt.sums.ling,
                       dt.summary.ling)

dt.stats.ling <- transpose(dt.stats.ling,
                           keep.names = "names")

setnames(dt.stats.ling, c("Variable",
                          "Total",
                          "Min",
                          "Quart1",
                          "Median",
                          "Mean",
```

```
                                "Quart3",
                                "Max"))
```

22.1.6 Show Summaries: English

```
kable(dt.stats.ling,
      format.args = list(big.mark = ","),
      format = "latex",
      booktabs = TRUE)
```

Variable	Total	Min	Quart1	Median	Mean	Quart3	Max
nchars	6,828,188	308	6,322.5	16,264	26,363.6602	35,296.5	180,844
ntokens	1,296,536	62	1,266.5	3,107	5,005.9305	6,719.5	33,626
ntypes	22,008	46	364.5	697	855.4672	1,170.5	3,145
nsentences	38,015	6	38.0	96	146.7761	200.5	824

22.1.7 Write Summaries to Disk: English

```
fwrite(dt.stats.ling,
       paste0(outputdir,
              datashort,
              "_EN_00_CorpusStatistics_Summaries_Linguistic.csv"),
       na = "NA")
```

22.1.8 Calculate Summaries: French

```
dt.summary.ling <- meta.tesseract.fr[, lapply(.SD,
                                              function(x)unclass(summary(x))),
                                     .SDcols = c("nchars",
                                                 "ntokens",
                                                 "ntypes",
                                                 "nsentences")]


dt.sums.ling <- meta.tesseract.fr[,
                                  lapply(.SD, sum),
                                  .SDcols = c("nchars",
                                              "ntokens",
                                              "ntypes",
                                              "nsentences")]


quanteda_options(tokens_locale = "fr") # Set Locale for Tokenization

tokens.temp <- tokens(corpus(data.tesseract.fr),
                      what = "word",
                      remove_punct = FALSE,
                      remove_symbols = FALSE,
                      remove_numbers = FALSE,
                      remove_url = FALSE,
                      remove_separators = TRUE,
                      split_hyphens = FALSE,
                      include_docvars = FALSE,
                      padding = FALSE
                      )

dt.sums.ling$ntypes <- nfeat(dfm(tokens.temp))




dt.stats.ling <- rbind(dt.sums.ling,
                       dt.summary.ling)

dt.stats.ling <- transpose(dt.stats.ling,
                           keep.names = "names")

setnames(dt.stats.ling, c("Variable",
                          "Total",
                          "Min",
                          "Quart1",
                          "Median",
                          "Mean",
                          "Quart3",
                          "Max"))
```

22.1.9 Show Summaries: French

```
kable(dt.stats.ling,
      format.args = list(big.mark = ","),
      format = "latex",
      booktabs = TRUE)
```

Variable	Total	Min	Quart1	Median	Mean	Quart3	Max
nchars	6,890,471	343	6,481	15,690	26,400.2720	36,362	182,540
ntokens	1,262,184	70	1,215	2,851	4,835.9540	6,523	32,769
ntypes	28,578	52	406	773	980.5019	1,373	3,703
nsentences	34,099	5	33	88	130.6475	177	745

22.1.10 Write Summaries to Disk: French

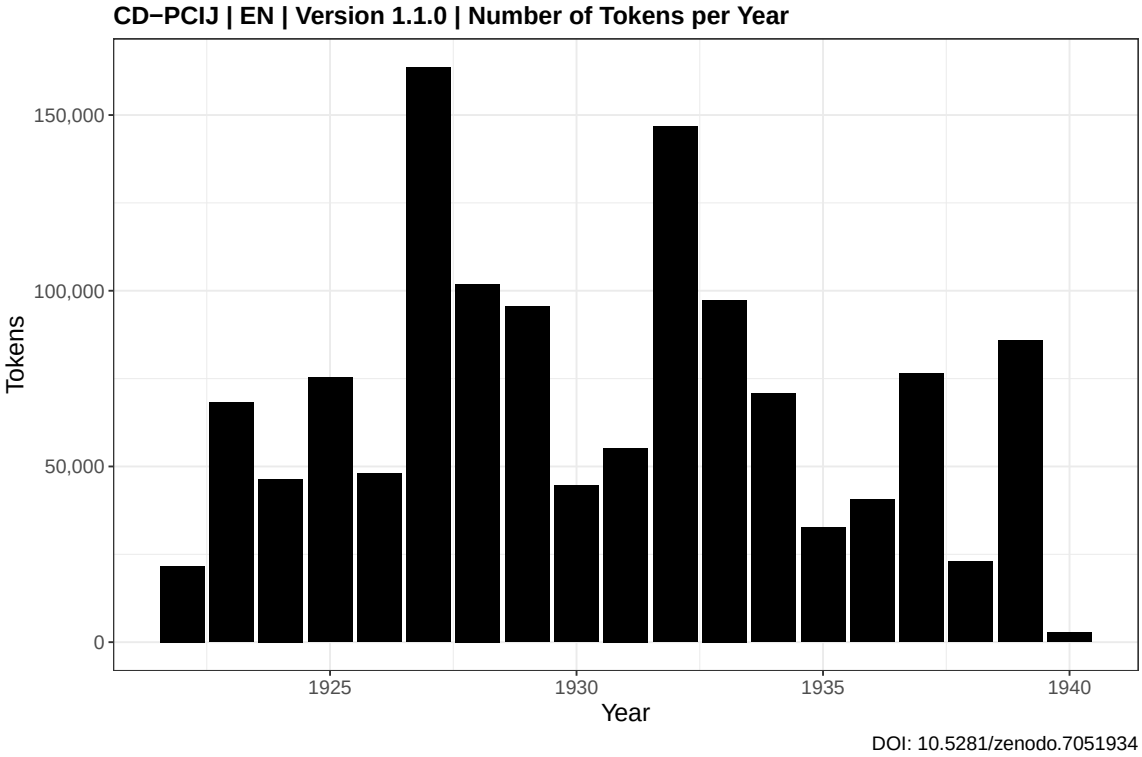
```
fwrite(dt.stats.ling,
       paste0(outputdir,
              datashort,
              "_FR_00_CorpusStatistics_Summaries_Linguistic.csv"),
       na = "NA")
```

22.2 Distributions

22.2.1 Tokens per Year: English

```
tokens.year.en <- meta.tesseract.en[,
                                    sum(ntokens),
                                    by = "year"]
```

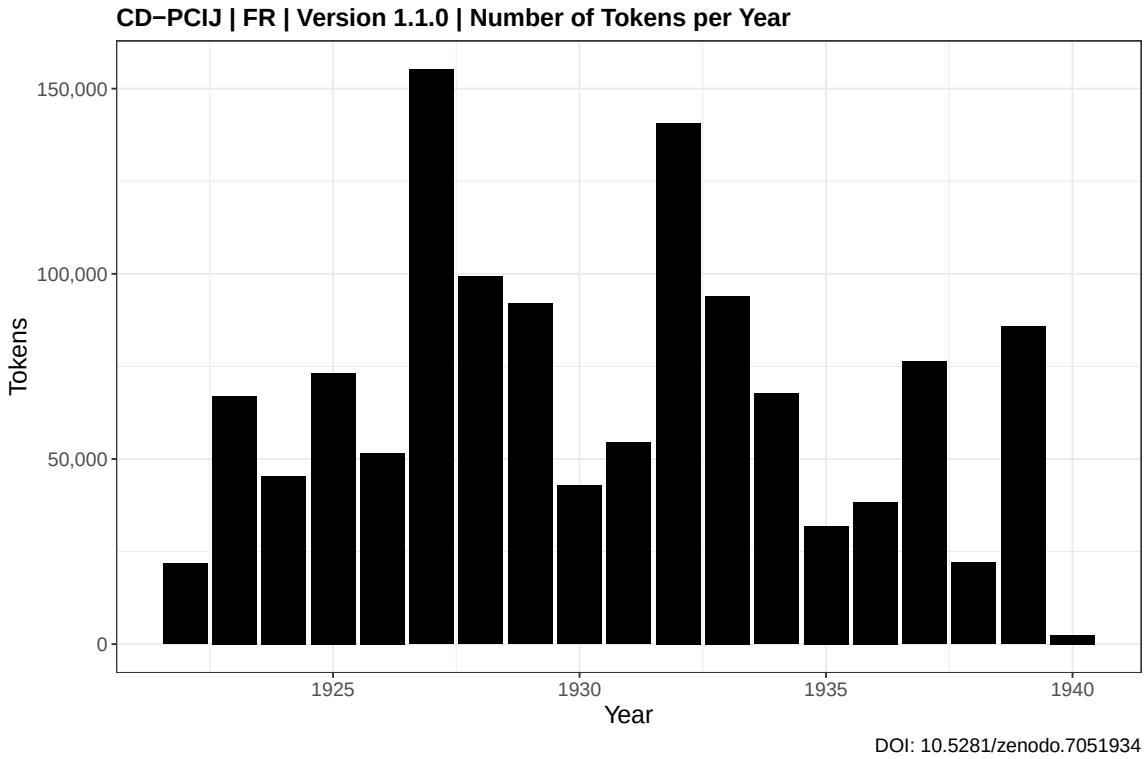
```
print(
    ggplot(data = tokens.year.en,
           aes(x = year,
               y = V1))+
    geom_bar(stat = "identity",
             fill = "black")+
    scale_y_continuous(labels = comma)+
    theme_bw()+
    labs(
        title = paste(datashort,
                      "| EN | Version",
                      version,
                      "| Number of Tokens per Year"),
        caption = paste("DOI:",
                        doi.version),
        x = "Year",
        y = "Tokens"
    )+
    theme(
        text = element_text(size = 14),
        plot.title = element_text(size = 14,
                                  face = "bold")
    )
)
```


CD−PCIJ | EN | Version 1.1.0 | Number of Tokens per Year
150,000
100,000
50,000
0
Tokens
1925
1930
1935
1940
Year
DOI: 10.5281/zenodo.7051934

22.2.2 Tokens per Year: French

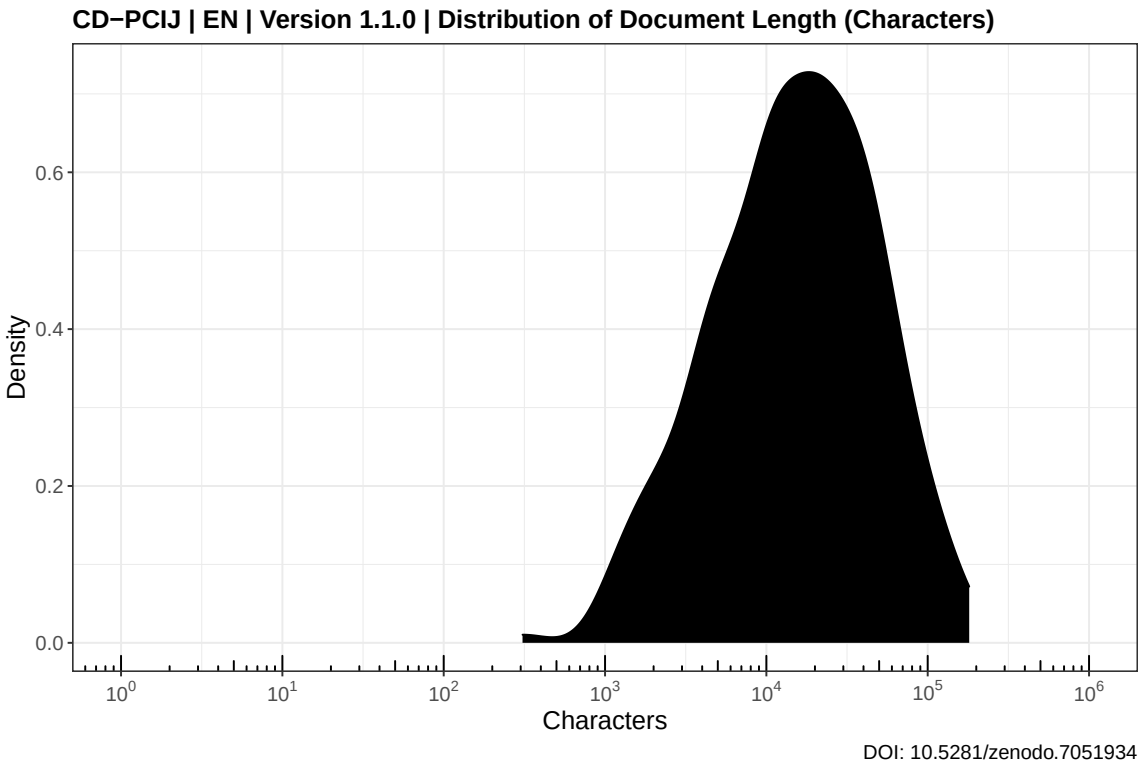
```
tokens.year.fr <- meta.tesseract.fr[,
                                    sum(ntokens),
                                    by = "year"]
```

```
print(
    ggplot(data = tokens.year.fr,
           aes(x = year,
               y = V1))+
    geom_bar(stat = "identity",
             fill = "black")+
    scale_y_continuous(labels = comma)+
    theme_bw()+
    labs(
        title = paste(datashort,
                      "| FR | Version",
                      version,
                      "| Number of Tokens per Year"),
        caption = paste("DOI:",
                        doi.version),
        x = "Year",
        y = "Tokens"
    )+
    theme(
        text = element_text(size = 14),
        plot.title = element_text(size = 14,
                                  face = "bold")
    )
)
```

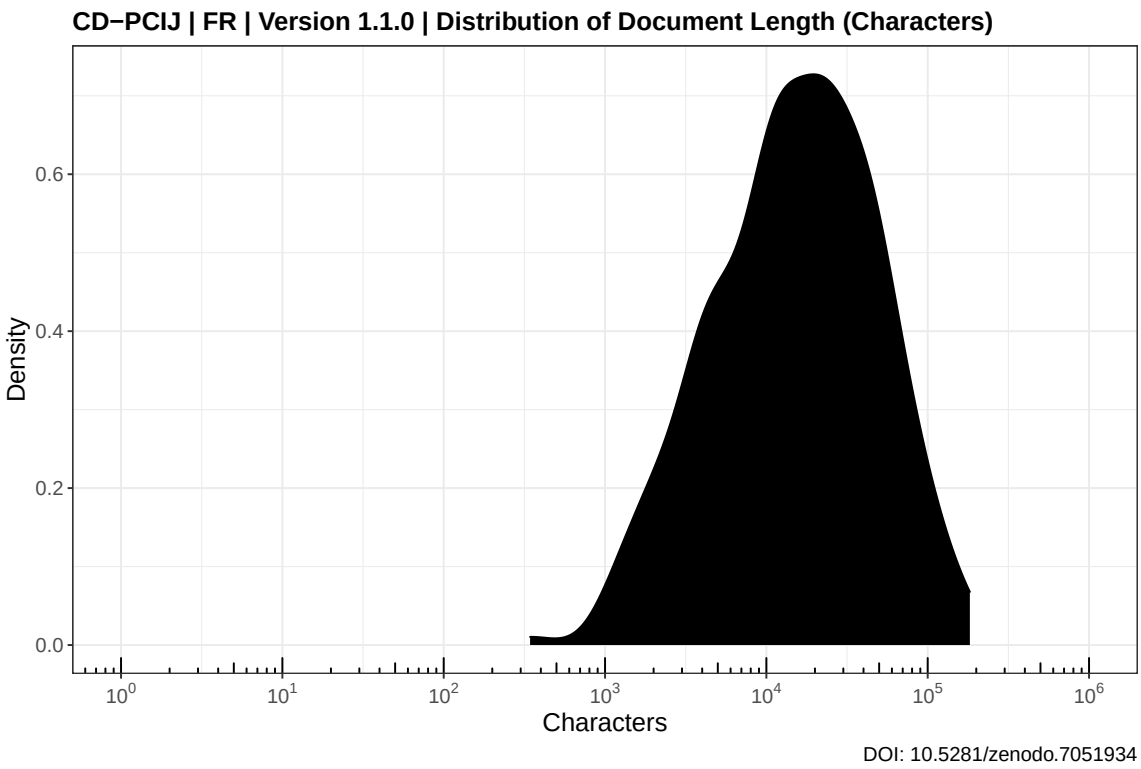

CD−PCIJ | FR | Version 1.1.0 | Number of Tokens per Year
150,000
100,000
50,000
0
Tokens
1925
1930
1935
1940
Year
DOI: 10.5281/zenodo.7051934

22.2.3 Density: Characters

```
ggplot(data = meta.tesseract.en) +
    geom_density(aes(x = nchars),
                 fill = "black") +
    scale_x_log10(breaks = trans_breaks("log10", function(x) 10^x),
                  labels = trans_format("log10", math_format(10^.x)))+
    annotation_logticks(sides = "b")+
    coord_cartesian(xlim = c(1, 10^6))+
    theme_bw() +
    labs(
        title = paste(datashort,
                      "| EN | Version",
                      version,
                      "| Distribution of Document Length (Characters)"),
        caption = paste("DOI:",
                        doi.version),
        x = "Characters",
        y = "Density"
    )+
    theme(
        text = element_text(size = 14),
        plot.title = element_text(size = 14,
                                  face = "bold"),
        legend.position = "none",
        plot.margin = margin(10, 20, 10, 10)
    )
```

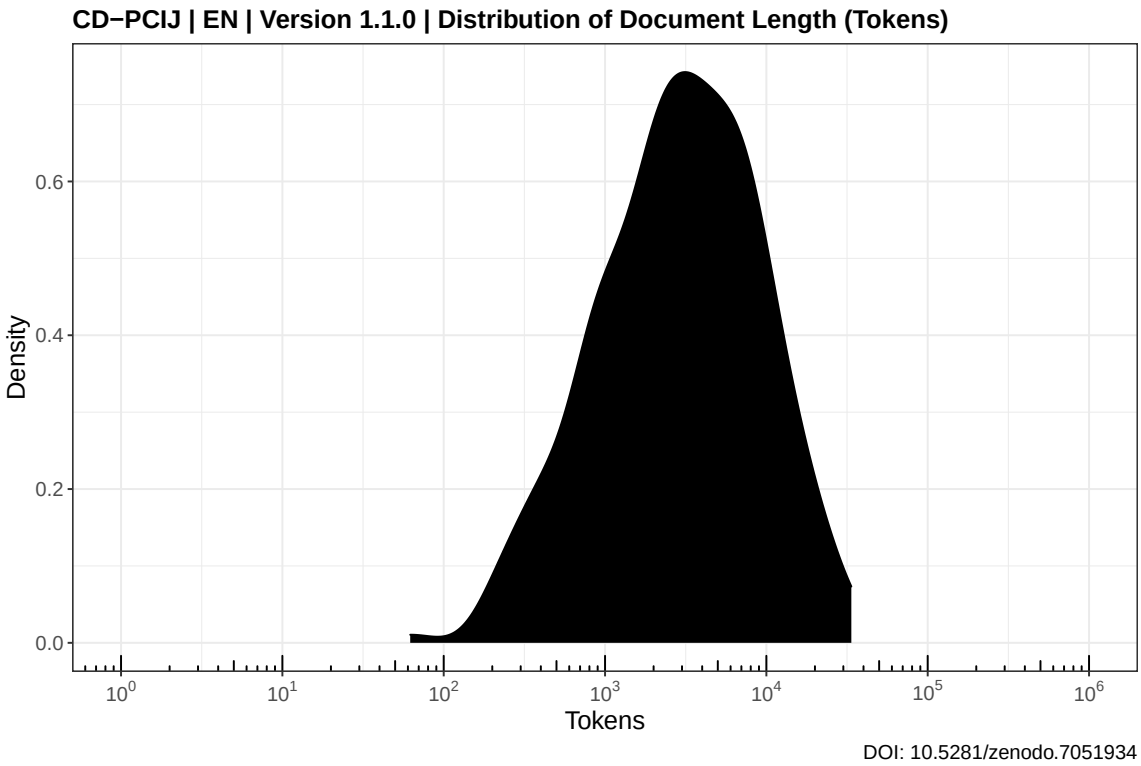


