Corpus of Decisions

International Court of Justice

(CD-ICJ-Source)

Compilation Report

Version 2022-09-07

DOI: 10.5281/zenodo.7051932

Title	Source Code for the 'Corpus of Decisions: International Court of Justice'
Abbreviation	CD-ICJ-Source
Author	Seán Fobbe
Version	2022-09-07
Download	https://doi.org/10.5281/zenodo.7051932
License	MIT No Attribution (MIT-0)

Citation

Seán Fobbe (2022). Source Code for the 'Corpus of Decisions: International Court of Justice' (CD-ICJ-Source). Version 2022-09-07. Zenodo. DOI: 10.5281/zenodo.7051932.

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 2022-09-07. 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

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

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

1.1 Overview

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

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.3826444

1.2 Functionality

This script will produce 21 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 of documents dated 2004 and earlier containing monolingual documents with enhanced OCR (English/French)
- 2 archives of PDF files as originally published by the ICJ (English/French)
- 2 archives of TXT files containing text as generated by Tesseract for documents dated 2004 or earlier (English/French)
- 2 archives of TXT files containing extracted text from the original documents (English/French)
- 1 archive PDF files that were unlabelled on the website (intended for replication and review only)
- 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 'BEST' 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 11 hours 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

This datestamp will be applied to all output files. It is set at the beginning of the script so it will be held constant for all output even if long runtime breaks the date barrier.

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

```
## [1] "2022-09-07"
```

2.2 Date and Time (Begin)

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

```
## [1] "2022-09-07 01:10:29 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
```

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

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

```
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
```

```
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.
```

```
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
```

```
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.

```
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.selectpdflinks.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 10
##  $ caseno  :List of 3
##   ..$ begin  : int 1
##   ..$ end    : int 183
##   ..$ exclude: int 2
##  $ cores   :List of 2
##   ..$ max    : logi TRUE
##   ..$ number: int 8
##  $ debug   :List of 2
##   ..$ sample: int 3
##   ..$ toggle: logi FALSE
##  $ doi     :List of 2
##   ..$ data    :List of 2
##   .. ..$ concept: chr "10.5281/zenodo.3826444"
##   .. ..$ version: chr "10.5281/zenodo.7051929"
##   ..$ software:List of 2
##   .. ..$ concept: chr "10.5281/zenodo.3977176"
##   .. ..$ version: chr "10.5281/zenodo.7051932"
##  $ 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: International Court of Justice"
##   ..$ shortname: chr "CD-ICJ"
```

3.3 Name of Data Set

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

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

3.4 DOI of Data Set Concept

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

```
## [1] "10.5281/zenodo.3826444"
```

3.5 DOI of Specific Version

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

```
## [1] "10.5281/zenodo.7051929"
```

3.6 License

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

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

3.7 Scope: Case Numbers

These variables define the scope of cases (by ordinal number) to be compiled into the data set.

Case number 2 appears to be unassigned. There is no information available on the ICJ website. It is therefore always excluded.

The variable for the final case number — caseno.end — must be set manually.

```
caseno.begin  <- config$caseno$begin
caseno.end <- config$caseno$end
caseno.exclude <- config$caseno$exclude

print(caseno.begin)
```

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

```
print(caseno.end)
```

```
## [1] 183
```

```
print(caseno.exclude)
```

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

3.8 Debugging Mode

The debugging mode will reduce the number of documents compiled significantly. The full complement of cases takes approximately 11 hours to process with 16 threads on a Ryzen 3700X. The reduced complement captures a variety of cases with key characteristics that are useful in testing all features. Testing should always include cases 116 and 146 or an error will occur.

In addition to the mandatory test cases debugging mode will draw two random samples of size *debug.sample*, one from older and one from more recent cases of the ICJ.

```
mode.debug.toggle <- config$debug$toggle
mode.debug.sample <- config$debug$sample

print(mode.debug.toggle)
```

```
## [1] FALSE
```

```
print(mode.debug.sample)
```

```
## [1] 3
```

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("EN_PDF_ORIGINAL_FULL",
            "FR_PDF_ORIGINAL_FULL",
            "EN_PDF_ENHANCED_max2004",
            "FR_PDF_ENHANCED_max2004",
            "EN_PDF_BEST_FULL",
            "FR_PDF_BEST_FULL",
            "EN_PDF_BEST_MajorityOpinions",
            "FR_PDF_BEST_MajorityOpinions",
            "EN_TXT_BEST_FULL",
            "FR_TXT_BEST_FULL",
            "EN_TXT_TESSERACT_max2004",
            "FR_TXT_TESSERACT_max2004",
            "EN_TXT_EXTRACTED_FULL",
            "FR_TXT_EXTRACTED_FULL")
```

4.2 Directory for Unlabelled Files

```
dir.unlabelled <- paste(datashort,
                        datestamp,
                        "UnlabelledFiles",
                        sep = "_")
```

4.3 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(dir.unlabelled, recursive = TRUE)
unlink("temp", recursive = TRUE)
```

4.4 Create directories

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

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

5 LaTeX Configuration

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}{",
                      datestamp,
                      "}"),
               "\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-ICJ_Source_TEX_Definitions.tex")
```

5.1 Write Package Citations

```
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
```

6.2 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
```

6.3 Data.table

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

6.4 Quanteda

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

7 Visualize Corpus Creation Process

7.1 Workflow Part 1

```
workflow1 <- "
digraph workflow {

  # a '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]

  100 [label = 'www.icj-cij.org']
  101 [label = 'Links to Raw PDF Files']
  102 [label = 'Unlabelled Files']
  103 [label = 'Labelling Information']
  104 [label = 'Labelled PDF Files']
  105 [label = 'Handcoded Case Names']

  106 [label = 'EN_PDF_ORIGINAL_FULL']
  107 [label = 'EN_TXT_EXTRACTED']
  108 [label = 'EN_TXT_TESSERACT_max2004']
  109 [label = 'EN_PDF_ENHANCED_Max2004']
  110 [label = 'EN_TXT_BEST']
  111 [label = 'EN_PDF_BEST_FULL']
  112 [label = 'EN_PDF_BEST_MajorityOpinions']

  113 [label = 'FR_PDF_ORIGINAL_FULL']
  114 [label = 'FR_TXT_EXTRACTED']
  115 [label = 'FR_TXT_TESSERACT_max2004']
  116 [label = 'FR_PDF_ENHANCED_Max2004']
  117 [label = 'FR_TXT_BEST']
  118 [label = 'FR_PDF_BEST_FULL']
  119 [label = 'FR_PDF_BEST_MajorityOpinions']


  # Action Nodes

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

  200 [label = 'Extract Links from HTML']
  201 [label = 'Detect Unlabelled Files']
  202 [label = 'Download Unlabelled Files']
```

```
  203 [label = 'Handcoding of Labels']
  204 [label = 'Apply Labelling']
  205 [label = 'Strict REGEX Validation: ICJ File Name Schema']
  206 [label = 'Download Module']
  207 [label = 'File Split Module']
  208 [label = 'Filename Enhancement Module']
  209 [label = 'Strict REGEX Validation: Codebook File Name Schema']
  210 [label = 'Detect Missing Language Counterparts']
  211 [label = 'Text Extraction Module']
  212 [label = 'Tesseract OCR Module']
  213 [label = 'Create Majority Variant']


  # Edge Statements
  100 -> 200 -> 101 -> 201 -> 202 -> 102
  102 -> 203 -> 103
  {101, 103} -> 204 -> 205 -> 206 -> 104  -> 207 -> 208 -> 209 -> {106,113} ->
    210 -> {211, 212}
  105 -> 208
  211 -> {107, 114}
  212 -> {108, 109, 115, 116}
  {107, 108} -> 110
  {106, 109} -> 111
  {114, 115} -> 117
  {113, 115} -> 118
  111 -> 213 -> 112
  118 -> 213 -> 119

}
"



grViz(workflow1) %>% export_svg %>% charToRaw %>% rsvg_pdf("ANALYSIS/CD-ICJ_
    Workflow_1.pdf")
grViz(workflow1) %>% export_svg %>% charToRaw %>% rsvg_png("ANALYSIS/CD-ICJ_
    Workflow_1.png")
```

7.2 Workflow Part 2

```
workflow2 <- "
digraph workflow {

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

  # Data Nodes

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

  100 [label = 'EN_TXT_BEST']
  101 [label = 'FR_TXT_BEST']
  102 [label = 'EN_TXT_EXTRACTED']
```

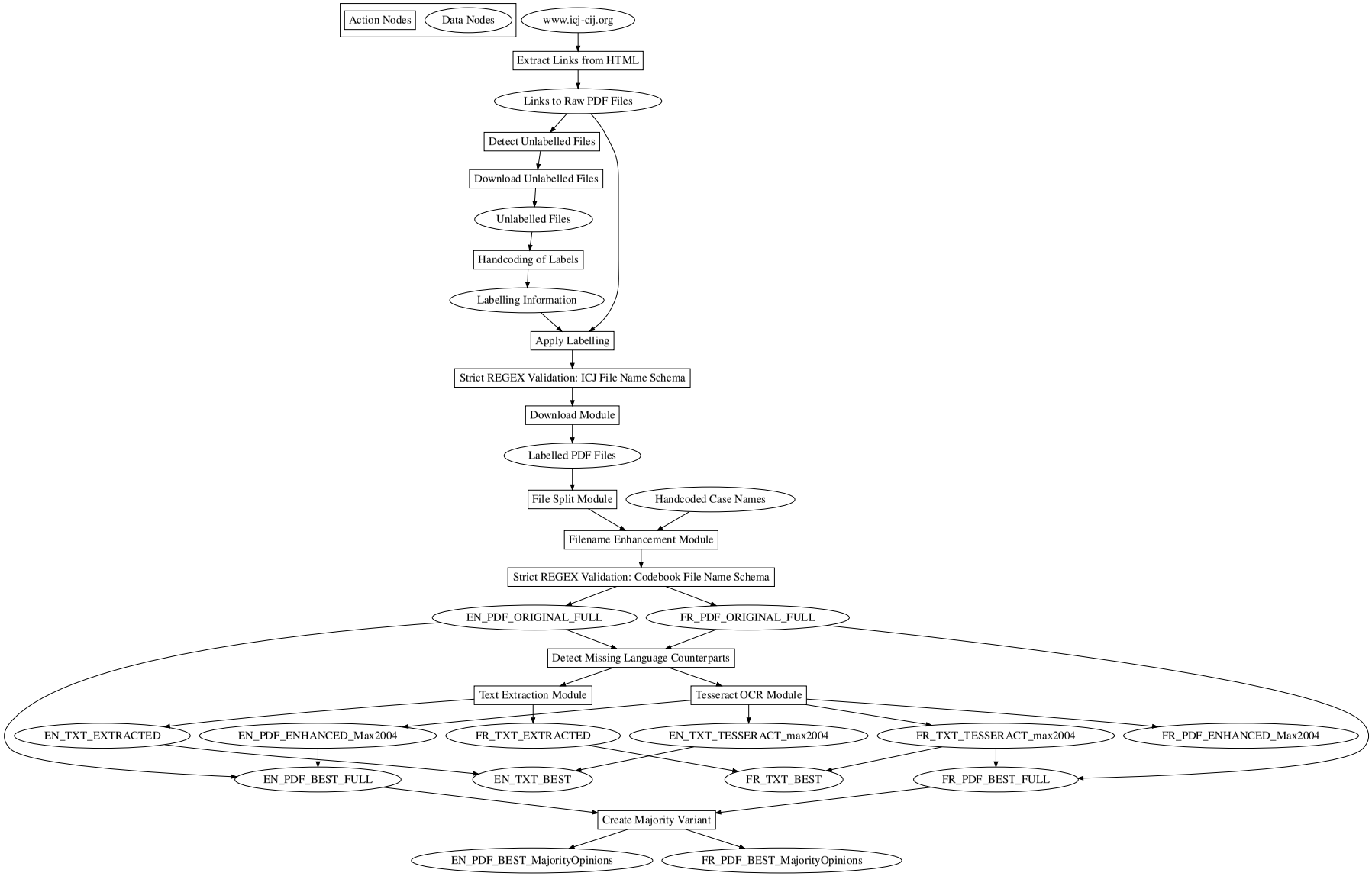


Figure 1: Workflow Part 1: Download, Labelling, Conversion and Sorting of Documents

```
  103 [label = 'FR_TXT_EXTRACTED']

  104 [label = 'EN_CSV_BEST_FULL']
  105 [label = 'FR_CSV_BEST_FULL']
  106 [label = 'EN_CSV_BEST_META']
  107 [label = 'FR_CSV_BEST_META']

  108 [label = 'ANALYSIS']
  109 [label = 'Frequency Tables']


  # Action Nodes

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

  200 [label = 'OCR Quality Control Module']
  201 [label = 'Clean Texts']
  202 [label = 'Language Purity Module']
  203 [label = 'Add Metadata']
  204 [label = 'Calculate Frequency Tables']
  205 [label = 'Visualize Frequency Tables']
  206 [label = 'Calculate and Add Summary Statistics']
  207 [label = 'Calculate Token Frequencies']
  208 [label = 'Calculate Document Similarity']
  209 [label = 'Write CSV Files']


  # Edge Statements

  {100, 101, 102, 103} -> 200
  {100, 101} -> 201 -> 202 -> 203
  203 -> 204 -> 109 -> 205
  203 -> 206 -> 209
  203 -> {207, 208}
  {109, 204, 205, 206, 207, 208} -> 108
  209 -> {104, 105, 106, 107}