```r
ggplot(data = meta.tesseract.fr) +
    geom_density(aes(x = nchars),
                 fill = "black") +
    scale_x_log10(breaks = trans_breaks("log10", function(x) 10^x),
                  labels = trans_format("log10", math_format(10^.x)))+
    annotation_logticks(sides = "b")+
    coord_cartesian(xlim = c(1, 10^6))+
    theme_bw() +
    labs(
        title = paste(datashort,
                      "| FR | Version",
                      version,
                      "| Distribution of Document Length (Characters)"),
        caption = paste("DOI:",
                        doi.version),
        x = "Characters",
        y = "Density"
    )+
    theme(
        text = element_text(size = 14),
        plot.title = element_text(size = 14,
                                  face = "bold"),
        legend.position = "none",
        plot.margin = margin(10, 20, 10, 10)
    )
```

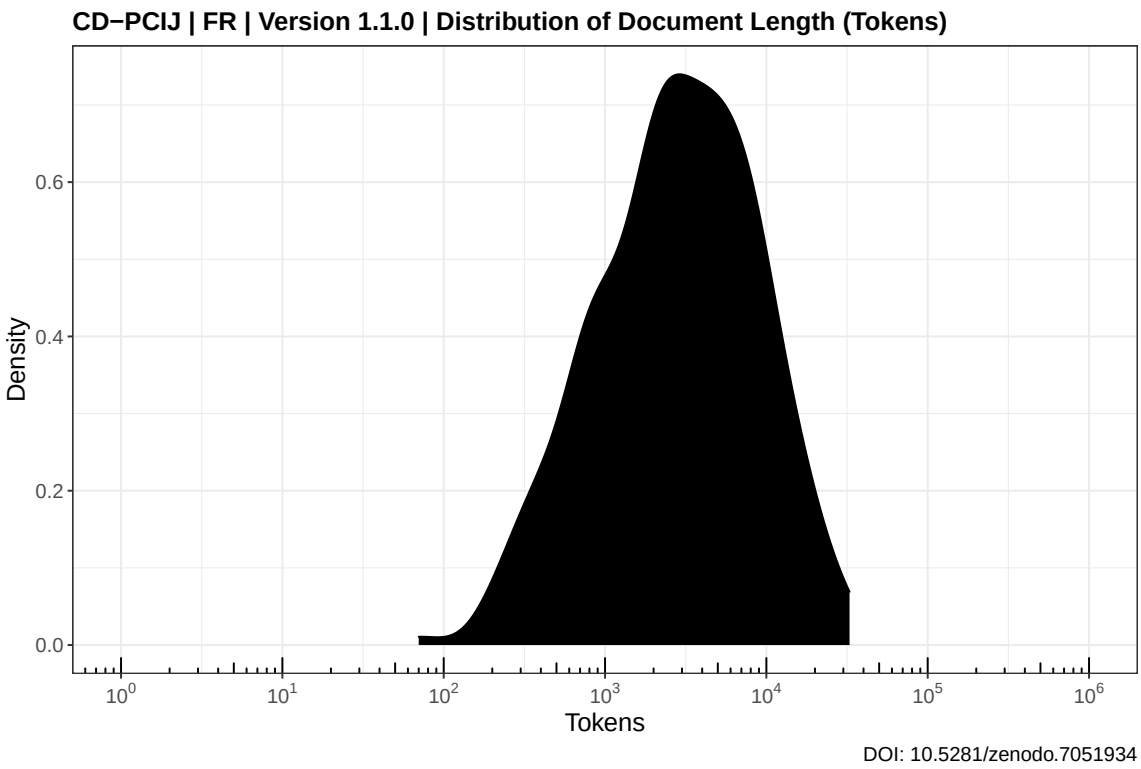


22.2.4 Density: Tokens

```
ggplot(data = meta.tesseract.en) +
    geom_density(aes(x = ntokens),
                 fill = "black") +
    scale_x_log10(breaks = trans_breaks("log10", function(x) 10^x),
                  labels = trans_format("log10", math_format(10^.x)))+
    annotation_logticks(sides = "b")+
    coord_cartesian(xlim = c(1, 10^6))+
    theme_bw() +
    labs(
        title = paste(datashort,
                      "| EN | Version",
                      version,
                      "| Distribution of Document Length (Tokens)"),
        caption = paste("DOI:",
                        doi.version),
        x = "Tokens",
        y = "Density"
    )+
    theme(
        text = element_text(size = 14),
        plot.title = element_text(size = 14,
                                  face = "bold"),
        legend.position = "none",
        plot.margin = margin(10, 20, 10, 10)
    )
```

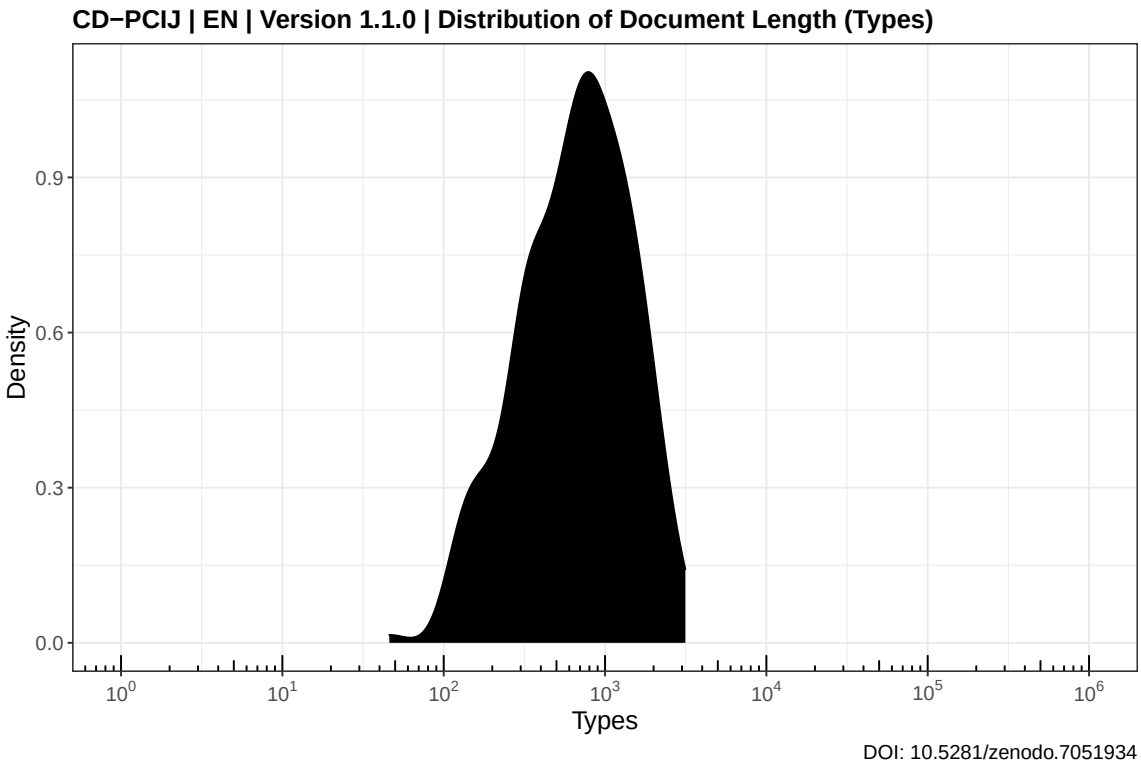


```r
ggplot(data = meta.tesseract.fr) +
    geom_density(aes(x = ntokens),
                 fill = "black") +
    scale_x_log10(breaks = trans_breaks("log10", function(x) 10^x),
                  labels = trans_format("log10", math_format(10^.x)))+
    annotation_logticks(sides = "b")+
    coord_cartesian(xlim = c(1, 10^6))+
    theme_bw() +
    labs(
        title = paste(datashort,
                      "| FR | Version",
                      version,
                      "| Distribution of Document Length (Tokens)"),
        caption = paste("DOI:",
                        doi.version),
        x = "Tokens",
        y = "Density"
    )+
    theme(
        text = element_text(size = 14),
        plot.title = element_text(size = 14,
                                  face = "bold"),
        legend.position = "none",
        plot.margin = margin(10, 20, 10, 10)
    )
```

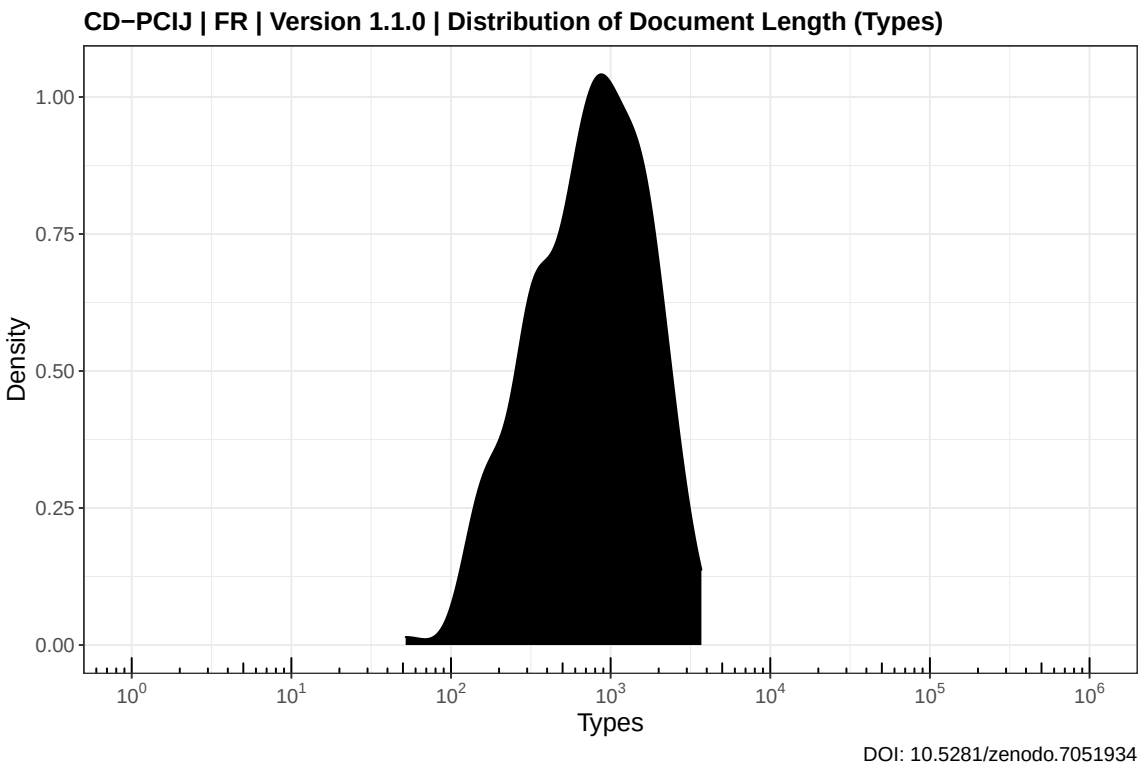


22.2.5 Density: Types

```
ggplot(data = meta.tesseract.en) +
    geom_density(aes(x = ntypes),
                 fill = "black") +
    scale_x_log10(breaks = trans_breaks("log10", function(x) 10^x),
                  labels = trans_format("log10", math_format(10^.x)))+
    annotation_logticks(sides = "b")+
    coord_cartesian(xlim = c(1, 10^6))+
    theme_bw() +
    labs(
        title = paste(datashort,
                      "| EN | Version",
                      version,
                      "| Distribution of Document Length (Types)"),
        caption = paste("DOI:",
                        doi.version),
        x = "Types",
        y = "Density"
    )+
    theme(
        text = element_text(size = 14),
        plot.title = element_text(size = 14,
                                  face = "bold"),
        legend.position = "none",
        plot.margin = margin(10, 20, 10, 10)
    )
```

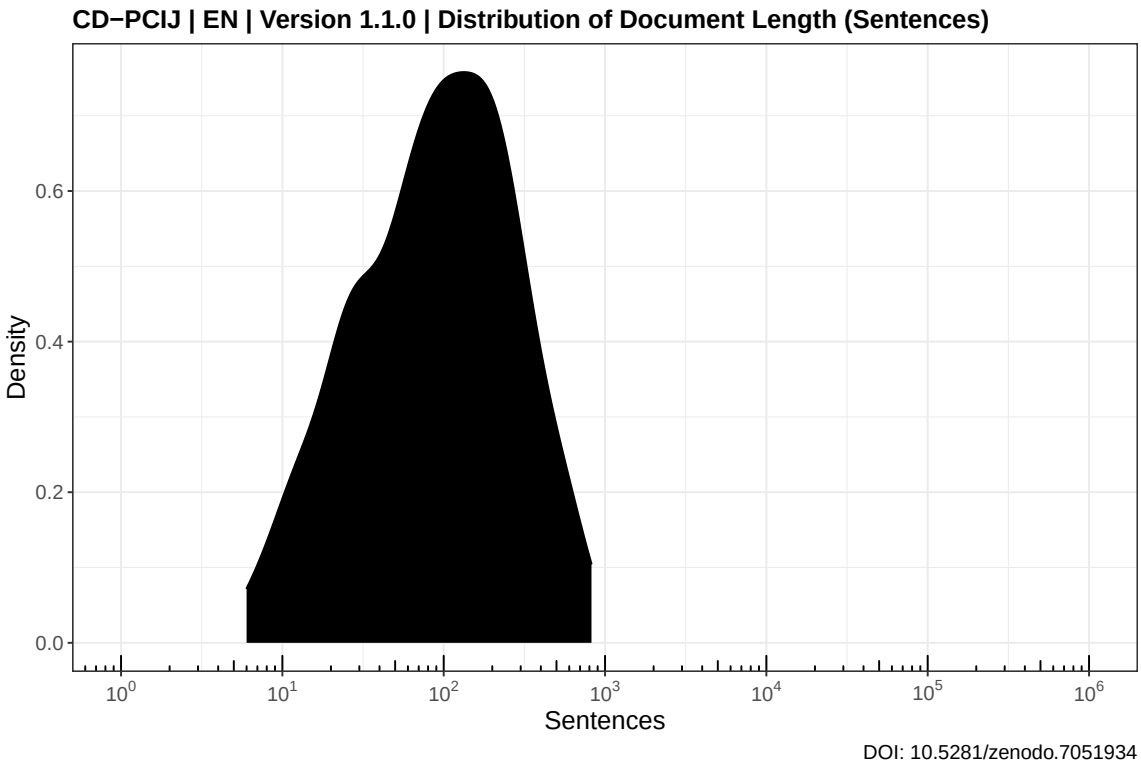


```r
ggplot(data = meta.tesseract.fr) +
    geom_density(aes(x = ntypes),
                 fill = "black") +
    scale_x_log10(breaks = trans_breaks("log10", function(x) 10^x),
                  labels = trans_format("log10", math_format(10^.x)))+
    annotation_logticks(sides = "b")+
    coord_cartesian(xlim = c(1, 10^6))+
    theme_bw() +
    labs(
        title = paste(datashort,
                      "| FR | Version",
                      version,
                      "| Distribution of Document Length (Types)"),
        caption = paste("DOI:",
                        doi.version),
        x = "Types",
        y = "Density"
    )+
    theme(
        text = element_text(size = 14),
        plot.title = element_text(size = 14,
                                  face = "bold"),
        legend.position = "none",
        plot.margin = margin(10, 20, 10, 10)
    )
```

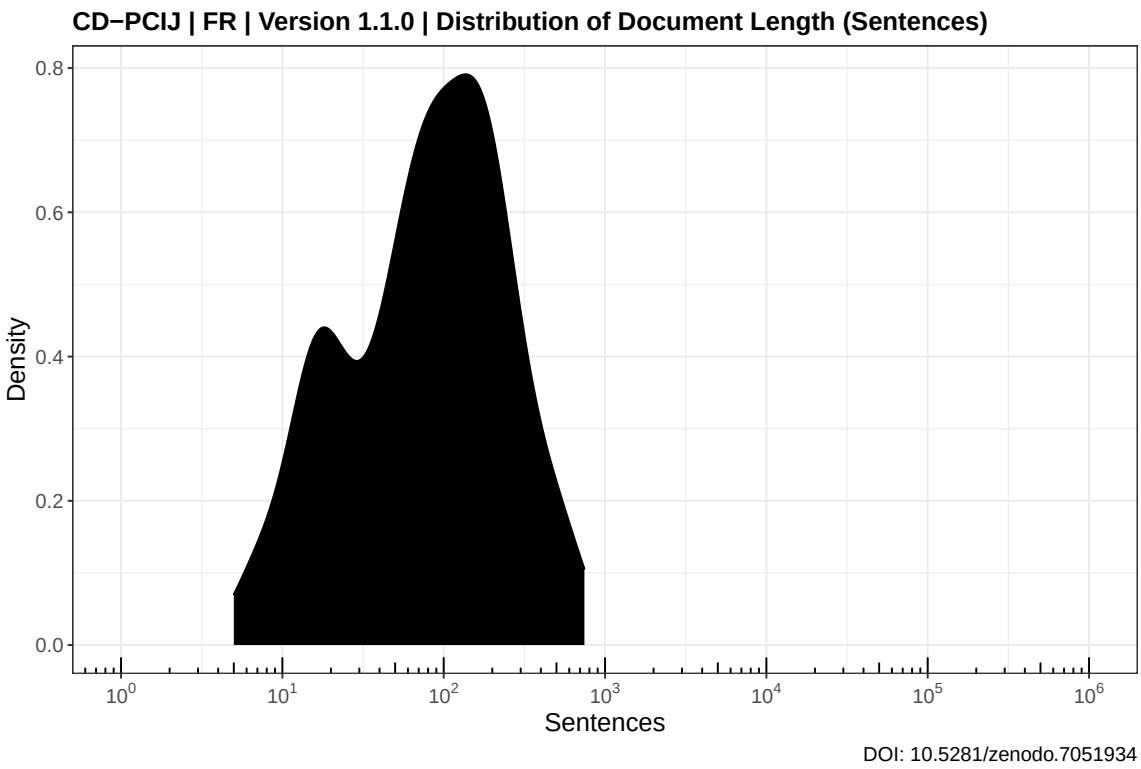


22.2.6 Density: Sentences