}
"

grViz(workflow2) %>% export_svg %>% charToRaw %>% rsvg_pdf("ANALYSIS/CD-ICJ_
    Workflow_2.pdf")
grViz(workflow2) %>% export_svg %>% charToRaw %>% rsvg_png("ANALYSIS/CD-ICJ_
    Workflow_2.png")
```

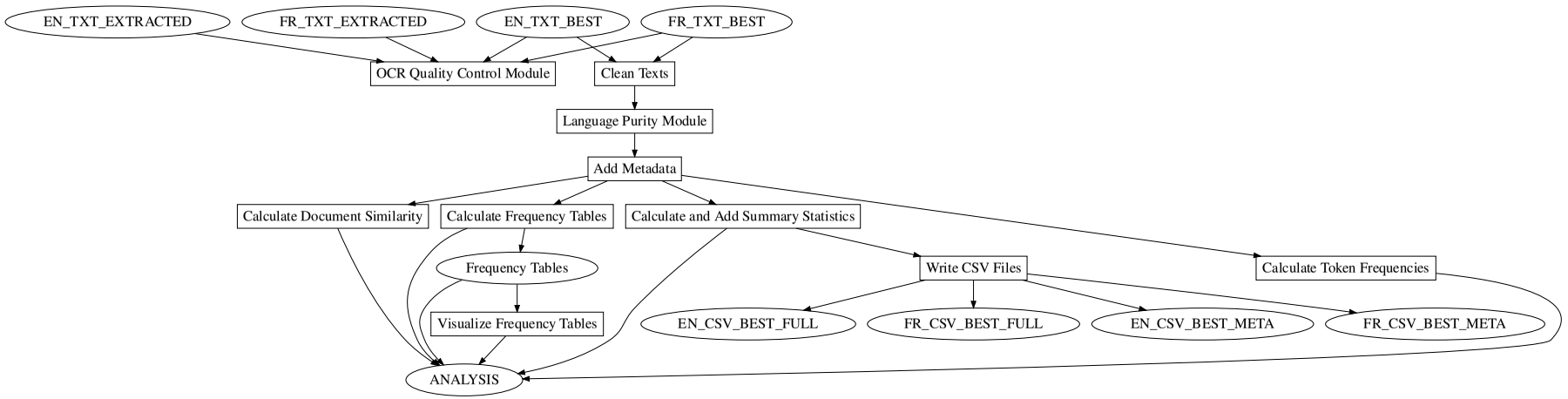


Figure 2: Workflow Part 2: Ingestion, Pre-Processing, Analysis and Creation of CSV Files

8 Prepare Download

8.1 Define Download Scope

```
caseno.full <- setdiff(caseno.begin:caseno.end,
                       caseno.exclude)
```

8.2 Debugging Mode — Reduced Scope

```
if(mode.debug.toggle == TRUE){
    caseno.full <- c(sample(3:41,
                            mode.debug.sample),
                     116,
                     146,
                     152,
                     sample(153:caseno.end,
                            mode.debug.sample),
                     175)
    caseno.full <- sort(caseno.full)
    }
```

8.3 Show Function: f.linkextract

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

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

8.4 Show Function: f.selectpdflinks

```
print(f.selectpdflinks)
```

```
## function(links){
##     temp <- grep ("case-related",
##                   links,
```

```
##                     ignore.case = TRUE,
##                     value = TRUE)
##     out <- grep ("BI.pdf",
##                  temp,
##                  ignore.case = TRUE,
##                  invert = TRUE,
##                  value = TRUE)
##     return(out)
## }
```

8.5 Prepare Empty Link List

```
links.list <- vector("list",
                     caseno.end)
```

8.6 Acquire Download Links

```
for (caseno in caseno.full) {

    URL.JUD <- sprintf("https://www.icj-cij.org/en/case/%d/judgments",
                    caseno)

    volatile <- f.linkextract(URL.JUD)
    links.jud <- f.selectpdflinks(volatile)


    URL.ORD <- sprintf("https://www.icj-cij.org/en/case/%d/orders",
                    caseno)

    volatile <- f.linkextract(URL.ORD)
    links.ord <- f.selectpdflinks(volatile)


    URL.ADV <- sprintf("https://www.icj-cij.org/en/case/%d/advisory-opinions",
                    caseno)

    volatile <- f.linkextract(URL.ADV)
    links.adv <- f.selectpdflinks(volatile)


    links.list[[caseno]] <- c(links.jud,
                              links.ord,
                              links.adv)
    print(caseno)

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

}
```

```
## [1] 1
## [1] 3
## [1] 4
## [1] 5
## [1] 6
## [1] 7
## [1] 8
## [1] 9
## [1] 10
## [1] 11
## [1] 12
## [1] 13
## [1] 14
## [1] 15
## [1] 16
## [1] 17
## [1] 18
## [1] 19
## [1] 20
## [1] 21
## [1] 22
## [1] 23
## [1] 24
## [1] 25
## [1] 26
## [1] 27
## [1] 28
## [1] 29
## [1] 30
## [1] 31
## [1] 32
## [1] 33
## [1] 34
## [1] 35
## [1] 36
## [1] 37
## [1] 38
## [1] 39
## [1] 40
## [1] 41
## [1] 42
## [1] 43
## [1] 44
## [1] 45
## [1] 46
## [1] 47
## [1] 48
## [1] 49
## [1] 50
## [1] 51
## [1] 52
## [1] 53
## [1] 54
## [1] 55
## [1] 56
```

```
## [1] 57
## [1] 58
## [1] 59
## [1] 60
## [1] 61
## [1] 62
## [1] 63
## [1] 64
## [1] 65
## [1] 66
## [1] 67
## [1] 68
## [1] 69
## [1] 70
## [1] 71
## [1] 72
## [1] 73
## [1] 74
## [1] 75
## [1] 76
## [1] 77
## [1] 78
## [1] 79
## [1] 80
## [1] 81
## [1] 82
## [1] 83
## [1] 84
## [1] 85
## [1] 86
## [1] 87
## [1] 88
## [1] 89
## [1] 90
## [1] 91
## [1] 92
## [1] 93
## [1] 94
## [1] 95
## [1] 96
## [1] 97
## [1] 98
## [1] 99
## [1] 100
## [1] 101
## [1] 102
## [1] 103
## [1] 104
## [1] 105
## [1] 106
## [1] 107
## [1] 108
## [1] 109
## [1] 110
## [1] 111
## [1] 112
```

```
## [1] 113
## [1] 114
## [1] 115
## [1] 116
## [1] 117
## [1] 118
## [1] 119
## [1] 120
## [1] 121
## [1] 122
## [1] 123
## [1] 124
## [1] 125
## [1] 126
## [1] 127
## [1] 128
## [1] 129
## [1] 130
## [1] 131
## [1] 132
## [1] 133
## [1] 134
## [1] 135
## [1] 136
## [1] 137
## [1] 138
## [1] 139
## [1] 140
## [1] 141
## [1] 142
## [1] 143
## [1] 144
## [1] 145
## [1] 146
## [1] 147
## [1] 148
## [1] 149
## [1] 150
## [1] 151
## [1] 152
## [1] 153
## [1] 154
## [1] 155
## [1] 156
## [1] 157
## [1] 158
## [1] 159
## [1] 160
## [1] 161
## [1] 162
## [1] 163
## [1] 164
## [1] 165
## [1] 166
## [1] 167
## [1] 168
```

```
## [1] 169
## [1] 170
## [1] 171
## [1] 172
## [1] 173
## [1] 174
## [1] 175
## [1] 176
## [1] 177
## [1] 178
## [1] 179
## [1] 180
## [1] 181
## [1] 182
## [1] 183
```

8.7 Clean Links

```
links <- unlist(links.list)

links.unique <- unique(links)

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

8.8 Remove Specific Links

Note 1: All files related to the advisory opinion in Case 146 are bilingual, even the supposedly monolingual variants. This removes the monolingual variants without replacement. True monolingual variants will be generated via splitting the bilingual variants at a later stage.

Note 2: The French files for cases 89, 125 and 156 are in fact mislabelled English variants. No French variants of the document are available on the website and even the bilingual variants are in fact entirely in English.

```
f1 <- "(089-19990629-ORD-01-00-FR)"
f2 <- "(125-20040709-ORD-01-00-FR)"
f3 <- "(146-20120201-ADV-01-00)"
f4 <- "(156-20150422-ORD-01-01-FR)"

links.download <- grep(paste(f1, f2, f3, f4, sep = "|"),
                       links.download,
                       invert = TRUE,
                       value = TRUE)
```

8.9 Add Specific Links

All files related to the advisory opinion in Case 146 are bilingual, even the supposedly monolingual variants. This adds the official bilingual advisory opinion and adds the

bilingual appended opinions which were not included in the original link list. These files will be split into monolingual variants at a later stage of the script.

```
links.download <- c(links.download,
                    "https://www.icj-cij.org/public/files/case-related/146/
    146-20120201-ADV-01-00-BI.pdf",
                    "https://www.icj-cij.org/public/files/case-related/146/
    146-20120201-ADV-01-01-BI.pdf",
                    "https://www.icj-cij.org/public/files/case-related/146/
    146-20120201-ADV-01-02-BI.pdf")
```

9 Labelling Module

Almost two dozen ICJ documents are unlabelled, i.e. they are provided with a computer-generated number only. Their filenames encode no semantic information. This module corrects the filenames and applies the standard naming scheme employed by the ICJ.

9.1 List Unlabelled Files

```
unlabelled.temp <- grep("EN|FR|BI",
                        links.unique,
                        invert = TRUE,
                        value = TRUE)

unlabelled.out <- data.table(sort(unlabelled.temp),
                             sort(unlabelled.temp))

print(unlabelled.temp)
```

```
##  [1] "/public/files/case-related/150/18852.pdf"
##  [2] "/public/files/case-related/152/18850.pdf"
##  [3] "/public/files/case-related/152/18852.pdf"
##  [4] "/public/files/case-related/152/18854.pdf"
##  [5] "/public/files/case-related/152/18856.pdf"
##  [6] "/public/files/case-related/152/18858.pdf"
##  [7] "/public/files/case-related/152/18860.pdf"
##  [8] "/public/files/case-related/152/18862.pdf"
##  [9] "/public/files/case-related/152/18864.pdf"
## [10] "/public/files/case-related/152/18867.pdf"
## [11] "/public/files/case-related/152/18868.pdf"
## [12] "/public/files/case-related/153/18748.pdf"
## [13] "/public/files/case-related/153/18749.pdf"
## [14] "/public/files/case-related/153/18750.pdf"
## [15] "/public/files/case-related/153/18751.pdf"
## [16] "/public/files/case-related/153/18752.pdf"
## [17] "/public/files/case-related/153/18753.pdf"
## [18] "/public/files/case-related/153/18754.pdf"
## [19] "/public/files/case-related/153/18755.pdf"
## [20] "/public/files/case-related/156/18638.pdf"
## [21] "/public/files/case-related/156/18640.pdf"
```

9.2 Write to Disk

```
fwrite(unlabelled.out,
       paste0(dir.unlabelled,
              "/",
              datashort,
              "_",
              datestamp,
              "_",
              "UnlabelledFiles.csv"))
```

9.3 Download Unlabelled Files

This is to prepare manual inspection and coding of unlabelled files.

9.3.1 Prepare

```
unlabelled.download.url <- paste0("https://www.icj-cij.org",
                                  unlabelled.temp)

unlabelled.download.name <- gsub("\\/", "\\_",
                                 unlabelled.temp)

unlabelled.download.name <- sub("\\_", "",
                                unlabelled.download.name)

dt <- data.table(unlabelled.download.url,
                 unlabelled.download.name)
```

9.3.2 Number of Unlabelled Files to Download

```
dt[,.N]
```

```
## [1] 21
```

9.3.3 Timestamp (Unlabelled Download Begin)

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

```
## [1] "2022-09-07 01:16:23 CEST"
```

9.3.4 Execute Download

Note: There is no download retry for this section, as these files are always inspected manually.

```
for (i in sample(dt[,.N])){
    download.file(dt$unlabelled.download.url[i],
                  dt$unlabelled.download.name[i])
    Sys.sleep(runif(1, 0.5, 1.5))
    }
```

9.3.5 Timestamp (Unlabelled Download End)

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

```
## [1] "2022-09-07 01:16:48 CEST"
```

9.3.6 Duration (Download)

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

```
## Time difference of 24.66496 secs
```

9.4 Download Result

9.4.1 Number of Files to Download

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

```
## [1] 21
```

9.4.2 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] 21
```

9.4.3 Number of Missing Files

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

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

9.4.4 Names of Missing Files

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

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

9.5 Store Unlabelled Files

```
file_move(files.pdf,
          dir.unlabelled)
```

9.6 Manual Coding

```
##########################################
###  HANDCODING OF UNLABELLED FILES
##########################################
```

9.7 Read in Corrected Labels

```
unlabelled.in <- fread("data/CD-ICJ_Source_UnlabelledFilesHandcoded.csv",
                       header = TRUE)
```

9.8 Apply Correct Labels to Link List

```
links.corrected <- mgsub(links.download,
                         unlabelled.in$old,
                         unlabelled.in$new)
```

9.9 Correct Underscores

```
links.corrected <- gsub("_", "-", links.corrected)
```

9.10 Correct Date Error

```
links.corrected <- gsub("202206613", "20220613", links.corrected)
```

9.11 REGEX VALIDATION 1: Strictly Validate Links against ICJ Naming Scheme

Test strict compliance of proposed download names with naming scheme used by ICJ. The result of a successful test should be an empty character vector!

9.11.1 Execute Validation

```
regex.test1 <- grep(paste0("^[0-9]{3}", # var: caseno
                           "_",
                           "[0-9]{8}", # var: date
                           "_",
                           "(JUD|ADV|ORD)", # var: doctype
                           "_",
                           "[0-9]{2}", # var: collision
                           "_",
                           "[0-9]{2}", # var: opinion
                           "_",
                           "(EN|FR|BI)", # var: language
                           ".pdf$"), # file extension,
                    basename(links.corrected),
                    invert = TRUE,
                    value = TRUE)
```

9.11.2 Results of Validation

```
print(regex.test1)
```

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

9.11.3 Stop Script on Failure

```
if (length(regex.test1) != 0){
    stop("REGEX VALIDATION 1 FAILED: LINKS NOT IN COMPLIANCE WITH ICJ SCHEMA!")
    }
```

9.12 Detect Duplicate Filenames

```
links.corrected[duplicated(links.corrected)]
```

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

9.13 Detect Missing Counterparts for each Language Version

```
linknames.en <- grep("EN.pdf",
                     links.corrected,
                     value=TRUE)

linknames.fr <- grep("FR.pdf",
                     links.corrected,
                     value=TRUE)
```

9.14 Difference in Number of Files

```
length(linknames.en) - length(linknames.fr)
```

```
## [1] 18
```

9.15 Show Missing French Documents

```
linknames.fr.temp <- gsub("FR",
                          "EN",
                          linknames.fr)

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

frenchmissing <- gsub("EN",
                      "FR",
                      frenchmissing)

print(frenchmissing)
```

```
##  [1] "https://www.icj-cij.org/public/files/case-related/89/089-19990629-ORD
    -01-00-FR.pdf"
##  [2] "https://www.icj-cij.org/public/files/case-related/116/116-20220209-JUD
    -01-02-FR.pdf"
##  [3] "https://www.icj-cij.org/public/files/case-related/116/116-20220209-JUD
    -01-03-FR.pdf"
##  [4] "https://www.icj-cij.org/public/files/case-related/116/116-20220209-JUD
    -01-05-FR.pdf"
##  [5] "https://www.icj-cij.org/public/files/case-related/125/125-20040709-ORD
    -01-00-FR.pdf"
##  [6] "https://www.icj-cij.org/public/files/case-related/155/155-20220421-JUD
    -01-01-FR.pdf"
##  [7] "https://www.icj-cij.org/public/files/case-related/155/155-20220421-JUD
    -01-02-FR.pdf"
```