```
ggplot(data = meta.tesseract.en) +
    geom_density(aes(x = nsentences),
                 fill = "black") +
    scale_x_log10(breaks = trans_breaks("log10", function(x) 10^x),
                  labels = trans_format("log10", math_format(10^.x)))+
    annotation_logticks(sides = "b")+
    coord_cartesian(xlim = c(1, 10^6))+
    theme_bw() +
    labs(
        title = paste(datashort,
                      "| EN | Version",
                      version,
                      "| Distribution of Document Length (Sentences)"),
        caption = paste("DOI:",
                        doi.version),
        x = "Sentences",
        y = "Density"
    )+
    theme(
        text = element_text(size = 14),
        plot.title = element_text(size = 14,
                                  face = "bold"),
        legend.position = "none",
        plot.margin = margin(10, 20, 10, 10)
    )
```



```r
ggplot(data = meta.tesseract.fr) +
    geom_density(aes(x = nsentences),
                 fill = "black") +
    scale_x_log10(breaks = trans_breaks("log10", function(x) 10^x),
                  labels = trans_format("log10", math_format(10^.x)))+
    annotation_logticks(sides = "b")+
    coord_cartesian(xlim = c(1, 10^6))+
    theme_bw() +
    labs(
        title = paste(datashort,
                      "| FR | Version",
                      version,
                      "| Distribution of Document Length (Sentences)"),
        caption = paste("DOI:",
                        doi.version),
        x = "Sentences",
        y = "Density"
    )+
    theme(
        text = element_text(size = 14),
        plot.title = element_text(size = 14,
                                  face = "bold"),
        legend.position = "none",
        plot.margin = margin(10, 20, 10, 10)
    )
```



22.2.7 All Distributions of Linguistic Metrics

When plotting a boxplot on a logarithmic scale the standard geom_boxplot() function from ggplot2 incorrectly performs the statistical transformation first before calculating the boxplot statistics. While median and quartiles are based on ordinal position the inter-quartile range differs depending on when statistical transformation is performed.

Solutions are based on this SO question: https://stackoverflow.com/questions/38753628/ggplot-boxplot-length-of-whiskers-with-logarithmic-axis

```
print(f.boxplot.body)
```

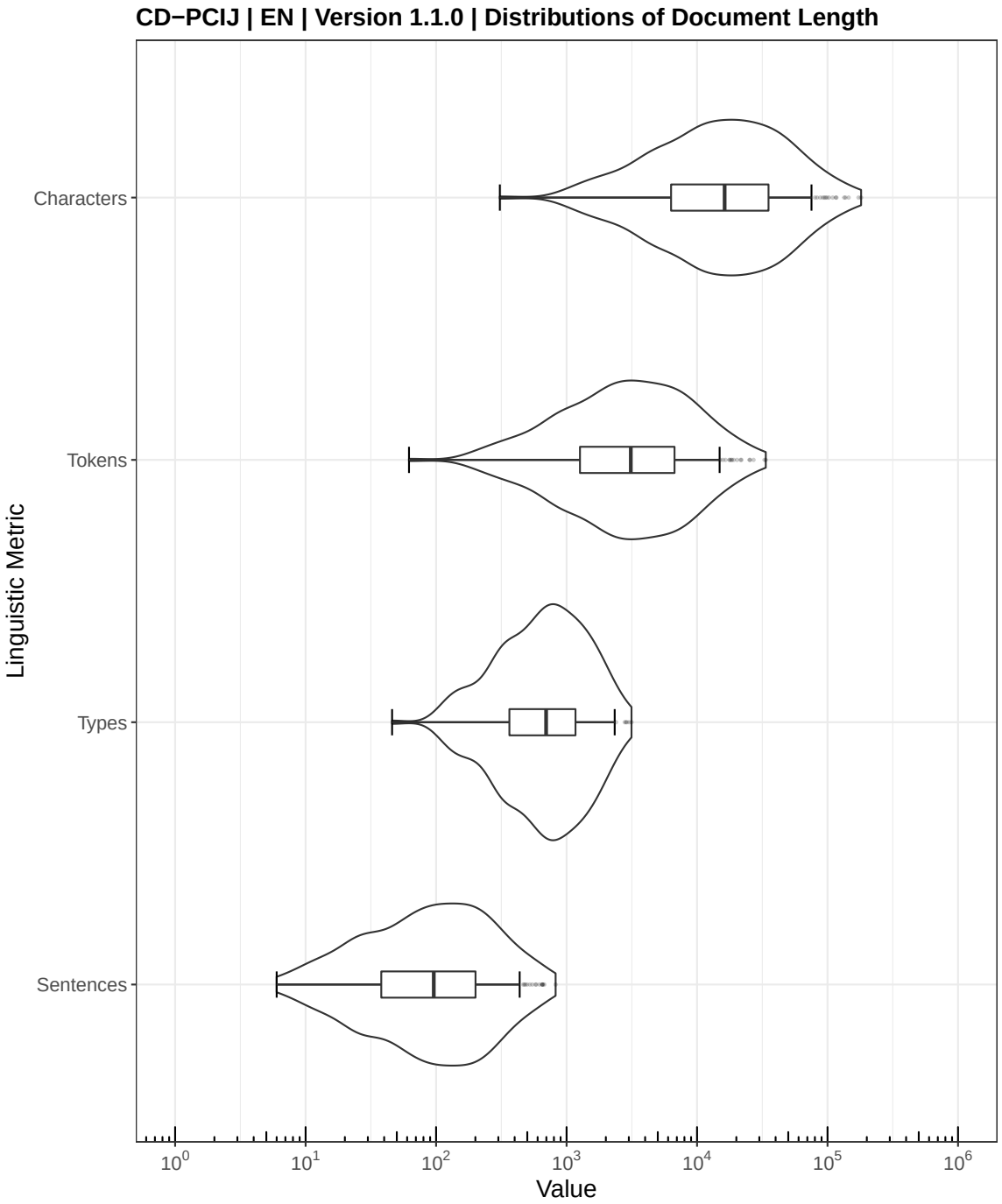
```
## function(x) {
##
##     body = log10(boxplot.stats(10^x)[["stats"]])
##
##     names(body) = c("ymin",
##                     "lower",
##                     "middle",
##                     "upper",
##                     "ymax")
##
##     return(body)
##
## }
```

```
print(f.boxplot.outliers)
```

```
## function(x) {
##
##     data.frame(y = log10(boxplot.stats(10^x)[["out"]]))
##
## }
```

```
dt.allmetrics.en <- melt(summary.corpus.en,
                         measure.vars = rev(c("nchars",
                                              "ntokens",
                                              "ntypes",
                                              "nsentences")))
```

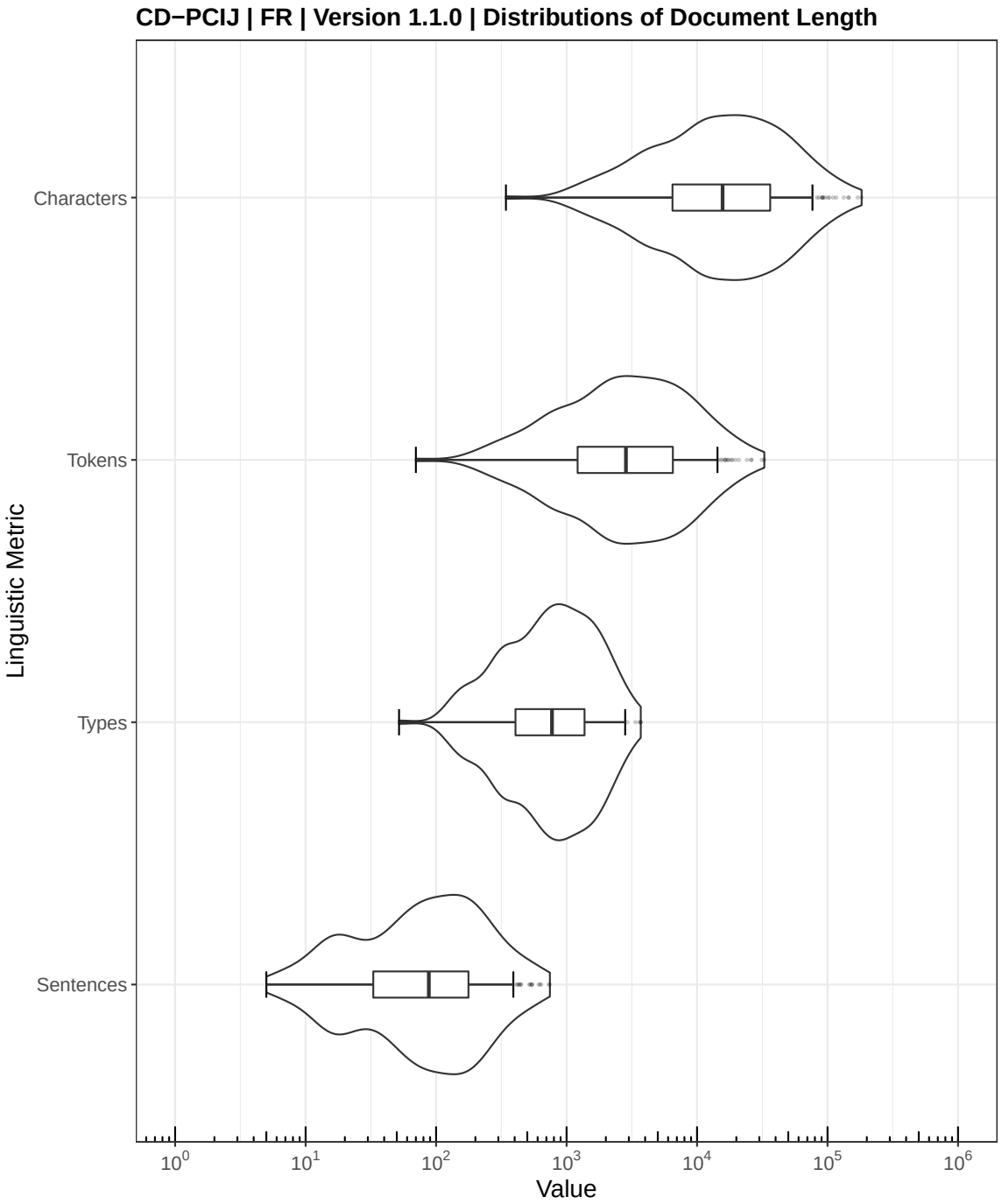
```
ggplot(dt.allmetrics.en, aes(x = value,
                             y = variable)) +
    geom_violin()+
    stat_summary(fun.data = f.boxplot.body,
                 geom = "errorbar",
                 width = 0.1) +
    stat_summary(fun.data = f.boxplot.body,
                 geom = "boxplot",
                 width = 0.1) +
    stat_summary(fun.data = f.boxplot.outliers,
                 geom = "point",
                 size =  0.5,
                 alpha = 0.2)+
    scale_x_log10(breaks = trans_breaks("log10", function(x) 10^x),
                  labels = trans_format("log10", math_format(10^.x)))+
    annotation_logticks(sides = "b")+
    coord_cartesian(xlim = c(1, 10^6))+
    scale_y_discrete(labels = rev(c("Characters",
                                    "Tokens",
                                    "Types",
                                    "Sentences")))+
    theme_bw() +
    labs(
        title = paste(datashort,
                      "| EN | Version",
                      version,
                      "| Distributions of Document Length"),
        caption = paste("DOI:",
                        doi.version),
        x = "Value",
        y = "Linguistic Metric"
    )+
    theme(
        text = element_text(size = 14),
        plot.title = element_text(size = 14,
                                  face = "bold"),
        legend.position = "none",
        plot.margin = margin(10, 20, 10, 10)
    )
```



DOI: 10.5281/zenodo.7051934

```r
dt.allmetrics.fr <- melt(summary.corpus.fr,
                         measure.vars = rev(c("nchars",
                                              "ntokens",
                                              "ntypes",
                                              "nsentences")))
```

```r
ggplot(dt.allmetrics.fr, aes(x = value,
                             y = variable)) +
    geom_violin()+
    stat_summary(fun.data = f.boxplot.body,
                 geom = "errorbar",
                 width = 0.1) +
    stat_summary(fun.data = f.boxplot.body,
                 geom = "boxplot",
                 width = 0.1) +
    stat_summary(fun.data = f.boxplot.outliers,
                 geom = "point",
                 size =  0.5,
                 alpha = 0.2)+
    scale_x_log10(breaks = trans_breaks("log10", function(x) 10^x),
                  labels = trans_format("log10", math_format(10^.x)))+
    annotation_logticks(sides = "b")+
    coord_cartesian(xlim = c(1, 10^6))+
    scale_y_discrete(labels = rev(c("Characters",
                                    "Tokens",
                                    "Types",
                                    "Sentences")))+
    theme_bw() +
    labs(
        title = paste(datashort,
                      "| FR | Version",
                      version,
                      "| Distributions of Document Length"),
        caption = paste("DOI:",
                        doi.version),
        x = "Value",
        y = "Linguistic Metric"
    )+
    theme(
        text = element_text(size = 14),
        plot.title = element_text(size = 14,
                                  face = "bold"),
        legend.position = "none",
        plot.margin = margin(10, 20, 10, 10)
    )
```


CD−PCIJ | FR | Version 1.1.0 | Distributions of Document Length
Characters
Tokens
Types
Sentences
Linguistic Metric
10^0
10^1
10^2
10^3
10^4
10^5
10^6
Value
DOI: 10.5281/zenodo.7051934

22.3 Number of Majority Opinions

22.3.1 English

```
dt.maj.disaggregated <- meta.tesseract.en[opinion == 0,
                                          .N,
                                          keyby = "doctype"]

sumrow <- data.table("Total",
                     sum(dt.maj.disaggregated$N))

dt.maj.disaggregated <- rbind(dt.maj.disaggregated,
                              sumrow,
                              use.names = FALSE)



kable(dt.maj.disaggregated,
      format = "latex",
      booktabs = TRUE,
      longtable = TRUE)
```

doctype	N
ADV	27
DEC	1
JUD	32
ORD	40
Total	100

```
fwrite(dt.maj.disaggregated,
       paste0(outputdir,
              datashort,
              "_EN_00_CorpusStatistics_Summaries_Majority.csv"),
       na = "NA")
```

22.3.2 French

```
dt.maj.disaggregated <- meta.tesseract.fr[opinion == 0,
                                          .N,
                                          keyby = "doctype"]

sumrow <- data.table("Total",
                     sum(dt.maj.disaggregated$N))

dt.maj.disaggregated <- rbind(dt.maj.disaggregated,
                              sumrow,
                              use.names=FALSE)


kable(dt.maj.disaggregated,
      format = "latex",
      booktabs = TRUE,
      longtable = TRUE)
```

doctype	N
ADV	27
DEC	1
JUD	32
ORD	40
Total	100

```
fwrite(dt.maj.disaggregated,
       paste0(outputdir,
              datashort,
              "_FR_00_CorpusStatistics_Summaries_Majority.csv"),
       na = "NA")
```

22.4 Number of Minority Opinions

22.4.1 English

```
dt.min.disaggregated <- meta.tesseract.en[opinion > 0,
                                          .N,
                                          keyby = "doctype"]

sumrow <- data.table("Total",
                     sum(dt.min.disaggregated$N))

dt.min.disaggregated <- rbind(dt.min.disaggregated,
                              sumrow,
                              use.names = FALSE)



kable(dt.min.disaggregated,
      format = "latex",
      booktabs = TRUE,
      longtable = TRUE)
```

doctype	N
ADV	17
JUD	81
ORD	11
Total	109

```
fwrite(dt.min.disaggregated,
       paste0(outputdir,
              datashort,
              "_EN_00_CorpusStatistics_Summaries_Minority.csv"),
       na = "NA")
```

22.4.2 French

```
dt.min.disaggregated <- meta.tesseract.fr[opinion > 0,
                                          .N,
                                          keyby = "doctype"]

sumrow <- data.table("Total",
                     sum(dt.min.disaggregated$N))

dt.min.disaggregated <- rbind(dt.min.disaggregated,
                              sumrow,
                              use.names = FALSE)


kable(dt.min.disaggregated,
      format = "latex",
      booktabs = TRUE,
      longtable = TRUE)
```

doctype	N
ADV	17
JUD	81
ORD	11
Total	109

```
fwrite(dt.min.disaggregated,
       paste0(outputdir,
              datashort,
              "_FR_00_CorpusStatistics_Summaries_Minority.csv"),
       na = "NA")
```

22.5 Year Range

```
summary(meta.tesseract.en$year) # English
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1922    1927    1931    1931    1934    1940
```

```
summary(meta.tesseract.fr$year) # French
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1922    1927    1931    1931    1934    1940
```

22.6 Date Range

```r
meta.tesseract.en$date <- as.Date(meta.tesseract.en$date)
meta.tesseract.fr$date <- as.Date(meta.tesseract.fr$date)

summary(meta.tesseract.en$date) # English
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## "1922-05-22" "1927-10-12" "1931-05-15" "1931-05-02" "1934-12-12" "1940-02-26"
```