```
##  [8] "https://www.icj-cij.org/public/files/case-related/155/155-20220421-JUD
    -01-05-FR.pdf"
##  [9] "https://www.icj-cij.org/public/files/case-related/155/155-20220421-JUD
    -01-06-FR.pdf"
## [10] "https://www.icj-cij.org/public/files/case-related/155/155-20220421-JUD
    -01-07-FR.pdf"
## [11] "https://www.icj-cij.org/public/files/case-related/155/155-20220421-JUD
    -01-08-FR.pdf"
## [12] "https://www.icj-cij.org/public/files/case-related/155/155-20220421-JUD
    -01-09-FR.pdf"
## [13] "https://www.icj-cij.org/public/files/case-related/155/155-20220421-JUD
    -01-10-FR.pdf"
## [14] "https://www.icj-cij.org/public/files/case-related/156/156-20150422-ORD
    -01-01-FR.pdf"
## [15] "https://www.icj-cij.org/public/files/case-related/178/178-20220722-JUD
    -01-01-FR.pdf"
## [16] "https://www.icj-cij.org/public/files/case-related/178/178-20220722-JUD
    -01-02-FR.pdf"
## [17] "https://www.icj-cij.org/public/files/case-related/182/182-20220316-ORD
    -01-01-FR.pdf"
## [18] "https://www.icj-cij.org/public/files/case-related/182/182-20220316-ORD
    -01-03-FR.pdf"
## [19] "https://www.icj-cij.org/public/files/case-related/182/182-20220316-ORD
    -01-04-FR.pdf"
## [20] "https://www.icj-cij.org/public/files/case-related/182/182-20220316-ORD
    -01-05-FR.pdf"
```

9.16 Show Missing English Documents

```
linknames.en.temp <- gsub("EN",
                          "FR",
                          linknames.en)

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

englishmissing <- gsub("FR",
                       "EN",
                       englishmissing)

print(englishmissing)
```

```
## [1] "https://www.icj-cij.org/public/files/case-related/116/116-20220209-JUD
    -01-06-EN.pdf"
## [2] "https://www.icj-cij.org/public/files/case-related/155/155-20220421-JUD
    -01-03-EN.pdf"
```

10 Download Module

10.1 Prepare Download Table

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

setnames(dt,
         new = c("links.download",
                 "names.download"))
```

10.2 Timestamp (Download Begin)

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

```
## [1] "2022-09-07 01:16:50 CEST"
```

10.3 Execute Download (All Files)

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

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

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

}
```

10.4 Timestamp (Download End)

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

```
## [1] "2022-09-07 02:46:38 CEST"
```

10.5 Duration (Download)

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

```
## Time difference of 1.496595 hours
```

10.6 Debugging Mode — Delete Random Files

This section deletes random files to test the result calculations and retry mode.

```
if (mode.debug.toggle == TRUE){

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

    unlink(sample(files.pdf, 5))

}
```

10.7 Download Result

10.7.1 Number of Files to Download

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

```
## [1] 4409
```

10.7.2 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] 4409
```

10.7.3 Number of Missing Files

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

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

10.7.4 Names of Missing Files

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

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

10.8 Timestamp (Retry Download Begin)

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

```
## [1] "2022-09-07 02:46:38 CEST"
```

10.9 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))
    }
}
```

10.10 Timestamp (Retry Download End)

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

```
## [1] "2022-09-07 02:46:38 CEST"
```

10.11 Duration (Retry Download)

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

```
## Time difference of 0.01119041 secs
```

10.12 Retry Result

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

10.12.1 Successful during Retry

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

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

10.12.2 Missing after Retry

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

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

10.13 Final Download Result

10.13.1 Number of Files to Download

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

```
## [1] 4409
```

10.13.2 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] 4409
```

10.13.3 Number of Missing Files

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

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

10.13.4 Names of Missing Files

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

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

11 File Split Module

11.1 Armed Activities Order

Note: this file contains the correct French document, but also an appended opinion in English, which is already correctly located in another file. Therefore the appended opinion is simply removed from the file.

```
filename <- "116-20161206-ORD-01-00-FR.pdf"

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

file.rename(filename, file.temp)
```

```
## [1] TRUE
```

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

```
## [1] "116-20161206-ORD-01-00-FR.pdf"
```

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

11.2 Case 146

Note: The files for the Advisory Opinion and appended opinions of Case 146 are all bilingual, including the supposedly monolingual versions. These need to be split into their component language versions. English is assumed to be on even pages for the majority opinion and on odd pages for the appended opinions. Both processes are looped in case further documents in need of splitting are discovered.

11.2.1 English on Even Pages

```
even.english <-  c("146-20120201-ADV-01-00-BI.pdf")

for (file in even.english){
    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)
    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)
    pdf_subset(file,
                pages = odd,
                output = odd.name)
}
```

11.2.2 English on Odd Pages

```
odd.english <- c("146-20120201-ADV-01-01-BI.pdf",
                 "146-20120201-ADV-01-02-BI.pdf")

for (file in odd.english){
    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)
    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)
    pdf_subset(file,
                pages = odd,
                output = odd.name)
}
```

11.2.3 Delete Bilingual Files

```
unlink(even.english)
unlink(odd.english)
```

11.3 Amity Treaty Order

Note: this file is bilingual. The English pages are removed manually.

```
filename <- "175-20210721-ORD-01-00-FR.pdf"

file.temp <- paste0(filename,
```

```
                      "-temp")

file.rename(filename, file.temp)
```

```
## [1] TRUE
```

```
pdf_subset(file.temp, c(1:4, 6, 8), filename)
```

```
## [1] "175-20210721-ORD-01-00-FR.pdf"
```

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

12 Filename Enhancement Module

This module applies a number of enhancements to the filenames:

- Better separators
- Case names
- Applicant ISO codes
- Respondent ISO codes
- Stage of proceedings

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

12.1 Enhance Syntax

```
filenames.enhanced1 <- gsub(paste0("([0-9]{3})", # var: caseno
                                   "_",
                                   "([0-9]{4})([0-9]{2})([0-9]{2})", # var: date
                                   "_",
                                   "([A-Z]{3})", # var: doctype
                                   "_",
                                   "([0-9]{2})", # var: collision
                                   "_",
                                   "([0-9]{2})", # var: opinion
                                   "_",
                                   "([A-Z]{2})"), # var: language
                            "\\1_\\2-\\3-\\4_\\5_\\6_\\7_\\8",
                            filenames.original)
```

12.2 Manual Coding

```
########## HAND CODING ####################
### - CASENAMES
### - Applicant Codes
### - Respondent Codes
### - Stage of Proceedings
###########################################
```

12.3 Read Hand Coded Data

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

12.4 Add Hand Coded Data to Filenames

Case names, Applicant codes and Respondent codes have been hand coded and are added in this step.

```
caseno.pad <- formatC(casenames$caseno,
                      width = 3,
                      flag = "0")

case.header <- paste0("ICJ_",
                      caseno.pad,
                      "_",
                      casenames$casename_short,
                      "_")

filenames.enhanced2 <- mgsub(filenames.enhanced1,
                             paste0("^",
                                    caseno.pad,
                                    "\\_"),
                             case.header)
```

12.5 Add Stage of Proceedings

```
stage <- fread("data/CD-ICJ_Source_Stages_Filenames.csv")



files <- list.files("CD-ICJ_2021-07-12_EN_TXT_BEST_FULL/EN_TXT_BEST_FULL/")

filenames.enhanced3 <- mgsub(filenames.enhanced2,
                             stage$old,
                             stage$new)

filenames.enhanced3 <- gsub("([0-9]{4}-[0-9]{2}-[0-9]{2}_[A-Z]{3}_[0-9]{2})(_
    [0-9]{2})",
                            "\\1_NA\\2",
                            filenames.enhanced3)
```

12.6 REGEX VALIDATION 2: Strictly Validate Naming Scheme against Codebook Schema

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

12.6.1 Execute Validation

```
regex.test2 <- grep(paste0("^ICJ", # var: court
                           "_",
                           "[0-9]{3}", # var: caseno
                           "_",
                           "[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
                           "_",
                           "(JUD|ADV|ORD)", # var: doctype
                           "_",
                           "[0-9]{2}", # var: collision
                           "_",
                           "(NA|PO|ME|IN|CO)", # var: stage
                           "_",
                           "[0-9]{2}", # var: opinion
                           "_",
                           "(EN|FR)", # var: language
                           ".pdf$"), # file extension
                    filenames.enhanced3,
                    value = TRUE,
                    invert = TRUE)
```

12.6.2 Results of Validation

```
print(regex.test2)
```

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

12.6.3 Stop Script on Failure

```
if (length(regex.test2) != 0){
    stop("REGEX VALIDATION 2 FAILED: FILE NAMES NOT IN COMPLIANCE WITH CODEBOOK
    SCHEMA!")
    }
```

12.7 Execute Rename

```
file.rename(filenames.original,
            filenames.enhanced3)
```

13 Detect Missing Counterparts for each Language Variant

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

13.1 Difference between French and English File Lists

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

```
## [1] 18
```

13.2 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] "ICJ_089_Lockerbie_LBY_USA_1999-06-29_ORD_01_NA_00_FR.pdf"
##  [2] "ICJ_116_ArmedActivities_COD_UGA_2022-02-09_JUD_01_NA_02_FR.pdf"
##  [3] "ICJ_116_ArmedActivities_COD_UGA_2022-02-09_JUD_01_NA_03_FR.pdf"
##  [4] "ICJ_116_ArmedActivities_COD_UGA_2022-02-09_JUD_01_NA_05_FR.pdf"
##  [5] "ICJ_125_FrontierDispute_BEN_NER_2004-07-09_ORD_01_NA_00_FR.pdf"
##  [6] "ICJ_155_SovereignRightsCaribbeanSea_NIC_COL_2022-04-21_JUD_01_NA_01_FR.
    pdf"
##  [7] "ICJ_155_SovereignRightsCaribbeanSea_NIC_COL_2022-04-21_JUD_01_NA_02_FR.
    pdf"
##  [8] "ICJ_155_SovereignRightsCaribbeanSea_NIC_COL_2022-04-21_JUD_01_NA_05_FR.
    pdf"
##  [9] "ICJ_155_SovereignRightsCaribbeanSea_NIC_COL_2022-04-21_JUD_01_NA_06_FR.
    pdf"
## [10] "ICJ_155_SovereignRightsCaribbeanSea_NIC_COL_2022-04-21_JUD_01_NA_07_FR.
    pdf"
## [11] "ICJ_155_SovereignRightsCaribbeanSea_NIC_COL_2022-04-21_JUD_01_NA_08_FR.
    pdf"
## [12] "ICJ_155_SovereignRightsCaribbeanSea_NIC_COL_2022-04-21_JUD_01_NA_09_FR.
    pdf"
```

```
## [13] "ICJ_155_SovereignRightsCaribbeanSea_NIC_COL_2022-04-21_JUD_01_NA_10_FR.
    pdf"
## [14] "ICJ_156_CertainDocumentsSeizure_TLS_AUS_2015-04-22_ORD_01_NA_01_FR.pdf"
## [15] "ICJ_178_ApplicationGenocideConvention_GMB_MMR_2022-07-22_JUD_01_NA_01_FR
    .pdf"
## [16] "ICJ_178_ApplicationGenocideConvention_GMB_MMR_2022-07-22_JUD_01_NA_02_FR
    .pdf"
## [17] "ICJ_182_ApplicationGenocideConvention_UKR_RUS_2022-03-16_ORD_01_NA_01_FR
    .pdf"
## [18] "ICJ_182_ApplicationGenocideConvention_UKR_RUS_2022-03-16_ORD_01_NA_03_FR
    .pdf"
## [19] "ICJ_182_ApplicationGenocideConvention_UKR_RUS_2022-03-16_ORD_01_NA_04_FR
    .pdf"
## [20] "ICJ_182_ApplicationGenocideConvention_UKR_RUS_2022-03-16_ORD_01_NA_05_FR
    .pdf"
```

13.3 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] "ICJ_116_ArmedActivities_COD_UGA_2022-02-09_JUD_01_NA_06_EN.pdf"
## [2] "ICJ_155_SovereignRightsCaribbeanSea_NIC_COL_2022-04-21_JUD_01_NA_03_EN.
    pdf"
```

14 Text Extraction Module

14.1 Define Set of Files to Process

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

14.2 Number of Files to Process

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

```
## [1] 4412
```

14.3 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)
}
```

}

14.4 Count Pages

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

```
## [1] "Parallel processing using 16 threads."
## [1] "Total number of pages: 64147"
```

```
## [1] 64147
```

14.5 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)
```

}

14.6 Extract Text

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

```
## [1] "Parallel processing using 16 threads. Begin at 2022-09-07 02:46:59"
## [1] "Processed 4412 files. Runtime was 11.82 secs. Ended at 2022-09-07
    02:47:10."
```

14.7 Copy and Move EXTRACTED TXT Files

This step copies all extracted TXT files from 2005 and later, which are assumed to be born-digital, to the BEST variant TXT folder. It further moves all TXT files to the "EXTRACTED" folder.

```
txt.best.en <- list.files(pattern = "_(200[5-9]|201[0-9]|202[0-5])-.*EN\\.txt")
txt.best.fr <- list.files(pattern = "_(200[5-9]|201[0-9]|202[0-5])-.*FR\\.txt")

file_copy(txt.best.en,
          "EN_TXT_BEST_FULL")
file_copy(txt.best.fr,
          "FR_TXT_BEST_FULL")

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")
```

15 Tesseract OCR Module

15.1 Mark Files for OCR

Only files which were published in 2004 or earlier are marked for optical character recognition (OCR) processing. Files from 2005 onwards are assumed to be born-digital and of perfect quality.