```r
summary(meta.tesseract.fr$date) # French
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## "1922-05-22" "1927-10-14" "1931-05-15" "1931-05-08" "1934-12-12" "1940-02-26"
```

23 Test and Sort Variable Names

23.1 Semantic Sorting of Variable Names

This step ensures that all variable names documented in the Codebook are present in the data set and sorted according to the order in the Codebook. Where variables are missing in the data or undocumented variables are present this step will throw an error.

23.1.1 Sort Variables: Full Data Set

```
setcolorder(data.tesseract.en, # English
            c("doc_id",
              "text",
              "court",
              "series",
              "seriesno",
              "caseno",
              "shortname",
              "fullname",
              "applicant",
              "respondent",
              "applicant_region",
              "respondent_region",
              "applicant_subregion",
              "respondent_subregion",
              "date",
              "doctype",
              "collision",
              "stage",
              "opinion",
              "language",
              "year",
              "minority",
              "nchars",
              "ntokens",
              "ntypes",
              "nsentences",
              "version",
              "doi_concept",
              "doi_version",
              "license"))
```

```r
setcolorder(data.tesseract.fr, # French
            c("doc_id",
              "text",
              "court",
              "series",
              "seriesno",
              "caseno",
              "shortname",
              "fullname",
              "applicant",
              "respondent",
              "applicant_region",
              "respondent_region",
              "applicant_subregion",
              "respondent_subregion",
              "date",
              "doctype",
              "collision",
              "stage",
              "opinion",
              "language",
              "year",
              "minority",
              "nchars",
              "ntokens",
              "ntypes",
              "nsentences",
              "version",
              "doi_concept",
              "doi_version",
              "license"))
```

23.1.2 Sort Variables: Metadata

```
setcolorder(meta.tesseract.en, # English
            c("doc_id",
              "court",
              "series",
              "seriesno",
              "caseno",
              "shortname",
              "fullname",
              "applicant",
              "respondent",
              "applicant_region",
              "respondent_region",
              "applicant_subregion",
              "respondent_subregion",
              "date",
              "doctype",
              "collision",
              "stage",
              "opinion",
              "language",
              "year",
              "minority",
              "nchars",
              "ntokens",
              "ntypes",
              "nsentences",
              "version",
              "doi_concept",
              "doi_version",
              "license"))
```

```r
setcolorder(meta.tesseract.fr, # French
            c("doc_id",
              "court",
              "series",
              "seriesno",
              "caseno",
              "shortname",
              "fullname",
              "applicant",
              "respondent",
              "applicant_region",
              "respondent_region",
              "applicant_subregion",
              "respondent_subregion",
              "date",
              "doctype",
              "collision",
              "stage",
              "opinion",
              "language",
              "year",
              "minority",
              "nchars",
              "ntokens",
              "ntypes",
              "nsentences",
              "version",
              "doi_concept",
              "doi_version",
              "license"))
```

23.2 Number of Variables: Full Data Set

```
length(data.tesseract.en) # English
```

```
## [1] 30
```

```
length(data.tesseract.fr) # French
```

```
## [1] 30
```

23.3 Number of Variables: Metadata

```
length(meta.tesseract.en) # English
```

```
## [1] 29
```

```
length(meta.tesseract.fr) # French
```

```
## [1] 29
```

23.4 List All Variables: Full Data Set

"doc_id" is the filename, "text" is the extracted plaintext, third variable onwards are the metadata variables ("docvars").

```
names(data.tesseract.en) # English
```

```
##  [1] "doc_id"               "text"                 "court"
##  [4] "series"               "seriesno"             "caseno"
##  [7] "shortname"            "fullname"             "applicant"
## [10] "respondent"           "applicant_region"     "respondent_region"
## [13] "applicant_subregion"  "respondent_subregion" "date"
## [16] "doctype"              "collision"            "stage"
## [19] "opinion"              "language"             "year"
## [22] "minority"             "nchars"               "ntokens"
## [25] "ntypes"               "nsentences"           "version"
## [28] "doi_concept"          "doi_version"          "license"
```

```r
names(data.tesseract.fr) # French
```

```
##  [1] "doc_id"               "text"                 "court"
##  [4] "series"               "seriesno"             "caseno"
##  [7] "shortname"            "fullname"             "applicant"
## [10] "respondent"           "applicant_region"     "respondent_region"
## [13] "applicant_subregion"  "respondent_subregion" "date"
## [16] "doctype"              "collision"            "stage"
## [19] "opinion"              "language"             "year"
## [22] "minority"             "nchars"               "ntokens"
## [25] "ntypes"               "nsentences"           "version"
## [28] "doi_concept"          "doi_version"          "license"
```

23.5 List All Variables: Metadata

```r
names(meta.tesseract.en) # English
```

```
##  [1] "doc_id"               "court"                "series"
##  [4] "seriesno"             "caseno"               "shortname"
##  [7] "fullname"             "applicant"            "respondent"
## [10] "applicant_region"     "respondent_region"    "applicant_subregion"
## [13] "respondent_subregion" "date"                 "doctype"
## [16] "collision"            "stage"                "opinion"
## [19] "language"             "year"                 "minority"
## [22] "nchars"               "ntokens"              "ntypes"
## [25] "nsentences"           "version"              "doi_concept"
## [28] "doi_version"          "license"
```

```r
names(meta.tesseract.fr) # French
```

```
##  [1] "doc_id"               "court"                "series"
##  [4] "seriesno"             "caseno"               "shortname"
##  [7] "fullname"             "applicant"            "respondent"
## [10] "applicant_region"     "respondent_region"    "applicant_subregion"
## [13] "respondent_subregion" "date"                 "doctype"
## [16] "collision"            "stage"                "opinion"
## [19] "language"             "year"                 "minority"
## [22] "nchars"               "ntokens"              "ntypes"
## [25] "nsentences"           "version"              "doi_concept"
## [28] "doi_version"          "license"
```

24 Calculate Detailed Token Frequencies

24.1 Create Corpora

```
corpus.en.b <- corpus(data.tesseract.en)
corpus.fr.b <- corpus(data.tesseract.fr)
```

24.2 Process Tokens

```
quanteda_options(tokens_locale = "en") # Set Locale for Tokenization
tokens.en <- f.token.processor(corpus.en.b)

quanteda_options(tokens_locale = "fr") # Set Locale for Tokenization
tokens.fr <- f.token.processor(corpus.fr.b)
```

24.3 Construct Document-Feature-Matrices

```
dfm.en <- dfm(tokens.en)
dfm.fr <- dfm(tokens.fr)

dfm.tfidf.en <- dfm_tfidf(dfm.en)
dfm.tfidf.fr <- dfm_tfidf(dfm.fr)
```

24.4 Most Frequent Tokens | TF Weighting | Tables

24.4.1 English

```
tstat.en <- textstat_frequency(dfm.en,
                               n = 100)

fwrite(tstat.en, paste0(outputdir,
                        datashort,
                        "_EN_11_Top100Tokens_TF-Weighting.csv"))

kable(tstat.en,
      format = "latex",
      align = "r",
      booktabs = TRUE,
      longtable = TRUE,
      col.names = c("Feature",
                    "Frequency",
                    "Rank",
                    "Docfreq",
                    "Group")) %>% kable_styling(latex_options = "repeat_header")
```

Feature	Frequency	Rank	Docfreq	Group
court	7586	1	245	all
article	6654	2	232	all
government	5986	3	229	all
treaty	3545	4	187	all
question	3390	5	213	all
case	3011	6	221	all
opinion	2764	7	209	all
law	2710	8	193	all
may	2690	9	230	all
international	2518	10	214	all
judgment	2396	11	186	all
polish	2335	12	77	all
agreement	2084	13	173	all
convention	2028	14	134	all
regard	2014	15	215	all
jurisdiction	1969	16	143	all
parties	1856	17	197	all
upon	1791	18	203	all
french	1739	19	163	all
two	1702	20	204	all
must	1679	21	187	all
paragraph	1665	22	169	all
one	1651	23	213	all
council	1649	24	124	all
commission	1604	25	95	all
application	1598	26	195	all
made	1583	27	209	all
state	1463	28	174	all
decision	1452	29	181	all
present	1442	30	199	all

(continued)

Feature	Frequency	Rank	Docfreq	Group
league	1386	31	147	all
part	1379	32	185	all
powers	1378	33	153	all
also	1375	34	191	all
german	1373	35	108	all
fact	1366	36	180	all
right	1360	37	168	all
nations	1354	38	162	all
special	1336	39	162	all
whether	1324	40	179	all
shall	1302	41	154	all
first	1282	42	183	all
free	1244	43	109	all
provisions	1244	43	170	all
order	1194	45	193	all
certain	1141	46	181	all
states	1134	47	166	all
proceedings	1129	48	180	all
view	1121	49	181	all
rights	1119	50	149	all
danzig	1094	51	29	all
july	1078	52	147	all
therefore	1073	53	172	all
belgian	1073	53	48	all
dispute	1058	55	144	all
poland	1043	56	69	all
b	1036	57	156	all
letter	1035	58	136	all
statute	1032	59	148	all

(continued)

Feature	Frequency	Rank	Docfreq	Group
time	1019	60	166	all
within	1011	61	160	all
point	1007	62	159	all
president	999	63	143	all
company	998	64	115	all
regards	987	65	155	all
following	982	66	188	all
interpretation	982	66	150	all
given	970	68	185	all
terms	964	69	168	all
general	956	70	181	all
legal	953	71	164	all
said	950	72	178	all
greek	948	73	42	all
articles	927	74	158	all
conditions	923	75	161	all
submitted	922	76	193	all
concerning	911	77	191	all
effect	904	78	175	all
territory	902	79	110	all
minister	895	80	103	all
taken	863	81	162	all
dated	861	82	144	all
subject	856	83	164	all
follows	850	84	166	all
place	831	85	158	all
versailles	826	86	79	all
paris	823	87	92	all
public	822	88	132	all

(continued)

Feature	Frequency	Rank	Docfreq	Group
march	819	89	149	all
note	811	90	132	all
necessary	802	91	167	all
june	800	92	157	all
can	794	93	160	all
governments	793	94	141	all
signed	789	95	240	all
questions	785	96	143	all
according	779	97	156	all
whereas	775	98	112	all
november	762	99	121	all
justice	759	100	161	all

24.4.2 French

```
tstat.fr <- textstat_frequency(dfm.fr,
                               n = 100)

fwrite(tstat.fr, paste0(outputdir,
                        datashort,
                        "_FR_11_Top100Tokens_TF-Weighting.csv"))

kable(tstat.fr,
      format = "latex",
      align = "r",
      booktabs = TRUE,
      longtable = TRUE,
      col.names = c("Feature",
                    "Frequency",
                    "Rank",
                    "Docfreq",
                    "Group")) %>% kable_styling(latex_options = "repeat_header")
```

Feature	Frequency	Rank	Docfreq	Group
cour	7654	1	248	all
gouvernement	6499	2	231	all
l'article	5287	3	224	all

(continued)

Feature	Frequency	Rank	Docfreq	Group
traité	3323	4	187	all
droit	3269	5	204	all
comme	2957	6	208	all
entre	2823	7	228	all
être	2805	8	188	all
question	2711	9	200	all
si	2689	10	197	all
fait	2617	11	215	all
d'une	2295	12	214	all
qu'il	2288	13	206	all
société	2175	14	184	all
convention	2119	15	146	all
dont	2089	16	213	all
deux	2018	17	212	all
conseil	1951	18	166	all
ainsi	1929	19	213	all
parties	1868	20	199	all
d'un	1849	21	206	all
cas	1754	22	188	all
commission	1714	23	100	all
plus	1705	24	178	all
polonais	1564	25	71	all
peut	1510	26	174	all
partie	1441	27	180	all
tout	1440	28	183	all
point	1426	29	174	all
nations	1362	30	161	all
premier	1349	31	169	all
c'est	1320	32	167	all

(continued)

Feature	Frequency	Rank	Docfreq	Group
décision	1304	33	168	all
compétence	1295	34	132	all
date	1288	35	181	all
lieu	1287	36	184	all
n'est	1281	37	170	all
arrêt	1239	38	135	all
pologne	1214	39	74	all
doit	1202	40	167	all
où	1201	41	173	all
non	1187	42	178	all
statut	1185	43	162	all
dispositions	1156	44	172	all
part	1133	45	174	all
droits	1124	46	155	all
vue	1111	47	180	all
avis	1110	48	153	all
juillet	1074	49	150	all
lettre	1073	50	138	all
bien	1073	50	166	all
sous	1064	52	179	all
requête	1059	53	148	all
termes	1055	54	181	all
article	1051	55	165	all
faire	1050	56	175	all
b	1040	57	157	all
président	1038	58	153	all
n'a	1036	59	167	all
alinéa	1030	60	137	all
autres	1018	61	193	all

(continued)

Feature	Frequency	Rank	Docfreq	Group
international	991	62	140	all
laquelle	971	63	175	all
loi	971	63	115	all
général	968	65	178	all
savoir	962	66	174	all
internationale	960	67	181	all
devant	960	67	176	all
demande	960	67	168	all
conclusion	953	70	144	all
articles	951	71	165	all
règlement	935	72	163	all
pays	926	73	124	all
belge	925	74	44	all
territoire	920	75	111	all
français	918	76	148	all
conditions	902	77	172	all
opinion	897	78	155	all
procédure	893	79	174	all
qu'elle	891	80	161	all
ministre	887	81	106	all
suisse	862	82	31	all
donc	860	83	149	all
toute	850	84	167	all
mai	849	85	155	all
texte	849	85	179	all
france	845	87	82	all
vertu	840	88	155	all
versailles	834	89	81	all
libre	831	90	96	all

(continued)

Feature	Frequency	Rank	Docfreq	Group
paris	830	91	92	all
mars	827	92	150	all
tous	824	93	153	all
dantzig	818	94	30	all
selon	814	95	140	all
compromis	812	96	86	all
juin	807	97	158	all
sens	806	98	139	all
contre	803	99	168	all
déclaration	800	100	128	all

24.5 Most Frequent Tokens | TFIDF Weighting | Tables

24.5.1 English

```
tstat.tfidf.en <- textstat_frequency(dfm.tfidf.en,
                                     n = 100,
                                     force = TRUE)

fwrite(tstat.en, paste0(outputdir,
                        datashort,
                        "_EN_12_Top100Tokens_TFIDF-Weighting.csv"))

kable(tstat.tfidf.en,
      format = "latex",
      align = "r",
      booktabs = TRUE,
      longtable = TRUE,
      col.names = c("Feature",
                    "Weight",
                    "Rank",
                    "Docfreq",
                    "Group")) %>% kable_styling(latex_options = "repeat_header")
```

Feature	Weight	Rank	Docfreq	Group
polish	1230.0991	1	77	all
danish	1085.5386	2	9	all
danzig	1040.2865	3	29	all

(continued)

Feature	Weight	Rank	Docfreq	Group
greenland	1035.0131	4	7	all
belgian	785.4988	5	48	all
danube	771.8413	6	9	all
greek	748.9678	7	42	all
commission	698.6642	8	95	all
zones	669.3398	9	18	all
european	652.4661	10	21	all
poland	599.1521	11	69	all
meuse	582.9410	12	8	all
convention	580.4034	13	134	all
water	566.2708	14	21	all
norwegian	563.9066	15	10	all
hungarian	541.7457	16	17	all
gold	537.3246	17	18	all
council	527.4790	18	124	all
concessions	523.8016	19	34	all
city	521.6272	20	40	all
german	521.5698	21	108	all
jurisdiction	507.9306	22	143	all
treaty	501.4692	23	187	all
contract	498.9855	24	57	all
bulgarian	498.6238	25	21	all
mavrommatis	492.0773	26	21	all
free	467.5863	27	109	all
netherlands	466.0410	28	20	all
memel	462.5211	29	9	all
switzerland	460.2285	30	26	all
labour	459.3449	31	21	all
turkish	443.5559	32	34	all

(continued)

Feature	Weight	Rank	Docfreq	Group
property	433.3720	33	67	all
canal	431.6772	34	16	all
protocol	431.2651	35	70	all
bonds	428.1920	36	17	all
versailles	425.9456	37	79	all
galatz	425.1386	38	6	all
mixed	420.3506	39	56	all
minority	419.8389	40	35	all
lithuanian	417.1670	41	16	all
swiss	415.3016	42	24	all
crete	409.3007	43	7	all
factory	408.0099	44	23	all
upper	405.7182	45	56	all
concession	384.2004	46	35	all
university	379.8329	47	18	all
customs	377.8435	48	41	all
navigation	374.5078	49	36	all
paris	369.9483	50	92	all
frontier	369.0661	51	28	all
directorate	368.6881	52	15	all
oberschlesische	365.9366	53	18	all
agreement	365.2286	54	173	all
gex	358.8779	55	12	all
minister	358.4140	56	103	all
geneva	356.3298	57	79	all
schools	354.8553	58	17	all
france	354.6358	59	76	all
conference	353.9140	60	87	all
company	351.8967	61	115	all

(continued)

Feature	Weight	Rank	Docfreq	Group
french	349.7340	62	163	all
law	346.1821	63	193	all
payment	345.1481	64	57	all
lighthouses	344.8451	65	10	all
judgment	344.5132	66	186	all
czechoslovak	344.3294	67	13	all
sovereignty	342.1203	68	62	all
league	340.9316	69	147	all
turkey	340.9081	70	38	all
port	340.4293	71	26	all
territory	335.4602	72	110	all
silesia	333.8137	73	37	all
mandate	330.1760	74	33	all
norway	329.2988	75	10	all
albanian	327.8344	76	4	all
lock	327.7542	77	7	all
maestricht	326.1860	78	7	all
agrarian	324.0602	79	16	all
government	320.0372	80	229	all
british	319.3449	81	54	all
denmark	319.1998	82	15	all
ottoman	318.1473	83	27	all
article	318.1396	84	232	all
palestine	316.8051	85	16	all
powers	315.0223	86	153	all
nationals	310.4821	87	93	all
loans	310.3513	88	18	all
contracts	309.3048	89	50	all
paragraph	308.7127	90	169	all

(continued)

Feature	Weight	Rank	Docfreq	Group
yugoslav	308.0993	91	10	all
greece	307.5159	92	39	all
braila	304.1376	93	6	all
administration	302.6999	94	79	all
francs	302.5374	95	20	all
arbitral	299.8986	96	62	all
territories	298.0958	97	64	all
loan	297.6130	98	18	all
shall	293.9643	99	154	all
savoy	293.5881	100	11	all

24.5.2 French

```
tstat.tfidf.fr <- textstat_frequency(dfm.tfidf.fr,
                                     n = 100,
                                     force = TRUE)

fwrite(tstat.fr, paste0(outputdir,
                        datashort,
                        "_FR_12_Top100Tokens_TFIDF-Weighting.csv"))

kable(tstat.tfidf.fr,
      format = "latex",
      align = "r",
      booktabs = TRUE,
      longtable = TRUE,
      col.names = c("Feature",
                    "Weight",
                    "Rank",
                    "Docfreq",
                    "Group")) %>% kable_styling(latex_options = "repeat_header")
```

Feature	Weight	Rank	Docfreq	Group
polonais	884.2577	1	71	all
européenne	814.5557	2	9	all
suisse	797.5903	3	31	all
dantzig	768.5267	4	30	all

(continued)

Feature	Weight	Rank	Docfreq	Group
danube	719.4998	5	9	all
belge	715.1987	6	44	all
commission	714.1218	7	100	all
groënland	705.6226	8	7	all
pologne	664.5543	9	74	all
zones	635.8980	10	20	all
danois	618.5944	11	9	all
meuse	582.0344	12	9	all
convention	534.5975	13	146	all
ville	517.0702	14	52	all
travail	516.0492	15	34	all
concessions	513.9452	16	35	all
hongrois	509.3546	17	13	all
phares	492.9909	18	10	all
mavrommatis	489.2063	19	21	all
hellénique	488.7409	20	39	all
traité	481.1667	21	187	all
contrat	475.3003	22	59	all
danemark	475.1304	23	15	all
memel	448.9562	24	9	all
concession	443.0741	25	36	all
protocole	439.6993	26	69	all
directoire	432.8754	27	8	all
franches	431.7111	28	15	all
france	424.8885	29	82	all
versailles	423.8017	30	81	all
canal	421.0980	31	17	all
polonaise	419.8922	32	67	all
bulgare	411.1243	33	18	all

(continued)

Feature	Weight	Rank	Docfreq	Group
être	399.6639	34	188	all
galatz	398.1529	35	6	all
mixte	395.6225	36	55	all
biens	394.9522	37	51	all
compromis	391.4993	38	86	all
conseil	383.4348	39	166	all
compétence	383.4062	40	132	all
allemand	381.0073	41	85	all
paris	375.8677	42	92	all
navigation	375.0099	43	37	all
belgique	372.7390	44	52	all
norvège	367.8013	45	12	all
d'eau	365.5077	46	32	all
souveraineté	362.9724	47	63	all
libre	360.9609	48	96	all
arrêt	354.7340	49	135	all
norvégien	353.9137	50	6	all
crète	353.5971	51	7	all
gex	351.7518	52	12	all
l'article	351.0171	53	224	all
turquie	350.2696	54	40	all
droit	349.8168	55	204	all
conférence	349.3379	56	85	all
ministre	347.1138	57	106	all
consultatif	346.5896	58	61	all
loi	345.6203	59	115	all
gouvernement	344.6324	60	231	all
territoire	341.6121	61	111	all
genève	339.9538	62	79	all

(continued)

Feature	Weight	Rank	Docfreq	Group
frontière	330.5935	63	28	all
société	330.2143	64	184	all
grèce	330.0507	65	42	all
si	328.5266	66	197	all
maestricht	328.4524	67	7	all
tribunal	322.0535	68	85	all
langue	320.1301	69	40	all
écoles	319.3762	70	18	all
minorités	318.8349	71	33	all
haut-commissaire	315.2886	72	27	all
port	315.0818	73	28	all
question	313.4201	74	200	all
territoires	311.4296	75	69	all
contrats	310.9182	76	49	all
emprunts	310.6431	77	19	all
paiement	307.5841	78	46	all
tchécoslovaque	306.1935	79	14	all
ressortissants	303.8062	80	94	all
britannique	303.2824	81	60	all
haute-silésie	303.0540	82	41	all
mandat	301.4462	83	41	all
roumanie	300.0992	84	20	all
l'accord	299.8117	85	81	all
l'autriche	299.6329	86	18	all
pays	299.3006	87	124	all
dissidente	297.3458	88	80	all
lettre	296.9650	89	138	all
pays-bas	296.5881	90	21	all
chinn	295.4500	91	7	all

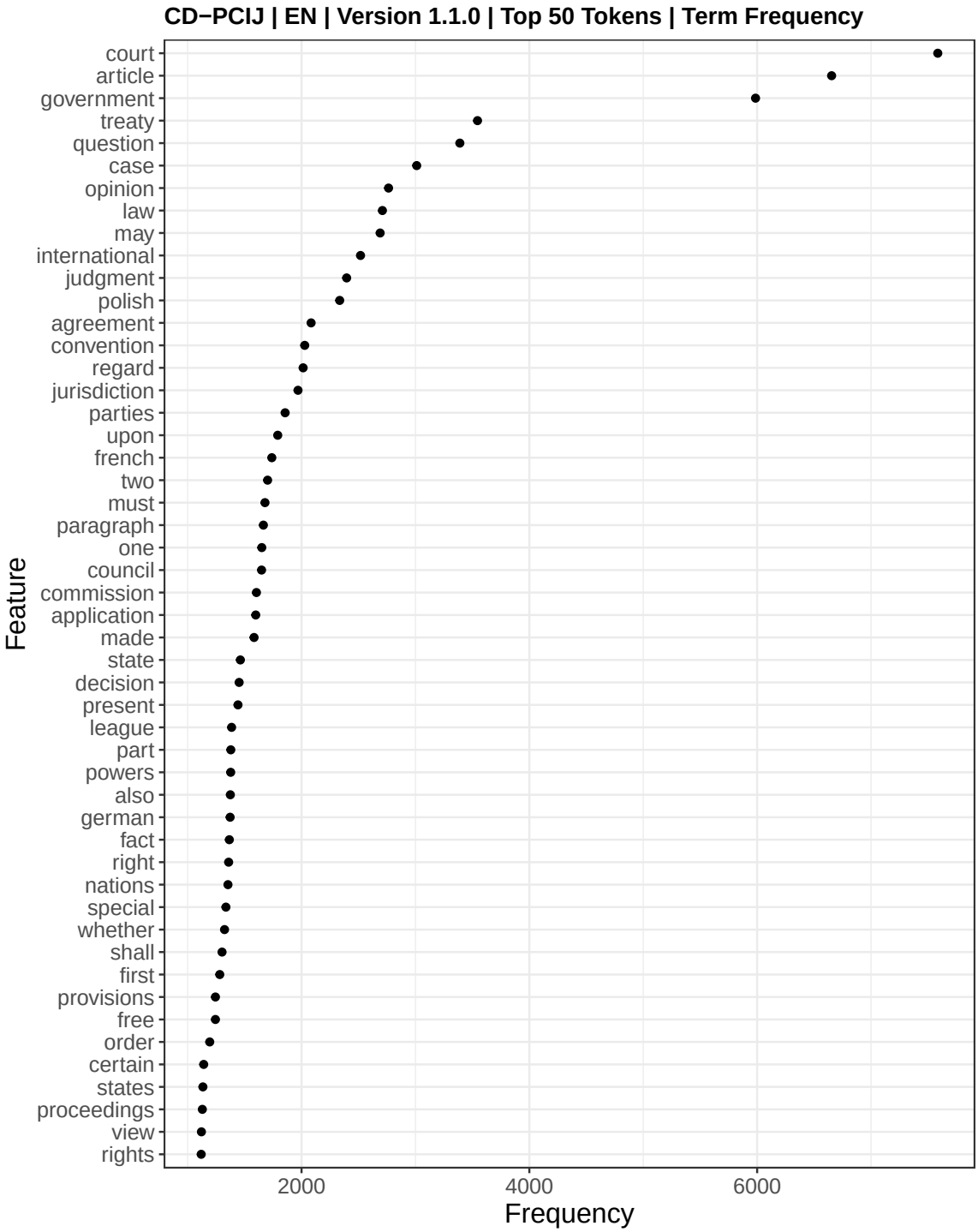
(continued)

Feature	Weight	Rank	Docfreq	Group
service	293.5595	92	61	all
groénland	293.2896	93	6	all
juridiction	292.7184	94	94	all
comme	291.4927	95	208	all
l'usine	288.9431	96	20	all
alinéa	288.3175	97	137	all
note	287.4126	98	108	all
lithuanien	286.5669	99	15	all
nations	285.7675	100	161	all

24.6 Most Frequent Tokens | TF Weighting | Scatterplots

24.6.1 English

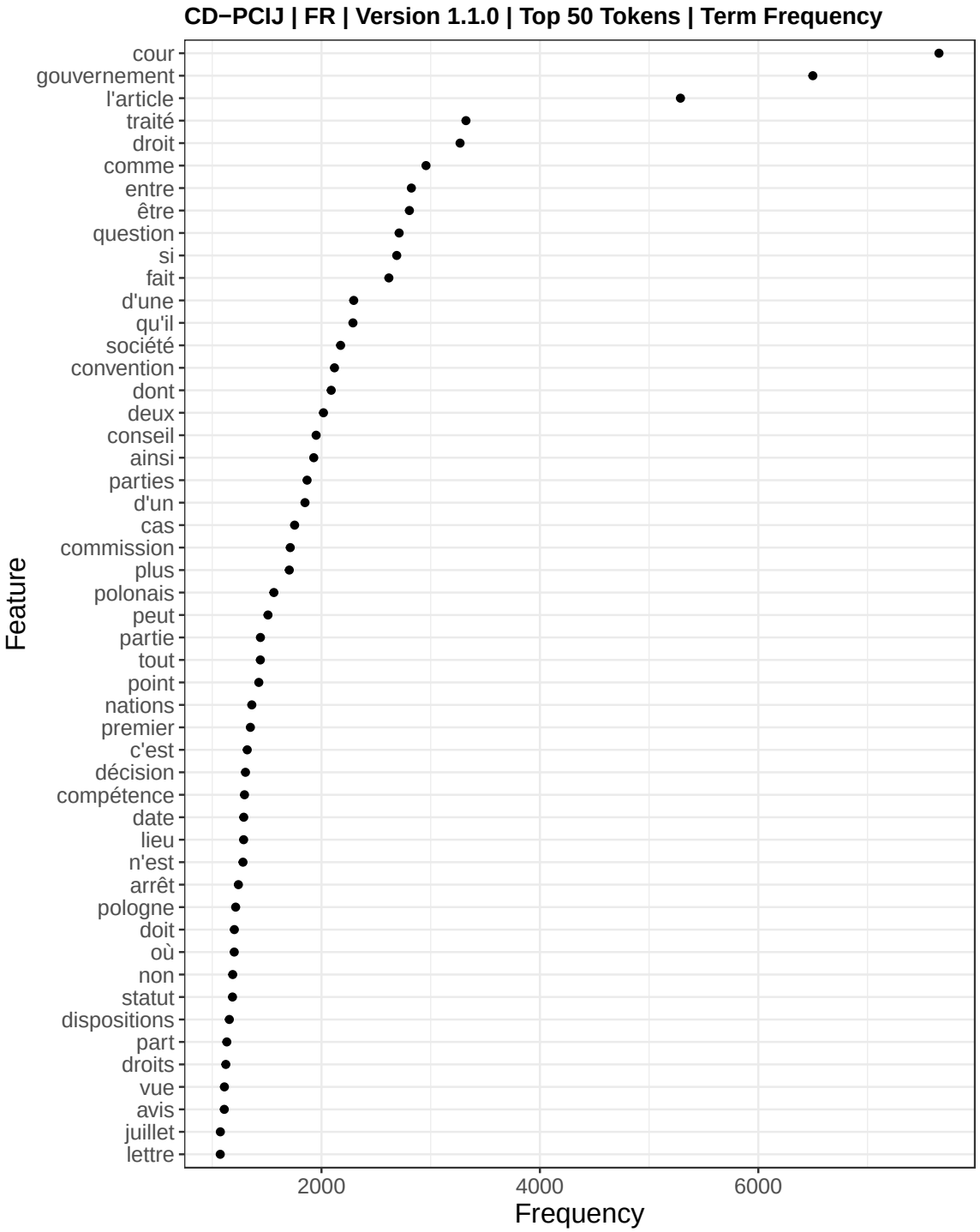
```
print(
    ggplot(data = tstat.en[1:50, ],
           aes(x = reorder(feature,
                           frequency),
               y = frequency))+
    geom_point()+
    coord_flip()+
    labs(
        title = paste(datashort,
                      "| EN | Version",
                      version,
                      "| Top 50 Tokens | Term Frequency"),
        caption = paste("DOI:",
                        doi.version),
        x = "Feature",
        y = "Frequency"
    )+
    theme_bw()+
    theme(
        text = element_text(size = 14),
        plot.title = element_text(size = 12,
                                  face = "bold")
    )
)
```



DOI: 10.5281/zenodo.7051934

24.6.2 French

```
print(
    ggplot(data = tstat.fr[1:50, ],
           aes(x = reorder(feature,
                           frequency),
               y = frequency))+
    geom_point()+
    coord_flip()+
    theme_bw()+
    labs(
        title = paste(datashort,
                      "| FR | Version",
                      version,
                      "| Top 50 Tokens | Term Frequency"),
        caption = paste("DOI:",
                        doi.version),
        x = "Feature",
        y = "Frequency"
    )+
    theme(
        text = element_text(size = 14),
        plot.title = element_text(size = 12,
                                  face = "bold")
    )
)
```

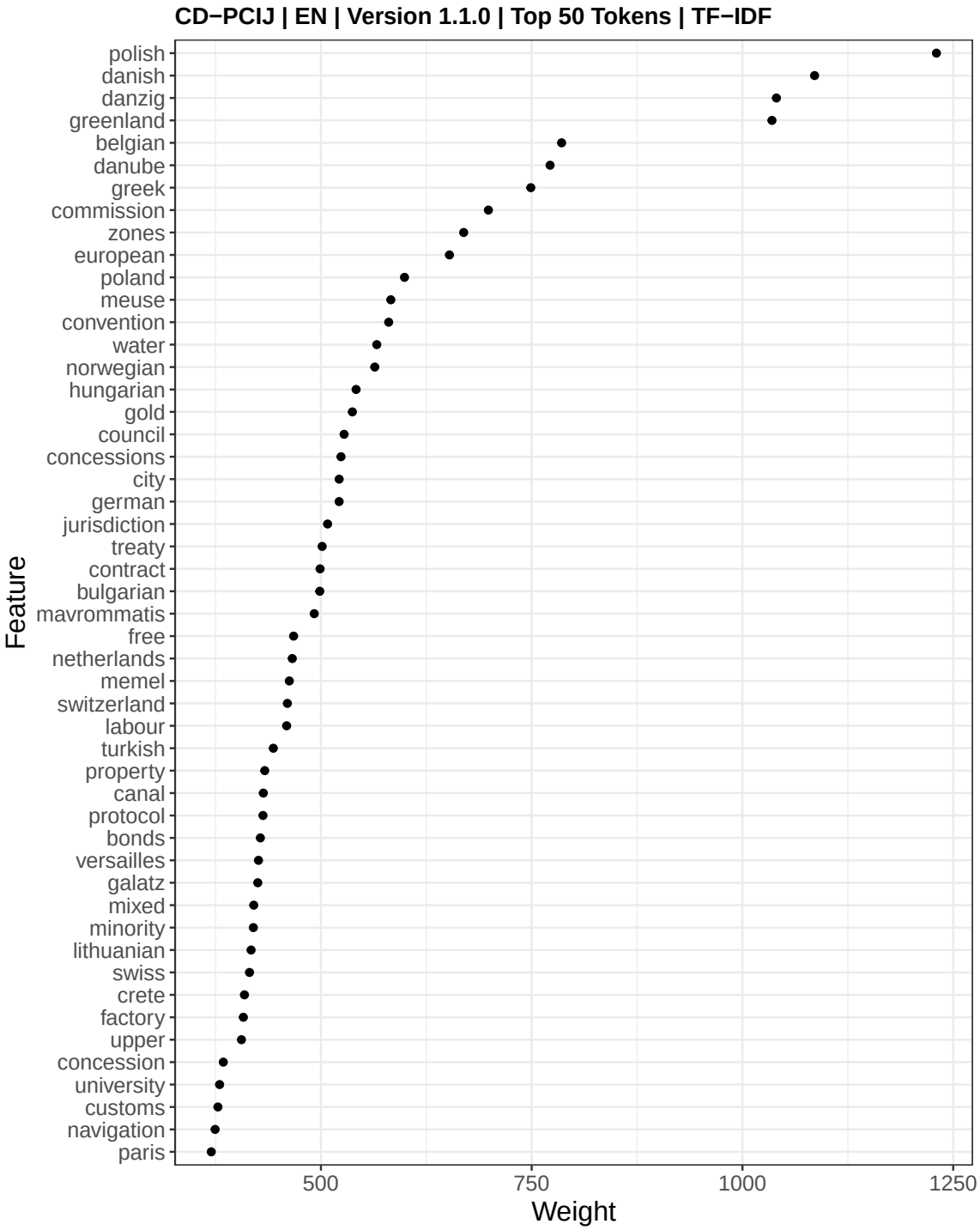


DOI: 10.5281/zenodo.7051934

24.7 Most Frequent Tokens | TFIDF Weighting | Scatterplots

24.7.1 English

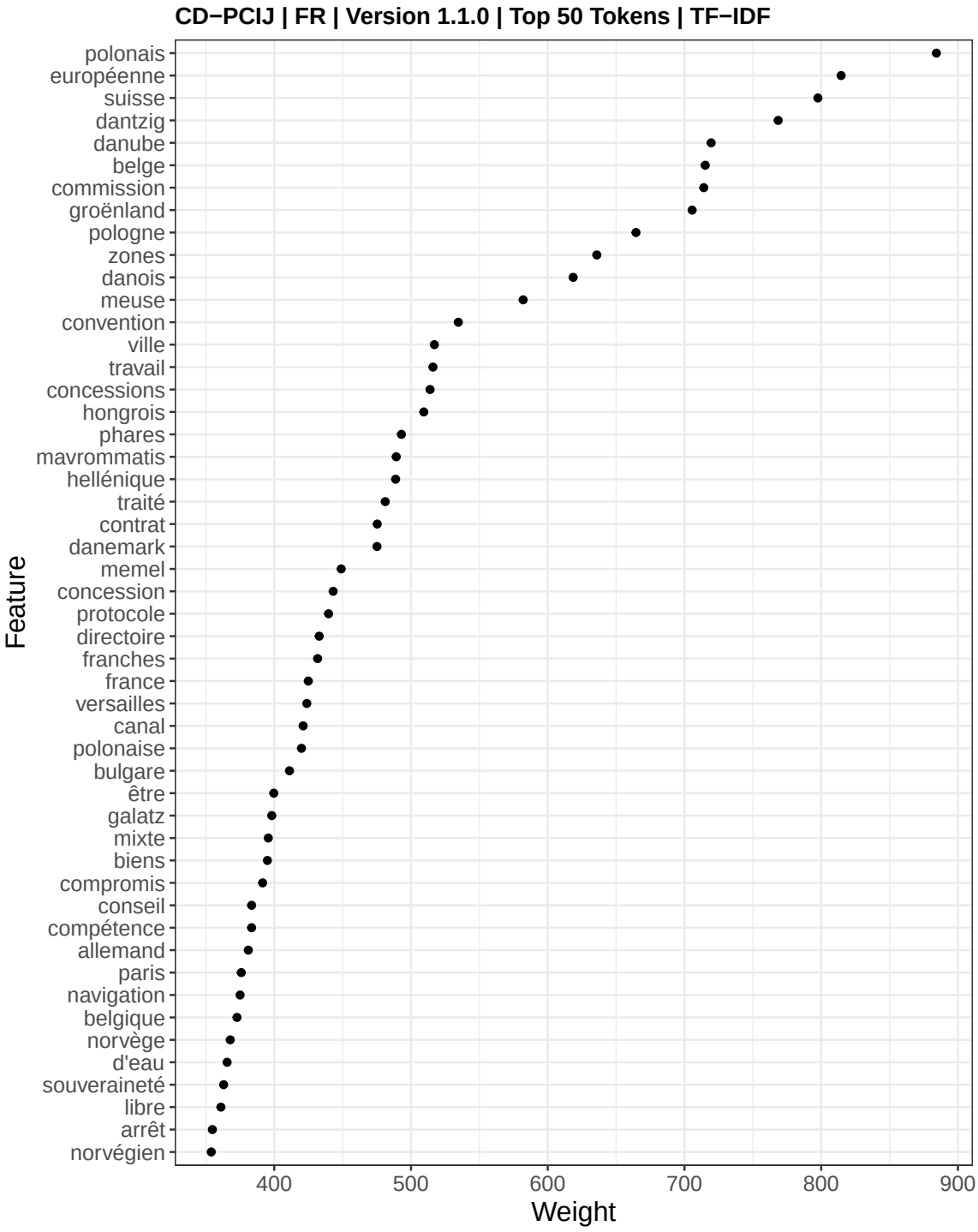
```
print(
    ggplot(data = tstat.tfidf.en[1:50, ],
           aes(x = reorder(feature,
                           frequency),
               y = frequency))+
    geom_point()+
    coord_flip()+
    theme_bw()+
    labs(
        title = paste(datashort,
                      "| EN | Version",
                      version,
                      "| Top 50 Tokens | TF-IDF"),
        caption = paste("DOI:",
                        doi.version),
        x = "Feature",
        y = "Weight"
    )+
    theme(
        text = element_text(size = 14),
        plot.title = element_text(size = 12,
                                  face = "bold")
    )
)
```



DOI: 10.5281/zenodo.7051934

24.7.2 French