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

files.ocr.en <- list.files(pattern = "_(19[4-8][0-9]|199[0-9]|200[0-4])-.*EN\\.
    pdf")
files.ocr.fr <- list.files(pattern = "_(19[4-8][0-9]|199[0-9]|200[0-4])-.*FR\\.
    pdf")
```

15.2 Copy and Move Born-Digital Files

```
files.pdf.best.en <- setdiff(files.pdf.en,
                             files.ocr.en)

files.pdf.best.fr <- setdiff(files.pdf.fr,
                             files.ocr.fr)

file_copy(files.pdf.best.en,
          "EN_PDF_BEST_FULL")
file_copy(files.pdf.best.fr,
          "FR_PDF_BEST_FULL")


file_move(files.pdf.best.en,
          "EN_PDF_ORIGINAL_FULL")
file_move(files.pdf.best.fr,
          "FR_PDF_ORIGINAL_FULL")
```

15.3 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)
```

}

15.4 English

15.4.1 Number of English Documents to Process

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

```
## [1] 1484
```

15.4.2 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: 20513"
```

```
## [1] 20513
```

15.4.3 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 2022-09-07 02:47:12"
## [1] "Processed 1484 files. Runtime was 2.76 hours. Ended at 2022-09-07
    05:32:48."
```

15.5 French

15.5.1 Number of French Documents to Process

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

```
## [1] 1482
```

15.5.2 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: 20502"
```

```
## [1] 20502
```

15.5.3 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 2022-09-07 05:32:48"
## [1] "Processed 1482 files. Runtime was 3.4 hours. Ended at 2022-09-07
    08:56:43."
```

15.6 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)
```

15.7 Copy and Move TXT Files

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

file_copy(files.ocr.txt.en,
          "EN_TXT_BEST_FULL")
file_copy(files.ocr.txt.fr,
          "FR_TXT_BEST_FULL")

file_move(files.ocr.txt.en,
          "EN_TXT_TESSERACT_max2004")
file_move(files.ocr.txt.fr,
          "FR_TXT_TESSERACT_max2004")
```

15.8 Copy and 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_copy(files.ocr.pdf.enhanced.en,
          "EN_PDF_BEST_FULL")
file_copy(files.ocr.pdf.enhanced.fr,
          "FR_PDF_BEST_FULL")

file_move(files.ocr.pdf.enhanced.en,
          "EN_PDF_ENHANCED_max2004")
file_move(files.ocr.pdf.enhanced.fr,
          "FR_PDF_ENHANCED_max2004")

file_move(files.ocr.pdf.original.en,
          "EN_PDF_ORIGINAL_FULL")
file_move(files.ocr.pdf.original.fr,
          "FR_PDF_ORIGINAL_FULL")
```

16 Create Majority-Only Variant

```
majonly.en <- list.files("EN_PDF_BEST_FULL",
                         full.names = TRUE,
                         pattern = "00_EN")

majonly.fr <- list.files("FR_PDF_BEST_FULL",
                         full.names = TRUE,
                         pattern = "00_FR")

file_copy(majonly.en,
          "EN_PDF_BEST_MajorityOpinions")
file_copy(majonly.fr,
          "FR_PDF_BEST_MajorityOpinions")
```

17 Read in TXT Files

17.1 Define Variable Names

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

17.2 BEST Variants

17.2.1 English

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

17.2.2 French

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

17.3 EXTRACTED Variants

17.3.1 English

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

17.3.2 French

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

17.4 Convert to Data Tables

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

18 Clean Texts

18.1 Remove Hyphenation across Linebreaks

Hyphenation across linebreaks is a serious issue in longer texts. 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.

18.1.1 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)
## }
```

18.1.2 Execute Function

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

18.2 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.

18.2.1 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)
## }
```

18.2.2 Execute Function

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

19 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.

Only documents from 2004 or earlier will be compared. This provides a more accurate measurement of the relative quality of the different OCR processes than if born-digital documents were to be included.

19.1 Create Corpora

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

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

19.2 Subset to 2004 and earlier

```
corpus.en.b.2004  <- corpus_subset(corpus.en.b, date < 2005)
corpus.en.e.2004  <- corpus_subset(corpus.en.e, date < 2005)

corpus.fr.b.2004  <- corpus_subset(corpus.fr.b, date < 2005)
corpus.fr.e.2004  <- corpus_subset(corpus.fr.e, date < 2005)
```

19.3 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)
##     }
```

19.4 Tokenize

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

tokens.en.b.2004 <- f.token.processor(corpus.en.b.2004)
tokens.en.e.2004 <- f.token.processor(corpus.en.e.2004)

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

tokens.fr.b.2004 <- f.token.processor(corpus.fr.b.2004)
tokens.fr.e.2004 <- f.token.processor(corpus.fr.e.2004)
```

19.5 Create Document-Feature-Matrices

```
dfm.en.b.2004 <- dfm(tokens.en.b.2004)
dfm.en.e.2004 <- dfm(tokens.en.e.2004)

dfm.fr.b.2004 <- dfm(tokens.fr.b.2004)
dfm.fr.e.2004 <- dfm(tokens.fr.e.2004)
```

19.6 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.

```
feat.languages <- c("English",
                    "French")

feat.extracted <- c(nfeat(dfm.en.e.2004),
                    nfeat(dfm.fr.e.2004))


feat.tesseract <- c(nfeat(dfm.en.b.2004),
                    nfeat(dfm.fr.b.2004))



feat.reduction.abs <- feat.extracted - feat.tesseract

feat.reduction.rel.pct <- (1 - (feat.tesseract / feat.extracted)) * 100


dt.ocrquality <- data.table(feat.languages,
                            feat.extracted,
                            feat.tesseract,
                            feat.reduction.abs,
                            paste(round(feat.reduction.rel.pct, 2), "%"))
```

```
kable(dt.ocrquality,
      format = "latex",
      align = "r",
      booktabs = TRUE,
      col.names = c("Language",
                    "Extracted Features",
                    "Tesseract Features",
                    "Difference (abs)",
                    "Difference (pct)"))
```

Language	Extracted Features	Tesseract Features	Difference (abs)	Difference (pct)
English	115706	56526	59180	51.15 %
French	135786	75403	60383	44.47 %

20 Language Purity Module

This module 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.

20.1 Limit Detection to English and French

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

20.2 Automatic Language Detection

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

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

20.3 Detected Languages

Note: Should only read 'english'

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

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

Note: Should only read 'french'

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

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

20.4 Show Mismatches

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

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

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

20.5 Final Note: Human Review of Mismatches

All documents flagged by textcat were reviewed and appropriate remedies devised. Some files were deleted from the corpus if no authentic language variant could be found. Monolingual files for case 146 are now generated from the bilingual originals. See the download section for details.

21 Add and Delete Variables

21.1 Delete Textcat Classifications

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

21.2 Add Variable "year"

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

21.3 Add Variable "minority"

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

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

21.4 Add Variable "fullname"

21.4.1 Read Hand Coded Data

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

21.4.2 Create Variable

```
data.best.en$fullname <- casenames$casename_full[match(data.best.en$caseno,
                                                       casenames$caseno)]

data.best.fr$fullname <- casenames$casename_full[match(data.best.fr$caseno,
                                                       casenames$caseno)]
```

21.5 Add Variable "applicant_region"

21.5.1 Read Hand Coded Data

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

21.5.2 Merge Regions for English Version

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

applicant_region <- gsub("CARAT|ECOSOC|IFAD|IMO|UNESCO|UNGA|UNSC|WHO",
                         "NA",
                         applicant_region)

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


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

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

21.5.3 Merge Regions for French Version

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

applicant_region <- gsub("CARAT|ECOSOC|IFAD|IMO|UNESCO|UNGA|UNSC|WHO",
                         "NA",
                         applicant_region)

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


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

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

21.6 Add Variable “respondent_region”

21.6.1 Read Hand Coded Data

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

21.6.2 Merge Regions for English Version

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

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

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

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

21.6.3 Merge Regions for French Version

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

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

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

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

21.7 Add Variable "applicant_subregion"

21.7.1 Read Hand Coded Data

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

21.7.2 Merge Subregions for English Version

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

applicant_subregion <- gsub("CARAT|ECOSOC|IFAD|IMO|UNESCO|UNGA|UNSC|WHO",
                            "NA",
                            applicant_subregion)

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


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

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

21.7.3 Merge Subregions for French Version

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

applicant_subregion <- gsub("CARAT|ECOSOC|IFAD|IMO|UNECO|UNGA|UNSC|WHO",
                            "NA",
                            applicant_subregion)

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


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

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

21.8 Add Variable "respondent_subregion"

21.8.1 Read Hand Coded Data

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

21.8.2 Merge Subregions for English Version

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

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

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

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

21.8.3 Merge Subregions for French Version

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

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

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

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

21.9 Add Variable "doi_concept"

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

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

21.10 Add Variable "doi_version"

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

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

21.11 Add Variable "version"

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

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

21.12 Add Variable "license"

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

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

22 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.

22.1 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)
}
```

}

22.2 English Corpus

22.2.1 Variables to Ignore

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

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

22.2.2 Variables to Analyze

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

varlist <- setdiff(varlist,
```

```
                    freq.var.ignore)

print(varlist)
```

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

22.2.3 Construct Frequency Tables

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

```
f.fast.freqtable(data.best.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
ICJ	2215	100	100	100
Total	2215	100	100	100

Frequency Table for Variable: caseno

182 unique value(s) detected.

caseno	N	exactpercent	roundedpercent	cumulpercent
1	19	0.8577878	0.86	0.86
3	7	0.3160271	0.32	1.17
4	7	0.3160271	0.32	1.49
5	9	0.4063205	0.41	1.90
6	1	0.0451467	0.05	1.94
7	9	0.4063205	0.41	2.35
8	10	0.4514673	0.45	2.80
9	4	0.1805869	0.18	2.98
10	7	0.3160271	0.32	3.30
11	6	0.2708804	0.27	3.57
12	4	0.1805869	0.18	3.75
13	1	0.0451467	0.05	3.79
14	2	0.0902935	0.09	3.88
15	16	0.7223476	0.72	4.60
16	12	0.5417607	0.54	5.15
17	7	0.3160271	0.32	5.46
18	11	0.4966140	0.50	5.96
19	7	0.3160271	0.32	6.28
20	3	0.1354402	0.14	6.41
21	6	0.2708804	0.27	6.68
22	1	0.0451467	0.05	6.73
23	1	0.0451467	0.05	6.77
24	5	0.2257336	0.23	7.00
25	1	0.0451467	0.05	7.04
26	1	0.0451467	0.05	7.09
27	1	0.0451467	0.05	7.13
28	1	0.0451467	0.05	7.18
29	11	0.4966140	0.50	7.67
30	9	0.4063205	0.41	8.08
31	6	0.2708804	0.27	8.35

(continued)

caseno	N	exactpercent	roundedpercent	cumulpercent
32	25	1.1286682	1.13	9.48
33	11	0.4966140	0.50	9.98
34	16	0.7223476	0.72	10.70
35	8	0.3611738	0.36	11.06
36	8	0.3611738	0.36	11.42
37	5	0.2257336	0.23	11.65
38	8	0.3611738	0.36	12.01
39	6	0.2708804	0.27	12.28
40	1	0.0451467	0.05	12.33
41	5	0.2257336	0.23	12.55
42	4	0.1805869	0.18	12.73
43	4	0.1805869	0.18	12.91
44	1	0.0451467	0.05	12.96
45	16	0.7223476	0.72	13.68
46	30	1.3544018	1.35	15.03
47	30	1.3544018	1.35	16.39
48	18	0.8126411	0.81	17.20
49	11	0.4966140	0.50	17.70
50	31	1.3995485	1.40	19.10
51	15	0.6772009	0.68	19.77
52	15	0.6772009	0.68	20.45
53	16	0.7223476	0.72	21.17
54	14	0.6320542	0.63	21.81
55	24	1.0835214	1.08	22.89
56	25	1.1286682	1.13	24.02
57	11	0.4966140	0.50	24.51
58	31	1.3995485	1.40	25.91
59	29	1.3092551	1.31	27.22
60	5	0.2257336	0.23	27.45

(continued)

caseno	N	exactpercent	roundedpercent	cumulpercent
61	15	0.6772009	0.68	28.13
62	20	0.9029345	0.90	29.03
63	14	0.6320542	0.63	29.66
64	7	0.3160271	0.32	29.98
65	11	0.4966140	0.50	30.47
66	11	0.4966140	0.50	30.97
67	12	0.5417607	0.54	31.51
68	22	0.9932280	0.99	32.51
69	8	0.3611738	0.36	32.87
70	32	1.4446953	1.44	34.31
71	5	0.2257336	0.23	34.54
72	10	0.4514673	0.45	34.99
73	3	0.1354402	0.14	35.12
74	11	0.4966140	0.50	35.62
75	20	0.9029345	0.90	36.52
76	6	0.2708804	0.27	36.79
77	6	0.2708804	0.27	37.07
78	13	0.5869074	0.59	37.65
79	9	0.4063205	0.41	38.06
80	11	0.4966140	0.50	38.56
81	5	0.2257336	0.23	38.78
82	15	0.6772009	0.68	39.46
83	8	0.3611738	0.36	39.82
84	10	0.4514673	0.45	40.27
85	1	0.0451467	0.05	40.32
86	7	0.3160271	0.32	40.63
87	29	1.3092551	1.31	41.94
88	28	1.2641084	1.26	43.21
89	26	1.1738149	1.17	44.38

(continued)

caseno	N	exactpercent	roundedpercent	cumulpercent
90	32	1.4446953	1.44	45.82
91	43	1.9413093	1.94	47.77
92	16	0.7223476	0.72	48.49
93	9	0.4063205	0.41	48.89
94	34	1.5349887	1.53	50.43
95	16	0.7223476	0.72	51.15
96	13	0.5869074	0.59	51.74
97	8	0.3611738	0.36	52.10
98	12	0.5417607	0.54	52.64
99	7	0.3160271	0.32	52.96
100	6	0.2708804	0.27	53.23
101	4	0.1805869	0.18	53.41
102	14	0.6320542	0.63	54.04
103	24	1.0835214	1.08	55.12
104	11	0.4966140	0.50	55.62
105	21	0.9480813	0.95	56.57
106	21	0.9480813	0.95	57.52
107	19	0.8577878	0.86	58.37
108	19	0.8577878	0.86	59.23
109	20	0.9029345	0.90	60.14
110	21	0.9480813	0.95	61.08
111	21	0.9480813	0.95	62.03
112	9	0.4063205	0.41	62.44
113	21	0.9480813	0.95	63.39
114	7	0.3160271	0.32	63.70
115	3	0.1354402	0.14	63.84
116	34	1.5349887	1.53	65.37
117	3	0.1354402	0.14	65.51
118	31	1.3995485	1.40	66.91