```
print(
    ggplot(data = tstat.tfidf.fr[1:50, ],
           aes(x = reorder(feature,
                           frequency),
               y = frequency)) +
    geom_point() +
    coord_flip() +
    labs(
        title = paste(datashort,
                      "| FR | Version",
                      version,
                      "| Top 50 Tokens | TF-IDF"),
        caption = paste("DOI:",
                        doi.version),
        x = "Feature",
        y = "Weight"
    )+
    theme_bw()+
    theme(
        text = element_text(size = 14),
        plot.title = element_text(size = 12,
                                  face = "bold")
    )
)
```



DOI: 10.5281/zenodo.7051934

24.8 Most Frequent Tokens | TF Weighting | Wordclouds

24.8.1 English

```
textplot_wordcloud(dfm.en,
                   max_words = 100,
                   min_size = 1,
                   max_size = 5,
                   random_order = FALSE,
                   rotation = 0,
                   color = brewer.pal(8, "Dark2"))
```

24.8.2 French

```
textplot_wordcloud(dfm.fr,
                   max_words = 100,
                   min_size = 1,
                   max_size = 5,
                   random_order = FALSE,
                   rotation = 0,
                   color = brewer.pal(8, "Dark2"))
```

24.9 Most Frequent Tokens | TFIDF Weighting | Wordclouds

24.9.1 English

```
textplot_wordcloud(dfm.tfidf.en,
                   max_words = 100,
                   min_size = 1,
                   max_size = 2,
                   random_order = FALSE,
                   rotation = 0,
                   color = brewer.pal(8, "Dark2"))
```

```
## Warning in dfm_trim.dfm(x, min_termfreq = min_count): dfm has been previously
## weighted
```

24.9.2 French

```
textplot_wordcloud(dfm.tfidf.fr,
                   max_words = 100,
                   min_size = 1,
                   max_size = 2,
                   random_order = FALSE,
                   rotation = 0,
                   color = brewer.pal(8, "Dark2"))
```

```
## Warning in dfm_trim.dfm(x, min_termfreq = min_count): dfm has been previously
## weighted
```

25 Document Similarity

This analysis computes the correlation similarity for all documents in each corpus, plots the number of documents to drop as a function of the correlation similarity threshold and outputs the document IDs for specific threshold values.

The similarity test uses the standard pre-processed unigram document-feature matrix created by the f.token.processor function for the analyses of detailed token frequencies, i.e. it includes removal of numbers, special characters, stopwords (English/French) and lowercasing. I investigated other pre-processing workflows without the removal of features or lowercasing, as well as bigrams and trigrams, but, based on a qualitative assessment of the results, these performed no better or even worse than the standard workflow. Further research will be required to provide a definitive recommendation on how to deduplicate the corpus.

I intentionally do not correct for length, as the analysis focuses on detecting duplicates and near-duplicates, not topical similarity.

25.1 Set Ranges

Note: These ranges should cover most use cases.

```
threshold.range <- seq(0.8, 1, 0.005)

threshold.N <- length(threshold.range)

print(threshold.range)
```

```
##  [1] 0.800 0.805 0.810 0.815 0.820 0.825 0.830 0.835 0.840 0.845 0.850 0.855
## [13] 0.860 0.865 0.870 0.875 0.880 0.885 0.890 0.895 0.900 0.905 0.910 0.915
## [25] 0.920 0.925 0.930 0.935 0.940 0.945 0.950 0.955 0.960 0.965 0.970 0.975
## [37] 0.980 0.985 0.990 0.995 1.000
```

```
print.range <- print.range <- seq(0.8, 0.99, 0.01)

print(print.range)
```

```
##  [1] 0.80 0.81 0.82 0.83 0.84 0.85 0.86 0.87 0.88 0.89 0.90 0.91 0.92 0.93
    0.94
## [16] 0.95 0.96 0.97 0.98 0.99
```

25.2 English

25.2.1 Calculate Similarity

```
sim <- textstat_simil(dfm.en,
                      margin = "documents",
                      method = "correlation")

sim.dt <- as.data.table(sim)
```

25.2.2 Create Empty Lists

```
list.ndrop <- vector("list",
                     threshold.N)

list.drop.ids <- vector("list",
                        threshold.N)

list.pair.ids <- vector("list",
                        threshold.N)
```

25.2.3 Build Tables

```
for (i in 1:threshold.N){

    threshold <- threshold.range[i]

    pair.ids <- sim.dt[correlation > threshold]

    list.pair.ids[[i]] <- pair.ids

    drop.ids <- sim.dt[correlation > threshold,
                       .(unique(document1))][order(V1)]

    list.drop.ids[[i]] <- drop.ids

    ndrop <- drop.ids[,.N]

    list.ndrop[[i]] <- data.table(threshold,
                                  ndrop)

}


dt.ndrop <- rbindlist(list.ndrop)
```

25.2.4 IDs of Paired Documents Above Threshold

IDs of document pairs, with one of them to drop, as function of correlation similarity.

```
for (i in print.range){

    index <- match(i, threshold.range)

    fwrite(list.pair.ids[[index]],
           paste0(outputdir,
                  datashort,
                  "_EN_17_DocumentSimilarity_Correlation_PairedDocIDs_",
                  str_pad(threshold.range[index],
                          width = 5,
                          side = "right",
                          pad = "0"),
                  ".csv"))
}
```

25.2.5 IDs of Duplicate Documents per Threshold

IDs of Documents to drop as function of correlation similarity.

```
for (i in print.range){

    index <- match(i, threshold.range)

    fwrite(list.drop.ids[[index]],
           paste0(outputdir,
                  datashort,
                  "_EN_17_DocumentSimilarity_Correlation_DuplicateDocIDs_",
                  str_pad(threshold.range[index],
                          width = 5,
                          side = "right",
                          pad = "0"),
                  ".csv"))
}
```

25.2.6 Count of Duplicate Documents per Threshold

Number of Documents to drop as function of correlation similarity

```
kable(dt.ndrop,
      format = "latex",
      align = "r",
      booktabs = TRUE,
      longtable = TRUE,
      col.names = c("Threshold",
                    "Number to Drop")) %>% kable_styling(latex_options = "repeat_
    header")
```

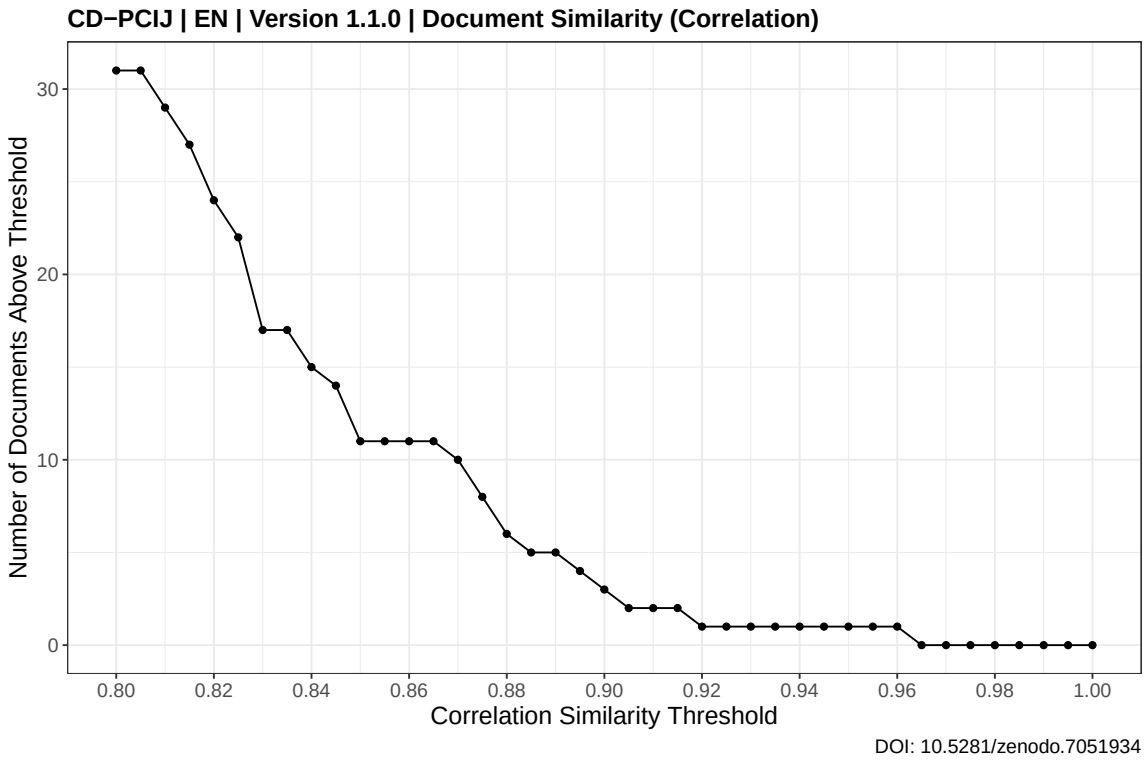
Threshold	Number to Drop
0.800	31
0.805	31
0.810	29
0.815	27
0.820	24
0.825	22
0.830	17
0.835	17
0.840	15
0.845	14
0.850	11
0.855	11
0.860	11
0.865	11
0.870	10
0.875	8
0.880	6
0.885	5
0.890	5
0.895	4
0.900	3
0.905	2
0.910	2
0.915	2
0.920	1
0.925	1
0.930	1
0.935	1
0.940	1
0.945	1

(continued)

Threshold	Number to Drop
0.950	1
0.955	1
0.960	1
0.965	0
0.970	0
0.975	0
0.980	0
0.985	0
0.990	0
0.995	0
1.000	0

```
fwrite(dt.ndrop,
       paste0(outputdir,
              datashort,
              "_EN_18_DocumentSimilarity_Correlation_Table.csv"))
```

```r
print(
    ggplot(data = dt.ndrop,
           aes(x = threshold,
               y = ndrop))+
    geom_line()+
    geom_point()+
    labs(
        title = paste(datashort,
                      "| EN | Version",
                      version,
                      "| Document Similarity (Correlation)"),
        caption = paste("DOI:",
                        doi.version),
        x = "Correlation Similarity Threshold",
        y = "Number of Documents Above Threshold"
    )+
    scale_x_continuous(breaks = seq(0.8, 1, 0.02))+
    theme_bw()+
    theme(
        text = element_text(size = 14),
        plot.title = element_text(size = 14,
                                  face = "bold"),
        legend.position = "bottom",
        legend.direction = "vertical"
    )
)
```



25.3 French

25.3.1 Calculate Similarity

```
sim <- textstat_simil(dfm.fr,
                      margin = "documents",
                      method = "correlation")

sim.dt <- as.data.table(sim)
```

25.3.2 Create Empty Lists

```
list.ndrop <- vector("list",
                     threshold.N)

list.drop.ids <- vector("list",
                        threshold.N)

list.pair.ids <- vector("list",
                        threshold.N)
```

25.3.3 Build Tables

```
for (i in 1:threshold.N){

    threshold <- threshold.range[i]

    pair.ids <- sim.dt[correlation > threshold]

    list.pair.ids[[i]] <- pair.ids

    drop.ids <- sim.dt[correlation > threshold,
                       .(unique(document1))][order(V1)]

    list.drop.ids[[i]] <- drop.ids

    ndrop <- drop.ids[,.N]

    list.ndrop[[i]] <- data.table(threshold,
                                  ndrop)

}

dt.ndrop <- rbindlist(list.ndrop)
```

25.3.4 IDs of Paired Documents Above Threshold

IDs of document pairs, with one of them to drop, as function of correlation similarity.

```
for (i in print.range){

    index <- match(i, threshold.range)

    fwrite(list.pair.ids[[index]],
           paste0(outputdir,
                  datashort,
                  "_FR_17_DocumentSimilarity_Correlation_PairedDocIDs_",
                  str_pad(threshold.range[index],
                          width = 5,
                          side = "right",
                          pad = "0"),
                  ".csv"))
}
```

25.3.5 IDs of Duplicate Documents per Threshold

IDs of Documents to drop as function of correlation similarity.

```
for (i in print.range){

    index <- match(i, threshold.range)

    fwrite(list.drop.ids[[index]],
           paste0(outputdir,
                  datashort,
                  "_FR_17_DocumentSimilarity_Correlation_DuplicateDocIDs_",
                  str_pad(threshold.range[index],
                          width = 5,
                          side = "right",
                          pad = "0"),
                  ".csv"))

}
```

25.3.6 Count of Duplicate Documents per Threshold

Number of Documents to drop as function of correlation similarity.