(continued)

caseno	N	exactpercent	roundedpercent	cumulpercent
119	7	0.3160271	0.32	67.22
120	8	0.3611738	0.36	67.58
121	22	0.9932280	0.99	68.58
122	6	0.2708804	0.27	68.85
123	8	0.3611738	0.36	69.21
124	35	1.5801354	1.58	70.79
125	7	0.3160271	0.32	71.11
126	16	0.7223476	0.72	71.83
127	4	0.1805869	0.18	72.01
128	11	0.4966140	0.50	72.51
129	11	0.4966140	0.50	73.00
130	9	0.4063205	0.41	73.41
131	11	0.4966140	0.50	73.91
132	4	0.1805869	0.18	74.09
133	6	0.2708804	0.27	74.36
134	1	0.0451467	0.05	74.40
135	20	0.9029345	0.90	75.30
136	11	0.4966140	0.50	75.80
137	13	0.5869074	0.59	76.39
138	4	0.1805869	0.18	76.57
139	8	0.3611738	0.36	76.93
140	16	0.7223476	0.72	77.65
141	11	0.4966140	0.50	78.15
142	8	0.3611738	0.36	78.51
143	15	0.6772009	0.68	79.19
144	17	0.7674944	0.77	79.95
145	3	0.1354402	0.14	80.09
146	5	0.2257336	0.23	80.32
147	1	0.0451467	0.05	80.36

(continued)

caseno	N	exactpercent	roundedpercent	cumulpercent
148	17	0.7674944	0.77	81.13
149	8	0.3611738	0.36	81.49
150	40	1.8058691	1.81	83.30
151	14	0.6320542	0.63	83.93
152	16	0.7223476	0.72	84.65
153	14	0.6320542	0.63	85.28
154	13	0.5869074	0.59	85.87
155	25	1.1286682	1.13	87.00
156	11	0.4966140	0.50	87.49
157	13	0.5869074	0.59	88.08
158	17	0.7674944	0.77	88.85
159	17	0.7674944	0.77	89.62
160	17	0.7674944	0.77	90.38
161	17	0.7674944	0.77	91.15
162	4	0.1805869	0.18	91.33
163	26	1.1738149	1.17	92.51
164	12	0.5417607	0.54	93.05
165	10	0.4514673	0.45	93.50
166	24	1.0835214	1.08	94.58
167	1	0.0451467	0.05	94.63
168	11	0.4966140	0.50	95.12
169	15	0.6772009	0.68	95.80
170	1	0.0451467	0.05	95.85
171	10	0.4514673	0.45	96.30
172	23	1.0383747	1.04	97.34
173	6	0.2708804	0.27	97.61
174	6	0.2708804	0.27	97.88
175	12	0.5417607	0.54	98.42
176	1	0.0451467	0.05	98.47

(continued)

caseno	N	exactpercent	roundedpercent	cumulpercent
177	3	0.1354402	0.14	98.60
178	11	0.4966140	0.50	99.10
179	2	0.0902935	0.09	99.19
180	5	0.2257336	0.23	99.41
181	3	0.1354402	0.14	99.55
182	8	0.3611738	0.36	99.91
183	2	0.0902935	0.09	100.00
Total	2215	100.0000000	100.00	100.00

Frequency Table for Variable: shortname

148 unique value(s) detected.

shortname	N	exactpercent	roundedpercent	cumulpercent
1955AmityTreaty	12	0.5417607	0.54	0.54
ATILO-UNESCO	9	0.4063205	0.41	0.95
AccessPacificOcean	14	0.6320542	0.63	1.58
AdmissionUN	7	0.3160271	0.32	1.90
AegeanSeaContinentalShelf	20	0.9029345	0.90	2.80
AerialHerbicideSpraying	4	0.1805869	0.18	2.98
AerialIncident1952	1	0.0451467	0.05	3.02
AerialIncident1953	1	0.0451467	0.05	3.07
AerialIncident1988	9	0.4063205	0.41	3.48
AerialIncident1999	7	0.3160271	0.32	3.79
AerialIncidentNov1954	1	0.0451467	0.05	3.84
AerialIncidentSept1954	1	0.0451467	0.05	3.88
AerialIndicent1955	21	0.9480813	0.95	4.83
Ambatielos	16	0.7223476	0.72	5.55
AngloIranianOil	12	0.5417607	0.54	6.09

(continued)

shortname	N	exactpercent	roundedpercent	cumulpercent
Antarctica	2	0.0902935	0.09	6.19
ApplicationCERD	31	1.3995485	1.40	7.58
ApplicationGenocideConvention	93	4.1986456	4.20	11.78
ApplicationGenocideConvention-Revision	6	0.2708804	0.27	12.05
ArbitralAward1899	10	0.4514673	0.45	12.51
ArbitralAward1989	15	0.6772009	0.68	13.18
ArbitralAwardKingOfSpain	6	0.2708804	0.27	13.45
ArbitrationUNHQAgreement	6	0.2708804	0.27	13.72
ArmedActivities	40	1.8058691	1.81	15.53
ArmedActivitiesApp2002	16	0.7223476	0.72	16.25
ArrestWarrant	22	0.9932280	0.99	17.25
Asylum	9	0.4063205	0.41	17.65
Asylum-Interpretation	1	0.0451467	0.05	17.70
Avena	11	0.4966140	0.50	18.19
Avena-Interpretation	8	0.3611738	0.36	18.56
BarcelonaTraction1958	5	0.2257336	0.23	18.78
BarcelonaTraction1962	31	1.3995485	1.40	20.18
CertainActivitiesBorderArea	40	1.8058691	1.81	21.99
CertainCriminalProceedings	11	0.4966140	0.50	22.48
CertainDocumentsSeizure	11	0.4966140	0.50	22.98
CertainExpensesUN	11	0.4966140	0.50	23.48
CertainPhosphateLands	11	0.4966140	0.50	23.97
CertainProperty	8	0.3611738	0.36	24.33
ChagosArchipelago	15	0.6772009	0.68	25.01
CompensationUNAT	6	0.2708804	0.27	25.28
CompetenceAdmissionGA	4	0.1805869	0.18	25.46
ConstitutionMaritimeSafetyCommittee	4	0.1805869	0.18	25.64
ConstructionWallOPT	11	0.4966140	0.50	26.14

(continued)

shortname	N	exactpercent	roundedpercent	cumulpercent
ContinentalShelf	36	1.6252822	1.63	27.77
ContinentalShelf-InterpretationRevision	5	0.2257336	0.23	27.99
ConventionPrivilegesImmunitiesUN	5	0.2257336	0.23	28.22
ConventionTerrorismFinancingCERD	24	1.0835214	1.08	29.30
CorfuChannel	19	0.8577878	0.86	30.16
DelimitationContinentalShelf	13	0.5869074	0.59	30.74
Diallo	24	1.0835214	1.08	31.83
DiplomaticEnvoyUN	1	0.0451467	0.05	31.87
DiplomaticRelations	1	0.0451467	0.05	31.92
ELSI	6	0.2708804	0.27	32.19
EastTimor	10	0.4514673	0.45	32.64
ElectriciteBeyrouth	3	0.1354402	0.14	32.78
Fisheries	9	0.4063205	0.41	33.18
FisheriesJurisdiction	62	2.7990971	2.80	35.98
FrenchNationalsEgypt	1	0.0451467	0.05	36.03
FrontierDispute	23	1.0383747	1.04	37.07
GabcikovoNagymaros	16	0.7223476	0.72	37.79
GuardianshipInfantsConvention	11	0.4966140	0.50	38.28
GuatemalaTerritorialInsularMaritimeClaim	3	0.1354402	0.14	38.42
GulfOfMaine	12	0.5417607	0.54	38.96
HayaDeLaTorre	2	0.0902935	0.09	39.05
ICAOCouncil	14	0.6320542	0.63	39.68
ICAOCouncil-CICA	6	0.2708804	0.27	39.95
ICAOCouncil-IASTA	6	0.2708804	0.27	40.23
ICERD	16	0.7223476	0.72	40.95
ImmunitiesCriminalProceedings	26	1.1738149	1.17	42.12
ImmunitySRCommHR	6	0.2708804	0.27	42.39
IndependenceDeclarationKosovo	11	0.4966140	0.50	42.89

(continued)

shortname	N	exactpercent	roundedpercent	cumulpercent
Interhandel	16	0.7223476	0.72	43.61
InterimAccord1995	8	0.3611738	0.36	43.97
IranianAssets	12	0.5417607	0.54	44.51
IslaPortillos	10	0.4514673	0.45	44.97
Jadhav	11	0.4966140	0.50	45.46
Judgment2867ATILO-IFAD	5	0.2257336	0.23	45.69
JudgmentsCivilCommercialMatters	3	0.1354402	0.14	45.82
JurisdictionalImmunities2008	15	0.6772009	0.68	46.50
JurisdictionalImmunities2022	2	0.0902935	0.09	46.59
KasikiliSedudu	12	0.5417607	0.54	47.13
LaGrand	11	0.4966140	0.50	47.63
LandIslandMaritimeFrontier	20	0.9029345	0.90	48.53
LandIslandMaritimeFrontier-Revision	4	0.1805869	0.18	48.71
LandMaritimeBoundary	34	1.5349887	1.53	50.25
LandMaritimeBoundary-Interpretation	4	0.1805869	0.18	50.43
LandMaritimeDelimitationSovereigntyIslands	2	0.0902935	0.09	50.52
LegalityNuclearWeaponsArmedConflict	9	0.4063205	0.41	50.93
LegalityThreatUseNuclearWeapons	16	0.7223476	0.72	51.65
Lockerbie	54	2.4379233	2.44	54.09
MaritimeDelimitation	30	1.3544018	1.35	55.44
MaritimeDelimitation-BlackSea	4	0.1805869	0.18	55.62
MaritimeDelimitation-CaribbeanPacific	13	0.5869074	0.59	56.21
MaritimeDelimitation-GreenlandJanMayen	13	0.5869074	0.59	56.79
MaritimeDelimitation-IndianOcean	17	0.7674944	0.77	57.56
MaritimeDispute	13	0.5869074	0.59	58.15

(continued)

shortname	N	exactpercent	roundedpercent	cumulpercent
MilitaryParamilitaryActivitiesNicaragua	32	1.4446953	1.44	59.59
MinquiersEcrehos	7	0.3160271	0.32	59.91
MonetaryGold	7	0.3160271	0.32	60.23
MutualAssistanceCriminalMatters	11	0.4966140	0.50	60.72
Namibia	16	0.7223476	0.72	61.44
NavigationalRights	6	0.2708804	0.27	61.72
NorthSeaContinentalShelf	30	1.3544018	1.35	63.07
NorthernCameroons	18	0.8126411	0.81	63.88
NorwegianLoans	11	0.4966140	0.50	64.38
Nottebohm	11	0.4966140	0.50	64.88
NuclearDisarmament	51	2.3024831	2.30	67.18
NuclearTests	60	2.7088036	2.71	69.89
NuclearTests-ExaminationSituation	8	0.3611738	0.36	70.25
ObligationProsecuteExtradite	17	0.7674944	0.77	71.02
OilPlatforms	32	1.4446953	1.44	72.46
PassageGreatBelt	7	0.3160271	0.32	72.78
PassageIndianTerritory	25	1.1286682	1.13	73.91
PeaceTreaties	10	0.4514673	0.45	74.36
PedraBranca	9	0.4063205	0.41	74.76
PedraBranca-Interpretation	1	0.0451467	0.05	74.81
PedraBranca-Revision	1	0.0451467	0.05	74.85
PetitionersComitteeSouthWestAfrica	6	0.2708804	0.27	75.12
PortBeyrouthSRO	4	0.1805869	0.18	75.30
PulpMills	20	0.9029345	0.90	76.21
RelocationEmbassyUSJerusalem	1	0.0451467	0.05	76.25
ReparationUN	7	0.3160271	0.32	76.57
ReservationsGenocideConvention	4	0.1805869	0.18	76.75
ReviewJudgment158UNAT	11	0.4966140	0.50	77.25

(continued)

shortname	N	exactpercent	roundedpercent	cumulpercent
ReviewJudgment273UNAT	11	0.4966140	0.50	77.74
ReviewJudgment333UNAT	10	0.4514673	0.45	78.19
SanJuanRiver	16	0.7223476	0.72	78.92
SilalaWaters	4	0.1805869	0.18	79.10
SouthWestAfrica	60	2.7088036	2.71	81.81
SovereignRightsCaribbeanSea	25	1.1286682	1.13	82.93
SovereigntyFrontierLand	8	0.3611738	0.36	83.30
SovereigntyPulau	14	0.6320542	0.63	83.93
StatusSouthWestAfrica	7	0.3160271	0.32	84.24
TemplePreahVihear	16	0.7223476	0.72	84.97
TemplePreahVihear-Interpretation	14	0.6320542	0.63	85.60
TerritorialDispute	43	1.9413093	1.94	87.54
TerritorialDispute-CaribbeanSea	8	0.3611738	0.36	87.90
TransborderArmedActions	14	0.6320542	0.63	88.53
TreatmentAirCrew	2	0.0902935	0.09	88.62
TrialPakistaniPOW	5	0.2257336	0.23	88.85
USDiplomaticStaffTehran	7	0.3160271	0.32	89.16
USNationalsMorocco	6	0.2708804	0.27	89.44
UseOfForce	179	8.0812641	8.08	97.52
ViennaConventionConsularRelations	7	0.3160271	0.32	97.83
VotingProcedureSouthWestAfrica	5	0.2257336	0.23	98.06
WHO-EgyptAgreement	11	0.4966140	0.50	98.56
WesternSahara	15	0.6772009	0.68	99.23
WhalingAntarctic	17	0.7674944	0.77	100.00
Total	2215	100.0000000	100.00	100.00

Frequency Table for Variable: applicant

81 unique value(s) detected.

applicant	N	exactpercent	roundedpercent	cumulpercent
ARG	20	0.9029345	0.90	0.90
ARM	5	0.2257336	0.23	1.13
AUS	48	2.1670429	2.17	3.30
AZE	3	0.1354402	0.14	3.43
BEL	64	2.8893905	2.89	6.32
BEN	7	0.3160271	0.32	6.64
BFA	16	0.7223476	0.72	7.36
BHR-EGY-ARE	6	0.2708804	0.27	7.63
BHR-EGY-SAU-ARE	6	0.2708804	0.27	7.90
BIH	49	2.2121896	2.21	10.11
BOL	14	0.6320542	0.63	10.74
BWA	12	0.5417607	0.54	11.29
CAN	12	0.5417607	0.54	11.83
CARAT	32	1.4446953	1.44	13.27
CHE	16	0.7223476	0.72	14.00
CHL	4	0.1805869	0.18	14.18
CMR	56	2.5282167	2.53	16.70
COD	89	4.0180587	4.02	20.72
COL	12	0.5417607	0.54	21.26
CRI	69	3.1151242	3.12	24.38
DEU	83	3.7471783	3.75	28.13
DJI	11	0.4966140	0.50	28.62
DMA	1	0.0451467	0.05	28.67
DNK	13	0.5869074	0.59	29.26
ECOSOC	11	0.4966140	0.50	29.75
ECU	4	0.1805869	0.18	29.93

(continued)

applicant	N	exactpercent	roundedpercent	cumulpercent
ESP	13	0.5869074	0.59	30.52
ETH	30	1.3544018	1.35	31.87
FIN	7	0.3160271	0.32	32.19
FRA	32	1.4446953	1.44	33.63
GAB	2	0.0902935	0.09	33.72
GBR	77	3.4762980	3.48	37.20
GEO	16	0.7223476	0.72	37.92
GIN	24	1.0835214	1.08	39.01
GMB	11	0.4966140	0.50	39.50
GNB	16	0.7223476	0.72	40.23
GNQ	26	1.1738149	1.17	41.40
GRC	36	1.6252822	1.63	43.02
GTM	3	0.1354402	0.14	43.16
GUY	10	0.4514673	0.45	43.61
HND	7	0.3160271	0.32	43.93
HRV	31	1.3995485	1.40	45.33
HUN	16	0.7223476	0.72	46.05
IDN	14	0.6320542	0.63	46.68
IFAD	5	0.2257336	0.23	46.91
IMO	4	0.1805869	0.18	47.09
IND	25	1.1286682	1.13	48.22
IRN	65	2.9345372	2.93	51.15
ISR	8	0.3611738	0.36	51.51
ITA	7	0.3160271	0.32	51.83
KHM	30	1.3544018	1.35	53.18
LBR	30	1.3544018	1.35	54.54
LBY	84	3.7923251	3.79	58.33
LIE	19	0.8577878	0.86	59.19
MEX	19	0.8577878	0.86	60.05

(continued)

applicant	N	exactpercent	roundedpercent	cumulpercent
MHL	51	2.3024831	2.30	62.35
MKD	8	0.3611738	0.36	62.71
MYS	11	0.4966140	0.50	63.21
NIC	143	6.4559819	6.46	69.66
NLD	11	0.4966140	0.50	70.16
NRU	11	0.4966140	0.50	70.65
NZL	37	1.6704289	1.67	72.33
PAK	12	0.5417607	0.54	72.87
PER	13	0.5869074	0.59	73.45
PRT	35	1.5801354	1.58	75.03
PRY	7	0.3160271	0.32	75.35
PSE	1	0.0451467	0.05	75.40
QAT	52	2.3476298	2.35	77.74
ROU	4	0.1805869	0.18	77.92
SCG	163	7.3589165	7.36	85.28
SLV	24	1.0835214	1.08	86.37
SOM	17	0.7674944	0.77	87.13
TLS	11	0.4966140	0.50	87.63
TUN	19	0.8577878	0.86	88.49
UKR	32	1.4446953	1.44	89.93
UNESCO	9	0.4063205	0.41	90.34
UNGA	141	6.3656885	6.37	96.70
UNSC	16	0.7223476	0.72	97.43
USA	21	0.9480813	0.95	98.37
WHO	20	0.9029345	0.90	99.28
YUG	16	0.7223476	0.72	100.00
Total	2215	100.0000000	100.00	100.00

Frequency Table for Variable: respondent

74 unique value(s) detected.

respondent	N	exactpercent	roundedpercent	cumulpercent
NA	238	10.7449210	10.74	10.74
ALB	19	0.8577878	0.86	11.60
ARE	23	1.0383747	1.04	12.64
ARG	1	0.0451467	0.05	12.69
ARM	3	0.1354402	0.14	12.82
AUS	32	1.4446953	1.44	14.27
AZE	5	0.2257336	0.23	14.49
BDI	3	0.1354402	0.14	14.63
BEL	43	1.9413093	1.94	16.57
BGR	21	0.9480813	0.95	17.52
BHR	29	1.3092551	1.31	18.83
BLZ	3	0.1354402	0.14	18.96
BOL	4	0.1805869	0.18	19.14
BRA	1	0.0451467	0.05	19.19
CAN	34	1.5349887	1.53	20.72
CHE	4	0.1805869	0.18	20.90
CHL	28	1.2641084	1.26	22.17
COD	24	1.0835214	1.08	23.25
COL	77	3.4762980	3.48	26.73
CRI	19	0.8577878	0.86	27.58
CSK	1	0.0451467	0.05	27.63
DEU	27	1.2189616	1.22	28.85
DNK	22	0.9932280	0.99	29.84
EGY	1	0.0451467	0.05	29.89
ESP	52	2.3476298	2.35	32.23
FRA	135	6.0948081	6.09	38.33

(continued)

respondent	N	exactpercent	roundedpercent	cumulpercent
FRA-GBR-USA	7	0.3160271	0.32	38.65
GBR	107	4.8306998	4.83	43.48
GNQ	2	0.0902935	0.09	43.57
GRC	8	0.3611738	0.36	43.93
GTM	11	0.4966140	0.50	44.42
HND	43	1.9413093	1.94	46.37
HUN	1	0.0451467	0.05	46.41
IND	54	2.4379233	2.44	48.85
IRN	19	0.8577878	0.86	49.71
ISL	49	2.2121896	2.21	51.92
ITA	43	1.9413093	1.94	53.86
JPN	17	0.7674944	0.77	54.63
KEN	17	0.7674944	0.77	55.40
LBN	7	0.3160271	0.32	55.71
LBY	19	0.8577878	0.86	56.57
MLI	8	0.3611738	0.36	56.93
MLT	22	0.9932280	0.99	57.92
MMR	11	0.4966140	0.50	58.42
MYS	14	0.6320542	0.63	59.05
NAM	12	0.5417607	0.54	59.59
NER	15	0.6772009	0.68	60.27
NGA	38	1.7155756	1.72	61.99
NIC	75	3.3860045	3.39	65.37
NLD	44	1.9864560	1.99	67.36
NOR	33	1.4898420	1.49	68.85
PAK	42	1.8961625	1.90	70.74
PER	12	0.5417607	0.54	71.29
PRT	21	0.9480813	0.95	72.23
QAT	12	0.5417607	0.54	72.78

(continued)

respondent	N	exactpercent	roundedpercent	cumulpercent
RUS	48	2.1670429	2.17	74.94
RWA	19	0.8577878	0.86	75.80
SCG	43	1.9413093	1.94	77.74
SEN	33	1.4898420	1.49	79.23
SGP	11	0.4966140	0.50	79.73
SRB	31	1.3995485	1.40	81.13
SUN	4	0.1805869	0.18	81.31
SVK	16	0.7223476	0.72	82.03
SWE	11	0.4966140	0.50	82.53
TCD	8	0.3611738	0.36	82.89
THA	30	1.3544018	1.35	84.24
TUR	20	0.9029345	0.90	85.15
UGA	34	1.5349887	1.53	86.68
UKR	4	0.1805869	0.18	86.86
URY	20	0.9029345	0.90	87.77
USA	195	8.8036117	8.80	96.57
VEN	10	0.4514673	0.45	97.02
YUG	6	0.2708804	0.27	97.29
ZAF	60	2.7088036	2.71	100.00
Total	2215	100.0000000	100.00	100.00

Frequency Table for Variable: doctype

3 unique value(s) detected.

doctype	N	exactpercent	roundedpercent	cumulpercent
ADV	194	8.758465	8.76	8.76
JUD	1062	47.945824	47.95	56.70
ORD	959	43.295711	43.30	100.00

(continued)

doctype	N	exactpercent	roundedpercent	cumulpercent
Total	2215	100.000000	100.00	100.00

Frequency Table for Variable: collision

3 unique value(s) detected.

collision	N	exactpercent	roundedpercent	cumulpercent
1	2198	99.2325056	99.23	99.23
2	16	0.7223476	0.72	99.95
3	1	0.0451467	0.05	100.00
Total	2215	100.0000000	100.00	100.00

Frequency Table for Variable: stage

5 unique value(s) detected.

stage	N	exactpercent	roundedpercent	cumulpercent
NA	1179	53.227991	53.23	53.23
CO	33	1.489842	1.49	54.72
IN	36	1.625282	1.63	56.34
ME	543	24.514673	24.51	80.86
PO	424	19.142212	19.14	100.00
Total	2215	100.000000	100.00	100.00

Frequency Table for Variable: opinion

15 unique value(s) detected.

opinion	N	exactpercent	roundedpercent	cumulpercent
0	782	35.3047404	35.30	35.30
1	260	11.7381490	11.74	47.04
2	233	10.5191874	10.52	57.56
3	203	9.1647856	9.16	66.73
4	173	7.8103837	7.81	74.54
5	150	6.7720090	6.77	81.31
6	127	5.7336343	5.73	87.04
7	96	4.3340858	4.33	91.38
8	72	3.2505643	3.25	94.63
9	58	2.6185102	2.62	97.25
10	31	1.3995485	1.40	98.65
11	14	0.6320542	0.63	99.28
12	8	0.3611738	0.36	99.64
13	4	0.1805869	0.18	99.82
14	4	0.1805869	0.18	100.00
Total	2215	100.0000000	100.00	100.00

Frequency Table for Variable: language

1 unique value(s) detected.

language	N	exactpercent	roundedpercent	cumulpercent
EN	2215	100	100	100
Total	2215	100	100	100

Frequency Table for Variable: year

76 unique value(s) detected.

year	N	exactpercent	roundedpercent	cumulpercent
1947	3	0.1354402	0.14	0.14
1948	12	0.5417607	0.54	0.68
1949	24	1.0835214	1.08	1.76
1950	31	1.3995485	1.40	3.16
1951	21	0.9480813	0.95	4.11
1952	26	1.1738149	1.17	5.28
1953	11	0.4966140	0.50	5.78
1954	19	0.8577878	0.86	6.64
1955	11	0.4966140	0.50	7.13
1956	22	0.9932280	0.99	8.13
1957	22	0.9932280	0.99	9.12
1958	28	1.2641084	1.26	10.38
1959	32	1.4446953	1.44	11.83
1960	26	1.1738149	1.17	13.00
1961	17	0.7674944	0.77	13.77
1962	42	1.8961625	1.90	15.67
1963	17	0.7674944	0.77	16.43
1964	15	0.6772009	0.68	17.11
1965	5	0.2257336	0.23	17.34
1966	24	1.0835214	1.08	18.42
1967	4	0.1805869	0.18	18.60
1968	5	0.2257336	0.23	18.83
1969	24	1.0835214	1.08	19.91
1970	14	0.6320542	0.63	20.54
1971	16	0.7223476	0.72	21.26
1972	23	1.0383747	1.04	22.30
1973	62	2.7990971	2.80	25.10
1974	52	2.3476298	2.35	27.45
1975	15	0.6772009	0.68	28.13
1976	11	0.4966140	0.50	28.62

(continued)

year	N	exactpercent	roundedpercent	cumulpercent
1977	1	0.0451467	0.05	28.67
1978	8	0.3611738	0.36	29.03
1979	3	0.1354402	0.14	29.16
1980	16	0.7223476	0.72	29.89
1981	8	0.3611738	0.36	30.25
1982	24	1.0835214	1.08	31.33
1983	2	0.0902935	0.09	31.42
1984	35	1.5801354	1.58	33.00
1985	18	0.8126411	0.81	33.81
1986	17	0.7674944	0.77	34.58
1987	17	0.7674944	0.77	35.35
1988	14	0.6320542	0.63	35.98
1989	21	0.9480813	0.95	36.93
1990	15	0.6772009	0.68	37.61
1991	23	1.0383747	1.04	38.65
1992	43	1.9413093	1.94	40.59
1993	29	1.3092551	1.31	41.90
1994	14	0.6320542	0.63	42.53
1995	29	1.3092551	1.31	43.84
1996	55	2.4830700	2.48	46.32
1997	19	0.8577878	0.86	47.18
1998	61	2.7539503	2.75	49.93
1999	135	6.0948081	6.09	56.03
2000	37	1.6704289	1.67	57.70
2001	44	1.9864560	1.99	59.68
2002	49	2.2121896	2.21	61.90
2003	35	1.5801354	1.58	63.48
2004	78	3.5214447	3.52	67.00
2005	21	0.9480813	0.95	67.95

(continued)

year	N	exactpercent	roundedpercent	cumulpercent
2006	18	0.8126411	0.81	68.76
2007	39	1.7607223	1.76	70.52
2008	42	1.8961625	1.90	72.42
2009	20	0.9029345	0.90	73.32
2010	42	1.8961625	1.90	75.21
2011	57	2.5733634	2.57	77.79
2012	36	1.6252822	1.63	79.41
2013	33	1.4898420	1.49	80.90
2014	40	1.8058691	1.81	82.71
2015	52	2.3476298	2.35	85.06
2016	75	3.3860045	3.39	88.44
2017	35	1.5801354	1.58	90.02
2018	64	2.8893905	2.89	92.91
2019	54	2.4379233	2.44	95.35
2020	35	1.5801354	1.58	96.93
2021	31	1.3995485	1.40	98.33
2022	37	1.6704289	1.67	100.00
Total	2215	100.0000000	100.00	100.00

Frequency Table for Variable: minority

2 unique value(s) detected.

minority	N	exactpercent	roundedpercent	cumulpercent
0	782	35.30474	35.3	35.3
1	1433	64.69526	64.7	100.0
Total	2215	100.00000	100.0	100.0

Frequency Table for Variable: fullname

182 unique value(s) detected.