```
kable(dt.ndrop,
      format = "latex",
      align = "r",
      booktabs = TRUE,
      longtable = TRUE,
      col.names = c("Threshold",
                    "Number to Drop")) %>% kable_styling(latex_options = "repeat_
    header")
```

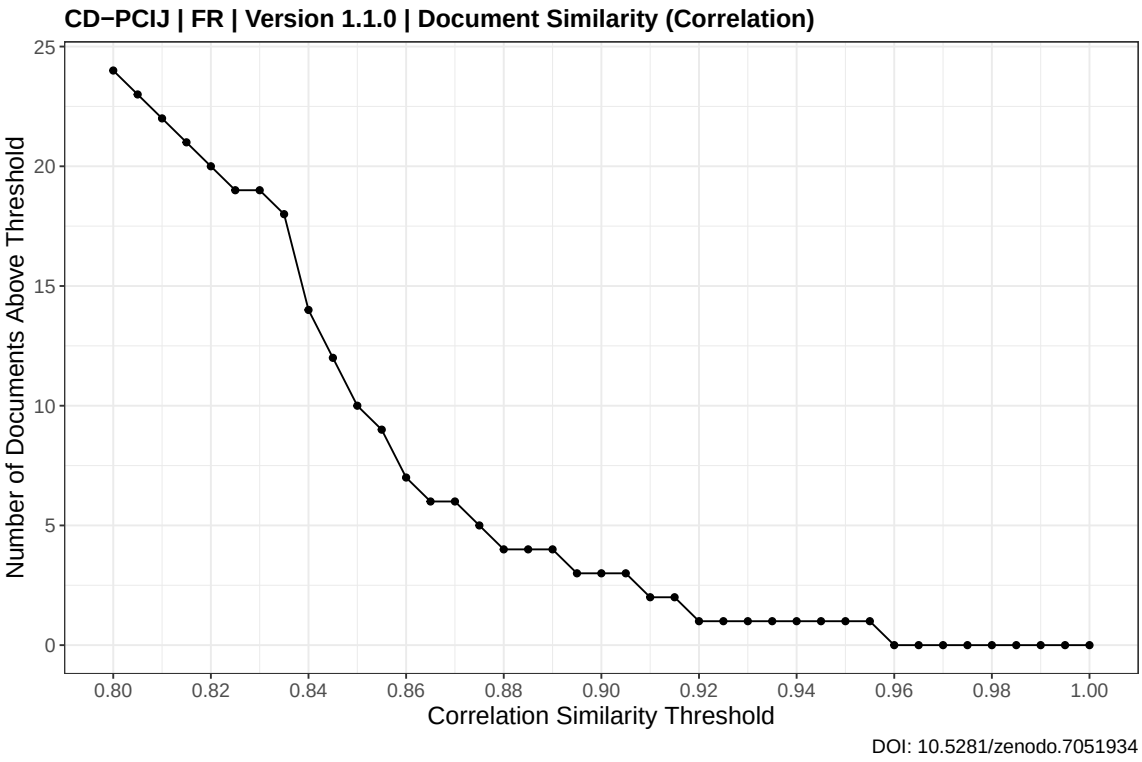
Threshold	Number to Drop
0.800	24
0.805	23
0.810	22
0.815	21
0.820	20
0.825	19
0.830	19
0.835	18
0.840	14
0.845	12
0.850	10
0.855	9
0.860	7
0.865	6
0.870	6
0.875	5
0.880	4
0.885	4
0.890	4
0.895	3
0.900	3
0.905	3
0.910	2
0.915	2
0.920	1
0.925	1
0.930	1
0.935	1
0.940	1
0.945	1

(continued)

Threshold	Number to Drop
0.950	1
0.955	1
0.960	0
0.965	0
0.970	0
0.975	0
0.980	0
0.985	0
0.990	0
0.995	0
1.000	0

```
fwrite(dt.ndrop,
       paste0(outputdir,
              datashort,
              "_FR_18_DocumentSimilarity_Correlation_Table.csv"))
```

```r
print(
    ggplot(data = dt.ndrop,
           aes(x = threshold,
               y = ndrop))+
    geom_line()+
    geom_point()+
    labs(
        title = paste(datashort,
                      "| FR | Version",
                      version,
                      "| Document Similarity (Correlation)"),
        caption = paste("DOI:",
                        doi.version),
        x = "Correlation Similarity Threshold",
        y = "Number of Documents Above Threshold"
    )+
    scale_x_continuous(breaks = seq(0.8, 1, 0.02))+
    theme_bw()+
    theme(
        text = element_text(size = 14),
        plot.title = element_text(size = 14,
                                  face = "bold"),
        legend.position="bottom",
        legend.direction = "vertical"
    )
)
```



26 Create CSV Files

26.1 Full Data Set

```
csvname.full.en <- paste(datashort,
                     version.dash,
                     "EN_CSV_TESSERACT_FULL.csv",
                     sep = "_")

csvname.full.fr <- paste(datashort,
                     version.dash,
                     "FR_CSV_TESSERACT_FULL.csv",
                     sep = "_")


fwrite(data.tesseract.en,
       csvname.full.en,
       na = "NA")

fwrite(data.tesseract.fr,
       csvname.full.fr,
       na = "NA")
```

26.2 Metadata Only

These files are the same as the full data set, minus the "text" variable.

```
csvname.meta.en <- paste(datashort,
                     version.dash,
                     "EN_CSV_TESSERACT_META.csv",
                     sep = "_")

csvname.meta.fr <- paste(datashort,
                     version.dash,
                     "FR_CSV_TESSERACT_META.csv",
                     sep = "_")


fwrite(meta.tesseract.en,
       csvname.meta.en,
       na = "NA")

fwrite(meta.tesseract.fr,
       csvname.meta.fr,
       na = "NA")
```

27 Final File Count per Folder

Note: Strictly speaking one of the German documents (the Danzig Courts file) is not a true original, as it was split from a bilingual file. However the quality of the document (scan and OCR) is original, so it is stored with the other originals to avoid creating another variant for a single document. This is the reason why the "MULT" variant contains 265 files instead of the maximum 264 that were downloaded from the ICJ website.

```
dir.table <- as.data.table(dirset)[, {
    filecount <- lapply(dirset,
                        function(x){length(list.files(x))})
    list(dirset, filecount)
}]


kable(dir.table,
      format = "latex",
      align = "r",
      booktabs = TRUE,
      longtable = TRUE,
      linesep = "",
      col.names = c("Directory",
                    "Filecount"))
```

Directory	Filecount
MULT_PDF_ORIGINAL_FULL	265
EN_PDF_ENHANCED_FULL	259
FR_PDF_ENHANCED_FULL	261
EN_PDF_ORIGINALSPLIT_FULL	259
FR_PDF_ORIGINALSPLIT_FULL	261
EN_PDF_ENHANCED_MajorityOpinions	100
FR_PDF_ENHANCED_MajorityOpinions	100
EN_TXT_TESSERACT_FULL	259
FR_TXT_TESSERACT_FULL	261
EN_TXT_EXTRACTED_FULL	259
FR_TXT_EXTRACTED_FULL	261

28 File Size Distribution

28.1 English

28.1.1 Corpus Object in RAM

```
print(object.size(corpus.en.b),
      humanReadable = TRUE,
      units = "MB")
```

```
## 6.8 Mb
```

28.1.2 Create Data Table of Filenames

```
enhanced <- list.files("EN_PDF_ENHANCED_FULL",
                       full.names = TRUE)

original <- list.files("MULT_PDF_ORIGINAL_FULL",
                       full.names = TRUE)

MB <- file.size(enhanced) / 10^6

dt1 <- data.table(MB, rep("ENHANCED",
                          length(MB)))


MB <- file.size(original) / 10^6

dt2 <- data.table(MB, rep("ORIGINAL",
                          length(MB)))


dt <- rbind(dt1,
            dt2)
setnames(dt,
         "V2",
         "variant")
```

28.1.3 Total Size Comparison

```
kable(dt[,
         .(MB_total = sum(MB)),
         keyby = variant],
      format = "latex",
      align = "r",
      booktabs = TRUE,
      longtable = TRUE)
```

variant	MB_total
ENHANCED	344.6932
ORIGINAL	246.7558

28.1.4 Analyze Files Larger than 10 MB

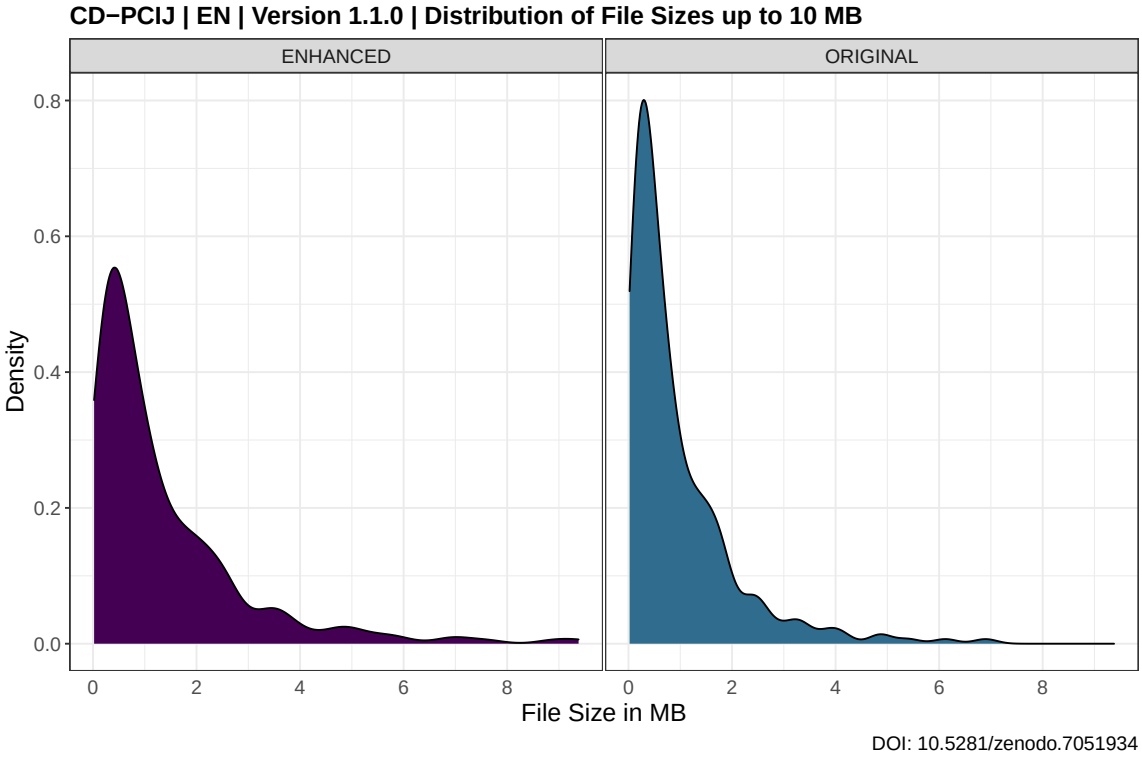
```
summary(dt[MB > 10]$MB)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##
```

28.1.5 Plot Density Distribution for Files 10MB or Less

```
dt.plot <- dt[MB <= 10]
```

```
print(
    ggplot(data = dt.plot,
           aes(x = MB,
               group = variant,
               fill = variant))+
    geom_density()+
    theme_bw()+
    facet_wrap(~variant,
               ncol = 2) +
    labs(
        title = paste(datashort,
                      "| EN | Version",
                      version,
                      "| Distribution of File Sizes up to 10 MB"),
        caption = paste("DOI:",
                        doi.version),
        x = "File Size in MB",
        y = "Density"
    )+
    scale_x_continuous(breaks = seq(0, 10, 2))+
    scale_fill_viridis(end = 0.35, discrete = TRUE) +
    scale_color_viridis(end = 0.35, discrete = TRUE) +
    theme(
        text = element_text(size=  14),
        plot.title = element_text(size = 14,
                                  face = "bold"),
        legend.position = "none",
        panel.spacing = unit(0.1,
                             "lines"),
        axis.ticks.x = element_blank()
    )
)
```


CD−PCIJ | EN | Version 1.1.0 | Distribution of File Sizes up to 10 MB
ENHANCED
ORIGINAL
Density
0.0
0.2
0.4
0.6
0.8
0
2
4
6
8
File Size in MB
DOI: 10.5281/zenodo.7051934

28.2 French

28.2.1 Corpus Object in RAM

```
print(object.size(corpus.fr.b),
      humanReadable = TRUE,
      units = "MB")
```

```
## 7.1 Mb
```

28.2.2 Create Data Table of filenames

```
enhanced <- list.files("FR_PDF_ENHANCED_FULL",
                       full.names = TRUE)

original <- list.files("MULT_PDF_ORIGINAL_FULL",
                       full.names = TRUE)


MB <- file.size(enhanced) / 10^6
dt1 <- data.table(MB,
                  rep("ENHANCED",
                      length(MB)))



MB <- file.size(original) / 10^6

dt2 <- data.table(MB,
                  rep("ORIGINAL",
                      length(MB)))


dt <- rbind(dt1,
            dt2)
setnames(dt,
         "V2",
         "variant")
```

28.2.3 Total Size Comparison

```
kable(dt[,
         .(MB_total = sum(MB)),
         keyby = variant],
      format = "latex",
      align = "r",
      booktabs = TRUE,
      longtable = TRUE)
```

variant	MB_total
ENHANCED	337.7872
ORIGINAL	246.7558

28.2.4 Analyze Files Larger than 10 MB

```
summary(dt[MB > 10]$MB)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##
```

28.2.5 Plot Density Distribution for Files 10MB or Less

```
dt.plot <- dt[MB <= 10]
```

```
print(
    ggplot(data = dt.plot,
           aes(x = MB,
               group = variant,
               fill = variant)) +
    geom_density() +
    theme_bw() +
    facet_wrap(~variant,
               ncol = 2) +
    labs(
        title = paste(datashort,
                      "| FR | Version",
                      version,
                      "| Distribution of File Sizes up to 10 MB"),
        caption = paste("DOI:",
                        doi.version),
        x = "File Size in MB",
        y = "Density"
    )+
    scale_fill_viridis(end = 0.35, discrete = TRUE) +
    scale_color_viridis(end = 0.35, discrete = TRUE) +
    scale_x_continuous(breaks = seq(0, 10, 2))+
    theme(
        text = element_text(size = 14),
        plot.title = element_text(size = 14,
                                  face = "bold"),
        legend.position = "none",
        panel.spacing = unit(0.1,
                             "lines"),
        axis.ticks.x = element_blank()
    )
)
```


CD−PCIJ | FR | Version 1.1.0 | Distribution of File Sizes up to 10 MB
ENHANCED
ORIGINAL
Density
0.0
0.2
0.4
0.6
0.8
0
2
4
6
8
File Size in MB
DOI: 10.5281/zenodo.7051934

29 Create ZIP Archives

29.1 ZIP CSV Files

```
csv.zip.name.full.en <- gsub(".csv",
                             "",
                             csvname.full.en)

csv.zip.name.full.fr <- gsub(".csv",
                             "",
                             csvname.full.fr)

csv.zip.name.meta.en <- gsub(".csv",
                             "",
                             csvname.meta.en)

csv.zip.name.meta.fr <- gsub(".csv",
                             "",
                             csvname.meta.fr)
```

```
zip(csv.zip.name.full.fr,
    csvname.full.fr)

zip(csv.zip.name.full.en,
    csvname.full.en)

zip(csv.zip.name.meta.fr,
    csvname.meta.fr)

zip(csv.zip.name.meta.en,
    csvname.meta.en)
```

29.2 ZIP Data Directories

Note: Vector of Directories was created at the beginning of the script.

```
for (dir in dirset){
    zip(paste(datashort,
              version.dash,
              dir,
              sep = "_"),
        dir)
}
```

29.3 ZIP ANALYSIS Directory

```
zip(paste(datashort,
          version.dash,
          "EN-FR",
          basename(outputdir),
          sep = "_"),
    basename(outputdir))
```

29.4 ZIP Source Files

```
files.source <-  c(list.files(pattern = "\\.R$|\\.toml$|\\.md$|\\.Rmd$"),
                   "data",
                   "functions",
                   "tex",
                   "buttons",
                   list.files(pattern = "renv\\.lock|\\.Rprofile",
                              all.files = TRUE),
                   list.files("renv",
                              pattern = "activate\\.R",
                              full.names = TRUE))




files.source <- grep("spin",
                     files.source,
                     value = TRUE,
                     ignore.case = TRUE,
                     invert = TRUE)

zip(paste(datashort,
          version.dash,
          "Source_Files.zip",
          sep = "_"),
    files.source)
```

30 Delete CSV and Directories

The metadata CSV files are retained for Codebook generation.

30.1 Delete CSV Data Set

```
unlink(csvname.full.fr)
unlink(csvname.full.en)
unlink(csvname.meta.fr)
unlink(csvname.meta.en)
```

30.2 Delete Data Directories

```
for (dir in dirset){
    unlink(dir,
           recursive = TRUE)
}
```

31 Cryptography Module

This module computes two types of hashes for every ZIP archive: SHA2-256 and SHA3-512. These are proof of the authenticity and integrity of data and document that the files are the result of this source code. The SHA-2 and SHA-3 family of algorithms are highly resistant to collision and pre-imaging attacks in reasonable scenarios and can therefore be considered secure according to current public cryptographic research. SHA3 hashes with an output length of 512 bit may even provide sufficient security when attacked with quantum cryptanalysis based on Grover's algorithm.

31.1 Create Set of ZIP Archives

```
files.zip <- list.files(pattern= "\\.zip$",
                        ignore.case = TRUE)
```

31.2 Show Function: f.dopar.multihashes

```
print(f.dopar.multihashes)
```

function(x, threads = detectCores()){

```
print(paste("Parallel processing using", threads, "threads."))

begin <- Sys.time()

cl <- makeForkCluster(threads)
registerDoParallel(cl)

multihashes <- foreach(filename = x,
                       .errorhandling = 'pass',
                       .combine = 'rbind') %dopar% {

                           sha2.256 <- system2("openssl",
                                                paste("sha256",
                                                      filename),
                                                stdout = TRUE)

                           sha2.256 <- gsub("^.*\\= ",
                                            "",
                                            sha2.256)

                           sha3.512 <- system2("openssl",
                                                paste("sha3-512",
                                                      filename),
                                                stdout = TRUE)

                           sha3.512 <- gsub("^.*\\= ",
                                            "",
```

```
                                                sha3.512)

                            out <- data.frame(filename,
                                              sha2.256,
                                              sha3.512)
                            return(out)
                        }
stopCluster(cl)

end <- Sys.time()
duration <- end - begin

print(paste0("Processed ",
             length(x),
             " files. Runtime was ",
             round(duration,
                   digits = 2),
             " ",
             attributes(duration)$units,
             "."))

return(multihashes)
```

}

31.3 Compute Hashes

```
multihashes <- f.dopar.multihashes(files.zip)
```

```
## [1] "Parallel processing using 16 threads."
## [1] "Processed 17 files. Runtime was 1.75 secs."
```

31.4 Convert to Data Table

```
setDT(multihashes)
```

31.5 Add Index

```
multihashes$index <- seq_len(multihashes[,.N])
```

31.6 Save to Disk

```
fwrite(multihashes,
       paste(datashort,
             version.dash,
             "CryptographicHashes.csv",
             sep = "_"),
       na = "NA")
```

31.7 Add Whitespace to Enable Automatic Linebreak

```
multihashes$sha3.512 <- paste(substr(multihashes$sha3.512, 1, 64),
                              substr(multihashes$sha3.512, 65, 128))
```

31.8 Print to Report

```
kable(multihashes[,.(index,filename)],
      format = "latex",
      align = c("p{1cm}",
                "p{13cm}"),
      booktabs = TRUE,
      longtable = TRUE)
```

index	filename
1	CD-PCIJ_1-1-0_EN_CSV_TESSERACT_FULL.zip
2	CD-PCIJ_1-1-0_EN_CSV_TESSERACT_META.zip
3	CD-PCIJ_1-1-0_EN_PDF_ENHANCED_FULL.zip
4	CD-PCIJ_1-1-0_EN_PDF_ENHANCED_MajorityOpinions.zip
5	CD-PCIJ_1-1-0_EN_PDF_ORIGINALSPLIT_FULL.zip
6	CD-PCIJ_1-1-0_EN_TXT_EXTRACTED_FULL.zip
7	CD-PCIJ_1-1-0_EN_TXT_TESSERACT_FULL.zip
8	CD-PCIJ_1-1-0_EN-FR_ANALYSIS.zip
9	CD-PCIJ_1-1-0_FR_CSV_TESSERACT_FULL.zip
10	CD-PCIJ_1-1-0_FR_CSV_TESSERACT_META.zip
11	CD-PCIJ_1-1-0_FR_PDF_ENHANCED_FULL.zip
12	CD-PCIJ_1-1-0_FR_PDF_ENHANCED_MajorityOpinions.zip
13	CD-PCIJ_1-1-0_FR_PDF_ORIGINALSPLIT_FULL.zip
14	CD-PCIJ_1-1-0_FR_TXT_EXTRACTED_FULL.zip
15	CD-PCIJ_1-1-0_FR_TXT_TESSERACT_FULL.zip
16	CD-PCIJ_1-1-0_MULT_PDF_ORIGINAL_FULL.zip
17	CD-PCIJ_1-1-0_Source_Files.zip