fullname	N	exactpercent	roundedpercent	cumulpercent
Accordance with international law of the unilateral declaration of independence in respect of Kosovo	11	0.4966140	0.50	0.50
Admissibility of Hearings of Petitioners by the Committee on South West Africa	6	0.2708804	0.27	0.77
Aegean Sea Continental Shelf (Greece v. Turkey)	20	0.9029345	0.90	1.67
Aerial Herbicide Spraying (Ecuador v. Colombia)	4	0.1805869	0.18	1.85
Aerial Incident of 10 August 1999 (Pakistan v. India)	7	0.3160271	0.32	2.17
Aerial Incident of 10 March 1953 (United States of America v. Czechoslovakia)	1	0.0451467	0.05	2.21
Aerial Incident of 27 July 1955 (Israel v. Bulgaria)	8	0.3611738	0.36	2.57
Aerial Incident of 27 July 1955 (United Kingdom v. Bulgaria)	5	0.2257336	0.23	2.80
Aerial Incident of 27 July 1955 (United States of America v. Bulgaria)	8	0.3611738	0.36	3.16
Aerial Incident of 3 July 1988 (Islamic Republic of Iran v. United States of America)	9	0.4063205	0.41	3.57
Aerial Incident of 4 September 1954 (United States of America v. Union of Soviet Socialist Republics)	1	0.0451467	0.05	3.61
Aerial Incident of 7 November 1954 (United States of America v. Union of Soviet Socialist Republics)	1	0.0451467	0.05	3.66

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Aerial Incident of 7 October 1952 (United States of America v. Union of Soviet Socialist Republics)	1	0.0451467	0.05	3.70
Ahmadou Sadio Diallo (Republic of Guinea v. Democratic Republic of the Congo)	24	1.0835214	1.08	4.79
Allegations of Genocide under the Convention on the Prevention and Punishment of the Crime of Genocide (Ukraine v. Russian Federation)	8	0.3611738	0.36	5.15
Alleged Violations of Sovereign Rights and Maritime Spaces in the Caribbean Sea (Nicaragua v. Colombia)	25	1.1286682	1.13	6.28
Alleged violations of the 1955 Treaty of Amity, Economic Relations, and Consular Rights (Islamic Republic of Iran v. United States of America)	12	0.5417607	0.54	6.82
Ambatielos (Greece v. United Kingdom)	16	0.7223476	0.72	7.54
Anglo-Iranian Oil Co. (United Kingdom v. Iran)	12	0.5417607	0.54	8.08
Antarctica (United Kingdom v. Argentina)	1	0.0451467	0.05	8.13
Antarctica (United Kingdom v. Chile)	1	0.0451467	0.05	8.17
Appeal Relating to the Jurisdiction of the ICAO Council (India v. Pakistan)	14	0.6320542	0.63	8.80
Appeal Relating to the Jurisdiction of the ICAO Council under Article 84 of the Convention on International Civil Aviation (Bahrain, Egypt, Saudi Arabia and United Arab Emirates v. Qatar)	6	0.2708804	0.27	9.07

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Appeal Relating to the Jurisdiction of the ICAO Council under Article II, Section 2, of the 1944 International Air Services Transit Agreement (Bahrain, Egypt and United Arab Emirates v. Qatar)	6	0.2708804	0.27	9.35
Applicability of Article VI, Section 22, of the Convention on the Privileges and Immunities of the United Nations	5	0.2257336	0.23	9.57
Applicability of the Obligation to Arbitrate under Section 21 of the United Nations Headquarters Agreement of 26 June 1947	6	0.2708804	0.27	9.84
Application for Review of Judgment No. 158 of the United Nations Administrative Tribunal	11	0.4966140	0.50	10.34
Application for Review of Judgment No. 273 of the United Nations Administrative Tribunal	11	0.4966140	0.50	10.84
Application for Review of Judgment No. 333 of the United Nations Administrative Tribunal	10	0.4514673	0.45	11.29
Application for Revision and Interpretation of the Judgment of 24 February 1982 in the Case concerning the Continental Shelf (Tunisia/Libyan Arab Jamahiriya) (Tunisia v. Libyan Arab Jamahiriya)	5	0.2257336	0.23	11.51

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Application for Revision of the Judgment of 11 July 1996 in the Case concerning Application of the Convention on the Prevention and Punishment of the Crime of Genocide (Bosnia and Herzegovina v. Yugoslavia), Preliminary Objections (Yugoslavia v. Bosnia and Herzegovina)	6	0.2708804	0.27	11.78
Application for Revision of the Judgment of 11 September 1992 in the Case concerning the Land, Island and Maritime Frontier Dispute (El Salvador/Honduras: Nicaragua intervening) (El Salvador v. Honduras)	4	0.1805869	0.18	11.96
Application for revision of the Judgment of 23 May 2008 in the case concerning Sovereignty over Pedra Branca/Pulau Batu Puteh, Middle Rocks and South Ledge (Malaysia/Singapore) (Malaysia v. Singapore)	1	0.0451467	0.05	12.01
Application of the Convention of 1902 Governing the Guardianship of Infants (Netherlands v. Sweden)	11	0.4966140	0.50	12.51
Application of the Convention on the Prevention and Punishment of the Crime of Genocide (Bosnia and Herzegovina v. Serbia and Montenegro)	43	1.9413093	1.94	14.45
Application of the Convention on the Prevention and Punishment of the Crime of Genocide (Croatia v. Serbia)	31	1.3995485	1.40	15.85

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Application of the Convention on the Prevention and Punishment of the Crime of Genocide (The Gambia v. Myanmar)	11	0.4966140	0.50	16.34
Application of the Interim Accord of 13 September 1995 (the former Yugoslav Republic of Macedonia v. Greece)	8	0.3611738	0.36	16.70
Application of the International Convention for the Suppression of the Financing of Terrorism and of the International Convention on the Elimination of All Forms of Racial Discrimination (Ukraine v. Russian Federation)	24	1.0835214	1.08	17.79
Application of the International Convention on the Elimination of All Forms of Racial Discrimination (Armenia v. Azerbaijan)	5	0.2257336	0.23	18.01
Application of the International Convention on the Elimination of All Forms of Racial Discrimination (Azerbaijan v. Armenia)	3	0.1354402	0.14	18.15
Application of the International Convention on the Elimination of All Forms of Racial Discrimination (Georgia v. Russian Federation)	16	0.7223476	0.72	18.87
Application of the International Convention on the Elimination of All Forms of Racial Discrimination (Qatar v. United Arab Emirates)	23	1.0383747	1.04	19.91
Arbitral Award Made by the King of Spain on 23 December 1906 (Honduras v. Nicaragua)	6	0.2708804	0.27	20.18
Arbitral Award of 3 October 1899 (Guyana v. Venezuela)	10	0.4514673	0.45	20.63

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Arbitral Award of 31 July 1989 (Guinea-Bissau v. Senegal)	15	0.6772009	0.68	21.31
Armed Activities on the Territory of the Congo (Democratic Republic of the Congo v. Burundi)	3	0.1354402	0.14	21.44
Armed Activities on the Territory of the Congo (Democratic Republic of the Congo v. Rwanda)	3	0.1354402	0.14	21.58
Armed Activities on the Territory of the Congo (Democratic Republic of the Congo v. Uganda)	34	1.5349887	1.53	23.12
Armed Activities on the Territory of the Congo (New Application: 2002) (Democratic Republic of the Congo v. Rwanda)	16	0.7223476	0.72	23.84
Arrest Warrant of 11 April 2000 (Democratic Republic of the Congo v. Belgium)	22	0.9932280	0.99	24.83
Asylum (Colombia v. Peru)	9	0.4063205	0.41	25.24
Avena and Other Mexican Nationals (Mexico v. United States of America)	11	0.4966140	0.50	25.73
Barcelona Traction, Light and Power Company, Limited (Belgium v. Spain)	5	0.2257336	0.23	25.96
Barcelona Traction, Light and Power Company, Limited (Belgium v. Spain) (New Application: 1962)	31	1.3995485	1.40	27.36
Border and Transborder Armed Actions (Nicaragua v. Costa Rica)	3	0.1354402	0.14	27.49
Border and Transborder Armed Actions (Nicaragua v. Honduras)	11	0.4966140	0.50	27.99

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Certain Activities Carried Out by Nicaragua in the Border Area (Costa Rica v. Nicaragua)	40	1.8058691	1.81	29.80
Certain Criminal Proceedings in France (Republic of the Congo v. France)	11	0.4966140	0.50	30.29
Certain Expenses of the United Nations (Article 17, paragraph 2, of the Charter)	11	0.4966140	0.50	30.79
Certain Iranian Assets (Islamic Republic of Iran v. United States of America)	12	0.5417607	0.54	31.33
Certain Norwegian Loans (France v. Norway)	11	0.4966140	0.50	31.83
Certain Phosphate Lands in Nauru (Nauru v. Australia)	11	0.4966140	0.50	32.33
Certain Property (Liechtenstein v. Germany)	8	0.3611738	0.36	32.69
Certain Questions concerning Diplomatic Relations (Honduras v. Brazil)	1	0.0451467	0.05	32.73
Certain Questions of Mutual Assistance in Criminal Matters (Djibouti v. France)	11	0.4966140	0.50	33.23
Compagnie du Port, des Quais et des Entrepôts de Beyrouth and Société Radio-Orient (France v. Lebanon)	4	0.1805869	0.18	33.41
Competence of the General Assembly for the Admission of a State to the United Nations	4	0.1805869	0.18	33.59
Conditions of Admission of a State to Membership in the United Nations (Article 4 of the Charter)	7	0.3160271	0.32	33.91

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Constitution of the Maritime Safety Committee of the Inter-Governmental Maritime Consultative Organization	4	0.1805869	0.18	34.09
Construction of a Road in Costa Rica along the San Juan River (Nicaragua v. Costa Rica)	16	0.7223476	0.72	34.81
Continental Shelf (Libyan Arab Jamahiriya/Malta)	22	0.9932280	0.99	35.80
Continental Shelf (Tunisia/Libyan Arab Jamahiriya)	14	0.6320542	0.63	36.43
Corfu Channel (United Kingdom of Great Britain and Northern Ireland v. Albania)	19	0.8577878	0.86	37.29
Delimitation of the Maritime Boundary in the Gulf of Maine Area (Canada/United States of America)	12	0.5417607	0.54	37.83
Difference Relating to Immunity from Legal Process of a Special Rapporteur of the Commission on Human Rights	6	0.2708804	0.27	38.10
Dispute over the Status and Use of the Waters of the Silala (Chile v. Bolivia)	4	0.1805869	0.18	38.28
Dispute regarding Navigational and Related Rights (Costa Rica v. Nicaragua)	6	0.2708804	0.27	38.56
East Timor (Portugal v. Australia)	10	0.4514673	0.45	39.01
Effect of Awards of Compensation Made by the United Nations Administrative Tribunal	6	0.2708804	0.27	39.28
Electricité de Beyrouth Company (France v. Lebanon)	3	0.1354402	0.14	39.41

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Elettronica Sicula S.p.A. (ELSI) (United States of America v. Italy)	6	0.2708804	0.27	39.68
Fisheries (United Kingdom v. Norway)	9	0.4063205	0.41	40.09
Fisheries Jurisdiction (Federal Republic of Germany v. Iceland)	25	1.1286682	1.13	41.22
Fisheries Jurisdiction (Spain v. Canada)	13	0.5869074	0.59	41.81
Fisheries Jurisdiction (United Kingdom v. Iceland)	24	1.0835214	1.08	42.89
Frontier Dispute (Benin/Niger)	7	0.3160271	0.32	43.21
Frontier Dispute (Burkina Faso/Niger)	8	0.3611738	0.36	43.57
Frontier Dispute (Burkina Faso/Republic of Mali)	8	0.3611738	0.36	43.93
Gabčíkovo-Nagymaros Project (Hungary/Slovakia)	16	0.7223476	0.72	44.65
Guatemala's Territorial, Insular and Maritime Claim (Guatemala/Belize)	3	0.1354402	0.14	44.79
Haya de la Torre (Colombia v. Peru)	2	0.0902935	0.09	44.88
Immunities and Criminal Proceedings (Equatorial Guinea v. France)	26	1.1738149	1.17	46.05
Interhandel (Switzerland v. United States of America)	16	0.7223476	0.72	46.77
International Status of South West Africa	7	0.3160271	0.32	47.09
Interpretation of Peace Treaties with Bulgaria, Hungary and Romania	10	0.4514673	0.45	47.54

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Interpretation of the Agreement of 25 March 1951 between the WHO and Egypt	11	0.4966140	0.50	48.04
Jadhav (India v. Pakistan)	11	0.4966140	0.50	48.53
Judgment No.2867 of the Administrative Tribunal of the International Labour Organization upon a Complaint Filed against the International Fund for Agricultural Development	5	0.2257336	0.23	48.76
Judgments of the Administrative Tribunal of the ILO upon Complaints Made against UNESCO	9	0.4063205	0.41	49.16
Jurisdiction and Enforcement of Judgments in Civil and Commercial Matters (Belgium v. Switzerland)	3	0.1354402	0.14	49.30
Jurisdictional Immunities of the State (Germany v. Italy: Greece intervening)	15	0.6772009	0.68	49.98
Kasikili/Sedudu Island (Botswana/Namibia)	12	0.5417607	0.54	50.52
LaGrand (Germany v. United States of America)	11	0.4966140	0.50	51.02
Land Boundary in the Northern Part of Isla Portillos (Costa Rica v. Nicaragua)	10	0.4514673	0.45	51.47
Land and Maritime Boundary between Cameroon and Nigeria (Cameroon v. Nigeria: Equatorial Guinea intervening)	34	1.5349887	1.53	53.00
Land and Maritime Delimitation and Sovereignty over Islands (Gabon/Equatorial Guinea)	2	0.0902935	0.09	53.09

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Land, Island and Maritime Frontier Dispute (El Salvador/Honduras: Nicaragua intervening)	20	0.9029345	0.90	54.00
Legal Consequences for States of the Continued Presence of South Africa in Namibia (South West Africa) notwithstanding Security Council Resolution 276 (1970)	16	0.7223476	0.72	54.72
Legal Consequences of the Construction of a Wall in the Occupied Palestinian Territory	11	0.4966140	0.50	55.21
Legal Consequences of the Separation of the Chagos Archipelago from Mauritius in 1965	15	0.6772009	0.68	55.89
Legality of Use of Force (Serbia and Montenegro v. Belgium)	21	0.9480813	0.95	56.84
Legality of Use of Force (Serbia and Montenegro v. Canada)	21	0.9480813	0.95	57.79
Legality of Use of Force (Serbia and Montenegro v. France)	19	0.8577878	0.86	58.65
Legality of Use of Force (Serbia and Montenegro v. Germany)	19	0.8577878	0.86	59.50
Legality of Use of Force (Serbia and Montenegro v. Italy)	20	0.9029345	0.90	60.41
Legality of Use of Force (Serbia and Montenegro v. Netherlands)	21	0.9480813	0.95	61.35
Legality of Use of Force (Serbia and Montenegro v. Portugal)	21	0.9480813	0.95	62.30