```
kable(multihashes[,.(index,sha2.256)],
      format = "latex",
      align = c("c",
                "p{13cm}"),
      booktabs = TRUE,
      longtable = TRUE)
```

index	sha2.256
1	ca74fe7ad4c353fa5d22fce9b39537f979ed5a12b589634ebaf995718910ad70
2	cc239328a814bc384ae216d401176d35de3313f6bf7bc74a32fe46731b180d4a
3	e43c7cd81432dbe2d9fe5044769bffaf141e063411f712ff6531047e4663d7f7
4	19c1e8a0094294432b7d58e8d719da51a85d7d6d63d236f70b52972c9e1286bb
5	485f44b710229e8475d20a404caff2d525dedac259e166603a63e6233b8bbedd
6	41ed3d7d33d4af42f7abd249c2fcaae3ac17dcda8f01701037c6b39d4303b1bb
7	2f4ffde1eae2dece3b1ed91b4dbd5dd50d1312c0894b9d2157b0afe95e16feba
8	91056e4f5ba538bd98384e10978e2422723d2effb8486911547075b167292dcf
9	64fbc15011015a69a897b46f5f46663fdf31930580a3318d1f47e8580611f896
10	340c13e875f0502e66a3ee7f77ca0dfb4681d0e9fe4ee08de02ba8f6444d1bec
11	79b0b616cb52f8d5a64f02ca3311ddbca73bdcb74c7503d4bc452af35e617832
12	e914d6c7dd758a160a5cfe3f8ba67aebe325049dc0359ce660f2ad921a7a959c
13	3eb58ce86b6e495d317bbedb25bfa0465a1e42127f7c4f9a6de4126ffff89b19
14	c06fbaef18d51e7d5d25ad27d43a4e304ca5b903b190bf50c9c4e4527814927e
15	97156d8fcc6e8a2c367d0ffa47d994e8fa91aea33481727b79bd6b9a280e0d56
16	633f429006543328d78d465b1559ae810183af062354d1fd04048a808ba7308f
17	baf3326bd59de41448b4aa1e111731d359288ea60304b84543342b64952f9750

```
kable(multihashes[,.(index,sha3.512)],
      format = "latex",
      align = c("c",
                "p{13cm}"),
      booktabs = TRUE,
      longtable = TRUE)
```

index	sha3.512
1	1ce867599f400ab32ae7efe87ee8fab1c8d59f9bd917423336aad16a867bb2b8 25fbb5df1992ac2c2a0abecd08485756dac2d38066d094f8a0598cdc620f074d
2	c3393f3654b7a68ccbef9d78635b1bb0cb3ea8930d0c08c8be5f5af56fed3fcb f24370d1bd268fa71a0d9d15c1047c1b719c136ce9097638b239fc1e345ca9c5
3	26c6e063ccebcb07b2e2a6691a5529a6ac5922905f27904ae2564f6c670e526c 4f6569078d66c4755631cc2201bf5901f16aabb1f3e1e5e9ac67ae02df5e003b
4	fed455437dd22303b1789c89af815388aa948dfdbafc8b6977ea1808016878d6 3bcee41cb63f5da0da5aca2b2d6819cbcbf977ef4e5b81d3a5ee042e931fed9d
5	562172dd70d6d6a270aae957f4daf22fa115761bfbd349332a5a2f319a026d28 ad2f6844e5e2663d1c0ffbf089c510600257e9ad4b8984a6cc9ad4cded5d966f
6	3263b2e752a80ae6876e43688c3dc4afaa47593109e8c33dd22a39ab40f4b401 18e7da0409966ddb99033deacc6b26f27d88dd25c2791b85675c540ce92903c1
7	83ad266d415dd3aa138caeba8e46289e34b5538cbd298b2b0781a007006271e2 c2a1de16b77f215c7bec6f7e0e7111b1e25555cb3e2c1a35c699c9e792fe90e8
8	9812add170f0e43a4deb57d00a9f3a7eb623de1e7eb91cc9efc47273950b34a6 542763671848604cafb9a1631668975043d592109d2b0f99534c27841ea82329
9	e865074d87d575d7a0c9525bce92023227150d2cea48b2c3c5d08dc3c68fd32f 52bb20cce71e1decdad63a989f350de6fe71c77091e4ea1c22ffa0eb2082e60e
10	1f4c1c986b5c512e676418f7e28be0a18abf179b556c5906c16029e2fef18910 cedee30b01b39c79836cf780e4c0fa89f8d8b4c4a85746d5be777b514937b85e
11	631454ccee6e0588830fb2b67bd2668cec9fec7c73b95ea56c3a4dd078c09e75 d74fa0f46ed6e5bcd604187aeab97a621c07c1d6c3a70b4c61986ed13626260f
12	bba85af95011f641dca8da6f52856f379c1350f12012a0d014f6ca1c435587e3 151cd84b270ef49da0ae712122594dafb849dceea4cdbaaf75393b35c4a7843a
13	b169b5acad8cecd2b1ff6a49d8e28c3c450ebe11a6cda7a08078922587030c71 88f87337d6756ec5bc3e8a7faeb7da7e2daf7f5339f7d9ce611df7f20ba24f16
14	8565435b976cc1f77c2c24e1ac4ba8e576c94a9397c887d33136ab4904d75a5d a9698ea99de814f73d334e68ca91f29c602a181bf225eee1c5246640891bb221
15	dc266dc09442babd9a1f6abdca9bb4e3274aaca4eae8e41c8830edca55b0c9b5 50f81430aa9aecb656a0590e8f3660dfea30c0bab21e81bb828e3076c4d8e0a9

16	48d03e9d00c2d4f4ed9c269467de97cd634391daf2f0953106fd1a0cea794f6b 3bf58d3297bb6b26781a403ea69465350ec71b85872ade54c0219a7d8d28f007
17	dd339eaf99af1b894e2c2d764e2fa7508b7f2702d7e6af33b9d2b6e350499bad a790f582e98a71112b06c4b8a0c2db5804ce40cc5cfb90d472e434a7dd08ad0b

32 Finalize

32.1 Datestamp

```
print(datestamp)
```

```
## [1] "2022-09-06"
```

32.2 Date and Time (Begin)

```
print(begin.script)
```

```
## [1] "2022-09-06 19:01:06 CEST"
```

32.3 Date and Time (End)

```
end.script <- Sys.time()
print(end.script)
```

```
## [1] "2022-09-06 20:02:09 CEST"
```

32.4 Script Runtime

```
print(end.script - begin.script)
```

```
## Time difference of 1.01756 hours
```

32.5 Warnings

```
warnings()
```

33 Strict Replication Parameters

```
sessionInfo()
```

```
## R version 4.1.3 (2022-03-10)
## Platform: x86_64-redhat-linux-gnu (64-bit)
## Running under: Fedora Linux 36 (Workstation Edition)
##
## Matrix products: default
## BLAS/LAPACK: /usr/lib64/libflexiblas.so.3.2
##
## locale:
##  [1] LC_CTYPE=en_US.utf8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.utf8        LC_COLLATE=en_US.utf8
##  [5] LC_MONETARY=en_US.utf8    LC_MESSAGES=en_US.utf8
##  [7] LC_PAPER=en_US.utf8       LC_NAME=C
##  [9] LC_ADDRESS=C              LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.utf8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] parallel  stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] doParallel_1.0.17         iterators_1.0.14
##  [3] foreach_1.5.2             data.table_1.14.2
##  [5] textcat_1.0-7             quanteda.textplots_0.94.2
##  [7] quanteda.textstats_0.95   quanteda_3.2.3
##  [9] readtext_0.81             RColorBrewer_1.1-3
## [11] viridis_0.6.2             viridisLite_0.4.1
## [13] scales_1.2.1              ggplot2_3.3.6
## [15] rsvg_2.3.1                DiagrammeRsvg_0.1
## [17] DiagrammeR_1.0.9          magick_2.7.3
## [19] kableExtra_1.3.4          knitr_1.40
## [21] fs_1.5.2                  pdftools_3.3.0
## [23] stringr_1.4.1             mgsub_1.7.3
## [25] rvest_1.0.3               httr_1.4.4
## [27] RcppTOML_0.1.7            rmarkdown_2.16
##
## loaded via a namespace (and not attached):
##  [1] jsonlite_1.8.0      RcppParallel_5.1.5  askpass_1.1
##  [4] highr_0.9           selectr_0.4-2       renv_0.15.5
##  [7] yaml_2.3.5          slam_0.1-50         qpdf_1.2.0
## [10] pillar_1.8.1        lattice_0.20-45     glue_1.6.2
## [13] digest_0.6.29       tau_0.0-24          colorspace_2.0-3
## [16] htmltools_0.5.3     Matrix_1.4-1        pkgconfig_2.0.3
## [19] ISOcodes_2022.01.10 purrr_0.3.4         webshot_0.5.3
## [22] svglite_2.1.0       nsyllable_1.0.1     tibble_3.1.8
## [25] farver_2.1.1        generics_0.1.3      withr_2.5.0
## [28] cli_3.3.0           magrittr_2.0.3      evaluate_0.16
## [31] stopwords_2.3       fansi_1.0.3         xml2_1.3.3
## [34] tools_4.1.3         lifecycle_1.0.1     V8_4.2.1
## [37] munsell_0.5.0       compiler_4.1.3      proxyC_0.3.2
```

```
## [40] systemfonts_1.0.4   rlang_1.0.5         grid_4.1.3
## [43] rstudioapi_0.14     htmlwidgets_1.5.4   visNetwork_2.1.0
## [46] labeling_0.4.2      gtable_0.3.1        codetools_0.2-18
## [49] curl_4.3.2          R6_2.5.1            gridExtra_2.3
## [52] dplyr_1.0.10        fastmap_1.1.0       utf8_1.2.2
## [55] fastmatch_1.1-3     stringi_1.7.8       Rcpp_1.0.9
## [58] vctrs_0.4.1         tidyselect_1.1.2    xfun_0.32
```

```
system2("openssl",
        "version",
        stdout = TRUE)
```

```
## [1] "OpenSSL 3.0.5 5 Jul 2022 (Library: OpenSSL 3.0.5 5 Jul 2022)"
```

```
system2("tesseract",
        "-v",
        stdout = TRUE)
```

```
## [1] "tesseract 5.0.1"
## [2] " leptonica-1.82.0"
## [3] "  libgif 5.2.1 : libjpeg 6b (libjpeg-turbo 2.1.2) : libpng 1.6.37 :
    libtiff 4.4.0 : zlib 1.2.11 : libwebp 1.2.4"
## [4] " Found AVX2"
## [5] " Found AVX"
## [6] " Found FMA"
## [7] " Found SSE4.1"
## [8] " Found OpenMP 201511"
```

```
system2("convert",
        "--version",
        stdout = TRUE)
```

```
## [1] "Version: ImageMagick 6.9.12-61 Q16 x86_64 17418 https://legacy.
    imagemagick.org"
## [2] "Copyright: (C) 1999 ImageMagick Studio LLC"
## [3] "License: https://imagemagick.org/script/license.php"
## [4] "Features: Cipher DPC Modules OpenMP(4.5) "
## [5] "Delegates (built-in): bzlib cairo djvu fontconfig freetype gslib gvc jbig
     jng jp2 jpeg lcms lqr ltdl lzma openexr pangocairo png ps raqm raw rsvg tiff
     webp wmf x xml zlib"
```

```
print(quanteda_options())
```

```
## $threads
## [1] 16
##
## $verbose
## [1] FALSE
##
## $print_dfm_max_ndoc
## [1] 6
##
## $print_dfm_max_nfeat
## [1] 10
##
## $print_dfm_summary
## [1] TRUE
##
## $print_corpus_max_ndoc
## [1] 6
##
## $print_corpus_max_nchar
## [1] 60
##
## $print_corpus_summary
## [1] TRUE
##
## $print_tokens_max_ndoc
## [1] 6
##
## $print_tokens_max_ntoken
## [1] 12
##
## $print_tokens_summary
## [1] TRUE
##
## $print_dictionary_max_nkey
## [1] 6
##
## $print_dictionary_max_nval
## [1] 20
##
## $print_dictionary_summary
## [1] TRUE
##
## $print_kwic_max_nrow
## [1] 1000
##
## $print_kwic_summary
## [1] TRUE
##
## $base_docname
## [1] "text"
##
## $base_featname
## [1] "feat"
##
## $base_compname
```

```
## [1] "comp"
##
## $language_stemmer
## [1] "english"
##
## $pattern_hashtag
## [1] "#\\w+#?"
##
## $pattern_username
## [1] "@[a-zA-Z0-9_]+"
##
## $tokens_block_size
## [1] 10000
##
## $tokens_locale
## [1] "fr"
```

References

Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2022. *Rmarkdown: Dynamic Documents for r*. https://CRAN.R-project.org/package=rmarkdown.

Analytics, Revolution, and Steve Weston. 2022. *Iterators: Provides Iterator Construct*. https://github.com/RevolutionAnalytics/iterators.

Benoit, Kenneth, and Adam Obeng. 2021. *Readtext: Import and Handling for Plain and Formatted Text Files*. https://github.com/quanteda/readtext.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Jiong Wei Lua, and Jouni Kuha. 2021. *Quanteda.textstats: Textual Statistics for the Quantitative Analysis of Textual Data*. https://quanteda.io.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018c. "Quanteda: An r Package for the Quantitative Analysis of Textual Data." *Journal of Open Source Software* 3 (30): 774. https://doi.org/10.21105/joss.00774.

———. 2018a. "Quanteda: An r Package for the Quantitative Analysis of Textual Data." *Journal of Open Source Software* 3 (30): 774. https://doi.org/10.21105/joss.00774.

———. 2018b. "Quanteda: An r Package for the Quantitative Analysis of Textual Data." *Journal of Open Source Software* 3 (30): 774. https://doi.org/10.21105/joss.00774.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, Akitaka Matsuo, and William Lowe. 2022. *Quanteda: Quantitative Analysis of Textual Data*. https://quanteda.io.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2022. *Quanteda.textplots: Plots for the Quantitative Analysis of Textual Data*. https://CRAN.R-project.org/package=quanteda.textplots.

Corporation, Microsoft, and Steve Weston. 2022. *doParallel: Foreach Parallel Adaptor for the Parallel Package*. https://github.com/RevolutionAnalytics/doparallel.

Dowle, Matt, and Arun Srinivasan. 2021. *Data.table: Extension of 'Data.frame'*. https://CRAN.R-project.org/package=data.table.

Eddelbuettel, Dirk. 2020. *RcppTOML: Rcpp Bindings to Parser for Tom's Obvious Markup Language*. http://dirk.eddelbuettel.com/code/rcpp.toml.html.

Ewing, Mark. 2021. *Mgsub: Safe, Multiple, Simultaneous String Substitution*. https://CRAN.R-project.org/package=mgsub.

Garnier, Simon. 2021. *Viridis: Colorblind-Friendly Color Maps for r*. https://CRAN.R-project.org/package=viridis.

———. 2022. *viridisLite: Colorblind-Friendly Color Maps (Lite Version)*. https://CRAN.R-project.org/package=viridisLite.

Hester, Jim, Hadley Wickham, and Gábor Csárdi. 2021. *Fs: Cross-Platform File System Operations Based on Libuv*. https://CRAN.R-project.org/package=fs.

Hornik, Kurt, Patrick Mair, Johannes Rauch, Wilhelm Geiger, Christian Buchta, and Ingo Feinerer. 2013. "The textcat Package for n-Gram Based Text Categorization in R." *Journal of Statistical Software* 52 (6): 1–17. https://doi.org/10.18637/jss.v052.i06.

Hornik, Kurt, Johannes Rauch, Christian Buchta, and Ingo Feinerer. 2020. *Textcat: N-Gram Based Text Categorization*. https://CRAN.R-project.org/package=textcat.

Iannone, Richard. 2016. *DiagrammeRsvg: Export DiagrammeR Graphviz Graphs as SVG*. https://github.com/rich-iannone/DiagrammeRsvg.

———. 2022. *DiagrammeR: Graph/Network Visualization*. https://github.com/rich-iannone/DiagrammeR.

Neuwirth, Erich. 2022. *RColorBrewer: ColorBrewer Palettes*. https://CRAN.R-project.or

g/package=RColorBrewer.
Ooms, Jeroen. 2021. *Magick: Advanced Graphics and Image-Processing in r*. https://CRAN.R-project.org/package=magick.
———. 2022a. *Pdftools: Text Extraction, Rendering and Converting of PDF Documents*. https://CRAN.R-project.org/package=pdftools.
———. 2022b. *Rsvg: Render SVG Images into PDF, PNG, (Encapsulated) PostScript, or Bitmap Arrays*. https://CRAN.R-project.org/package=rsvg.
R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.
Revolution Analytics, and Steve Weston. n.d. *Foreach: Provides Foreach Looping Construct*.
Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.
———. 2022a. *Httr: Tools for Working with URLs and HTTP*. https://CRAN.R-project.org/package=httr.
———. 2022b. *Rvest: Easily Harvest (Scrape) Web Pages*. https://CRAN.R-project.org/package=rvest.
———. 2022c. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://CRAN.R-project.org/package=stringr.
Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. 2022. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://CRAN.R-project.org/package=ggplot2.
Wickham, Hadley, and Dana Seidel. 2022. *Scales: Scale Functions for Visualization*. https://CRAN.R-project.org/package=scales.
Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.
———. 2015. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.org/knitr/.
———. 2022. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.
Xie, Yihui, J. J. Allaire, and Garrett Grolemund. 2018. *R Markdown: The Definitive Guide*. Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown.
Xie, Yihui, Christophe Dervieux, and Emily Riederer. 2020. *R Markdown Cookbook*. Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown-cookbook.
Zhu, Hao. 2021. *kableExtra: Construct Complex Table with Kable and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.