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Legality of Use of Force (Serbia and Montenegro v. United Kingdom)	21	0.9480813	0.95	63.25
Legality of Use of Force (Yugoslavia v. Spain)	9	0.4063205	0.41	63.66
Legality of Use of Force (Yugoslavia v. United States of America)	7	0.3160271	0.32	63.97
Legality of the Threat or Use of Nuclear Weapons	16	0.7223476	0.72	64.70
Legality of the Use by a State of Nuclear Weapons in Armed Conflict	9	0.4063205	0.41	65.10
Maritime Delimitation and Territorial Questions between Qatar and Bahrain (Qatar v. Bahrain)	29	1.3092551	1.31	66.41
Maritime Delimitation between Guinea-Bissau and Senegal (Guinea-Bissau v. Senegal)	1	0.0451467	0.05	66.46
Maritime Delimitation in the Area between Greenland and Jan Mayen (Denmark v. Norway)	13	0.5869074	0.59	67.04
Maritime Delimitation in the Black Sea (Romania v. Ukraine)	4	0.1805869	0.18	67.22
Maritime Delimitation in the Caribbean Sea and the Pacific Ocean (Costa Rica v. Nicaragua)	13	0.5869074	0.59	67.81
Maritime Delimitation in the Indian Ocean (Somalia v. Kenya)	17	0.7674944	0.77	68.58
Maritime Dispute (Peru v. Chile)	13	0.5869074	0.59	69.16

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Military and Paramilitary Activities in and against Nicaragua (Nicaragua v. United States of America)	32	1.4446953	1.44	70.61
Minquiers and Ecrehos (France/United Kingdom)	7	0.3160271	0.32	70.93
Monetary Gold Removed from Rome in 1943 (Italy v. France, United Kingdom of Great Britain and Northern Ireland and United States of America)	7	0.3160271	0.32	71.24
North Sea Continental Shelf (Federal Republic of Germany/Denmark)	15	0.6772009	0.68	71.92
North Sea Continental Shelf (Federal Republic of Germany/Netherlands)	15	0.6772009	0.68	72.60
Northern Cameroons (Cameroon v. United Kingdom)	18	0.8126411	0.81	73.41
Nottebohm (Liechtenstein v. Guatemala)	11	0.4966140	0.50	73.91
Nuclear Tests (Australia v. France)	31	1.3995485	1.40	75.30
Nuclear Tests (New Zealand v. France)	29	1.3092551	1.31	76.61
Obligation to Negotiate Access to the Pacific Ocean (Bolivia v. Chile)	14	0.6320542	0.63	77.25
Obligations concerning Negotiations relating to Cessation of the Nuclear Arms Race and to Nuclear Disarmament (Marshall Islands v. India)	17	0.7674944	0.77	78.01
Obligations concerning Negotiations relating to Cessation of the Nuclear Arms Race and to Nuclear Disarmament (Marshall Islands v. Pakistan)	17	0.7674944	0.77	78.78

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Obligations concerning Negotiations relating to Cessation of the Nuclear Arms Race and to Nuclear Disarmament (Marshall Islands v. United Kingdom)	17	0.7674944	0.77	79.55
Oil Platforms (Islamic Republic of Iran v. United States of America)	32	1.4446953	1.44	80.99
Passage through the Great Belt (Finland v. Denmark)	7	0.3160271	0.32	81.31
Protection of French Nationals and Protected Persons in Egypt (France v. Egypt)	1	0.0451467	0.05	81.35
Pulp Mills on the River Uruguay (Argentina v. Uruguay)	20	0.9029345	0.90	82.26
Question of the Delimitation of the Continental Shelf between Nicaragua and Colombia beyond 200 nautical miles from the Nicaraguan Coast (Nicaragua v. Colombia)	13	0.5869074	0.59	82.84
Questions of Interpretation and Application of the 1971 Montreal Convention arising from the Aerial Incident at Lockerbie (Libyan Arab Jamahiriya v. United Kingdom)	28	1.2641084	1.26	84.11
Questions of Interpretation and Application of the 1971 Montreal Convention arising from the Aerial Incident at Lockerbie (Libyan Arab Jamahiriya v. United States of America)	26	1.1738149	1.17	85.28

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Questions of jurisdictional immunities of the State and measures of constraint against State-owned property (Germany v. Italy)	2	0.0902935	0.09	85.37
Questions relating to the Obligation to Prosecute or Extradite (Belgium v. Senegal)	17	0.7674944	0.77	86.14
Questions relating to the Seizure and Detention of Certain Documents and Data (Timor-Leste v. Australia)	11	0.4966140	0.50	86.64
Relocation of the United States Embassy to Jerusalem (Palestine v. United States of America)	1	0.0451467	0.05	86.68
Reparation for Injuries Suffered in the Service of the United Nations	7	0.3160271	0.32	87.00
Request for Interpretation of the Judgment of 11 June 1998 in the Case concerning the Land and Maritime Boundary between Cameroon and Nigeria (Cameroon v. Nigeria), Preliminary Objections (Nigeria v. Cameroon)	4	0.1805869	0.18	87.18
Request for Interpretation of the Judgment of 15 June 1962 in the Case concerning the Temple of Preah Vihear (Cambodia v. Thailand) (Cambodia v. Thailand)	14	0.6320542	0.63	87.81
Request for Interpretation of the Judgment of 20 November 1950 in the Asylum Case (Colombia v. Peru)	1	0.0451467	0.05	87.86

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Request for Interpretation of the Judgment of 23 May 2008 in the case concerning Sovereignty over Pedra Branca/Pulau Batu Puteh, Middle Rocks and South Ledge (Malaysia/Singapore) (Malaysia v. Singapore)	1	0.0451467	0.05	87.90
Request for Interpretation of the Judgment of 31 March 2004 in the Case concerning Avena and Other Mexican Nationals (Mexico v. United States of America) (Mexico v. United States of America)	8	0.3611738	0.36	88.26
Request for an Examination of the Situation in Accordance with Paragraph 63 of the Court's Judgment of 20 December 1974 in the Nuclear Tests (New Zealand v. France) Case	8	0.3611738	0.36	88.62
Reservations to the Convention on the Prevention and Punishment of the Crime of Genocide	4	0.1805869	0.18	88.80
Right of Passage over Indian Territory (Portugal v. India)	25	1.1286682	1.13	89.93
Rights of Nationals of the United States of America in Morocco (France v. United States of America)	6	0.2708804	0.27	90.20
South West Africa (Ethiopia v. South Africa)	30	1.3544018	1.35	91.56
South West Africa (Liberia v. South Africa)	30	1.3544018	1.35	92.91
Sovereignty over Certain Frontier Land (Belgium/Netherlands)	8	0.3611738	0.36	93.27

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Sovereignty over Pedra Branca/Pulau Batu Puteh, Middle Rocks and South Ledge (Malaysia/Singapore)	9	0.4063205	0.41	93.68
Sovereignty over Pulau Ligitan and Pulau Sipadan (Indonesia/Malaysia)	14	0.6320542	0.63	94.31
Status vis-à-vis the Host State of a Diplomatic Envoy to the United Nations (Commonwealth of Dominica v. Switzerland)	1	0.0451467	0.05	94.36
Temple of Preah Vihear (Cambodia v. Thailand)	16	0.7223476	0.72	95.08
Territorial Dispute (Libyan Arab Jamahiriya/Chad)	8	0.3611738	0.36	95.44
Territorial and Maritime Dispute (Nicaragua v. Colombia)	35	1.5801354	1.58	97.02
Territorial and Maritime Dispute between Nicaragua and Honduras in the Caribbean Sea (Nicaragua v. Honduras)	8	0.3611738	0.36	97.38
Treatment in Hungary of Aircraft and Crew of United States of America (United States of America v. Hungarian People's Republic)	1	0.0451467	0.05	97.43
Treatment in Hungary of Aircraft and Crew of United States of America (United States of America v. Union of Soviet Socialist Republics)	1	0.0451467	0.05	97.47
Trial of Pakistani Prisoners of War (Pakistan v. India)	5	0.2257336	0.23	97.70
United States Diplomatic and Consular Staff in Tehran (United States of America v. Iran)	7	0.3160271	0.32	98.01

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Vienna Convention on Consular Relations (Paraguay v. United States of America)	7	0.3160271	0.32	98.33
Voting Procedure on Questions relating to Reports and Petitions concerning the Territory of South West Africa	5	0.2257336	0.23	98.56
Western Sahara	15	0.6772009	0.68	99.23
Whaling in the Antarctic (Australia v. Japan: New Zealand intervening)	17	0.7674944	0.77	100.00
Total	2215	100.0000000	100.00	100.00

Frequency Table for Variable: applicant_region

8 unique value(s) detected.

applicant_region	N	exactpercent	roundedpercent	cumulpercent
Africa	450	20.3160271	20.32	20.32
Americas	383	17.2911964	17.29	37.61
Asia	253	11.4221219	11.42	49.03
Asia|Africa|Asia	6	0.2708804	0.27	49.30
Asia|Africa|Asia|Asia	6	0.2708804	0.27	49.57
Europe	732	33.0474041	33.05	82.62
NA	238	10.7449210	10.74	93.36
Oceania	147	6.6365688	6.64	100.00
Total	2215	100.0000000	100.00	100.00

Frequency Table for Variable: respondent_region

7 unique value(s) detected.

respondent_region	N	exactpercent	roundedpercent	cumulpercent
NA	238	10.7449210	10.74	10.74
Africa	293	13.2279910	13.23	23.97
Americas	533	24.0632054	24.06	48.04
Asia	297	13.4085779	13.41	61.44
Europe	815	36.7945824	36.79	98.24
Europe|Europe|Americas	7	0.3160271	0.32	98.56
Oceania	32	1.4446953	1.44	100.00
Total	2215	100.0000000	100.00	100.00

Frequency Table for Variable: applicant_subregion

16 unique value(s) detected.

applicant_subregion	N	exactpercent	roundedpercent	cumulpercent
Australia and New Zealand	85	3.8374718	3.84	3.84
Eastern Europe	52	2.3476298	2.35	6.19
Latin America and the Caribbean	350	15.8013544	15.80	21.99
Micronesia	62	2.7990971	2.80	24.79
NA	238	10.7449210	10.74	35.53
Northern Africa	103	4.6501129	4.65	40.18
Northern America	33	1.4898420	1.49	41.67
Northern Europe	97	4.3792325	4.38	46.05
South-eastern Asia	66	2.9796840	2.98	49.03
Southern Asia	102	4.6049661	4.60	53.63
Southern Europe	358	16.1625282	16.16	69.80
Sub-Saharan Africa	347	15.6659142	15.67	85.46
Western Asia	85	3.8374718	3.84	89.30
Western Asia|Northern Africa|Western Asia	6	0.2708804	0.27	89.57

(continued)

applicant_subregion	N	exactpercent	roundedpercent	cumulpercent
Western Asia|Northern Africa|Western Asia|Western Asia	6	0.2708804	0.27	89.84
Western Europe	225	10.1580135	10.16	100.00
Total	2215	100.0000000	100.00	100.00

Frequency Table for Variable: respondent_subregion

15 unique value(s) detected.

respondent_subregion	N	exactpercent	roundedpercent	cumulpercent
NA	238	10.7449210	10.74	10.74
Australia and New Zealand	32	1.4446953	1.44	12.19
Eastern Asia	17	0.7674944	0.77	12.96
Eastern Europe	95	4.2889391	4.29	17.25
Latin America and the Caribbean	304	13.7246050	13.72	30.97
Northern Africa	20	0.9029345	0.90	31.87
Northern America	229	10.3386005	10.34	42.21
Northern Europe	222	10.0225734	10.02	52.23
South-eastern Asia	66	2.9796840	2.98	55.21
Southern Asia	115	5.1918736	5.19	60.41
Southern Europe	245	11.0609481	11.06	71.47
Sub-Saharan Africa	273	12.3250564	12.33	83.79
Western Asia	99	4.4695260	4.47	88.26
Western Europe	253	11.4221219	11.42	99.68
Western Europe|Northern Europe|Northern America	7	0.3160271	0.32	100.00
Total	2215	100.0000000	100.00	100.00

Frequency Table for Variable: doi_concept

1 unique value(s) detected.

doi_concept	N	exactpercent	roundedpercent	cumulpercent
10.5281/zenodo.3826444	2215	100	100	100
Total	2215	100	100	100

Frequency Table for Variable: doi_version

1 unique value(s) detected.

doi_version	N	exactpercent	roundedpercent	cumulpercent
10.5281/zenodo.7051929	2215	100	100	100
Total	2215	100	100	100

Frequency Table for Variable: version

1 unique value(s) detected.

version	N	exactpercent	roundedpercent	cumulpercent
2022-09-07	2215	100	100	100
Total	2215	100	100	100

Frequency Table for Variable: license

1 unique value(s) detected.

license	N	exactpercent	roundedpercent	cumulpercent
Creative Commons Zero 1.0 Universal	2215	100	100	100
Total	2215	100	100	100

(continued)

license	N	exactpercent	roundedpercent	cumulpercent

22.3 French Corpus

22.3.1 Variables to Ignore

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

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

22.3.2 Variables to Analyze

```
varlist <- names(data.best.fr)

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

print(varlist)
```

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

22.3.3 Construct Frequency Tables

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

```
f.fast.freqtable(data.best.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
ICJ	2197	100	100	100
Total	2197	100	100	100

Frequency Table for Variable: caseno

182 unique value(s) detected.

caseno	N	exactpercent	roundedpercent	cumulpercent
1	19	0.8648157	0.86	0.86
3	7	0.3186163	0.32	1.18
4	7	0.3186163	0.32	1.50
5	9	0.4096495	0.41	1.91
6	1	0.0455166	0.05	1.96
7	9	0.4096495	0.41	2.37
8	10	0.4551661	0.46	2.82
9	4	0.1820665	0.18	3.00
10	7	0.3186163	0.32	3.32
11	6	0.2730997	0.27	3.60
12	4	0.1820665	0.18	3.78
13	1	0.0455166	0.05	3.82
14	2	0.0910332	0.09	3.91
15	16	0.7282658	0.73	4.64
16	12	0.5461994	0.55	5.19
17	7	0.3186163	0.32	5.51
18	11	0.5006827	0.50	6.01
19	7	0.3186163	0.32	6.33
20	3	0.1365498	0.14	6.46

(continued)

caseno	N	exactpercent	roundedpercent	cumulpercent
21	6	0.2730997	0.27	6.74
22	1	0.0455166	0.05	6.78
23	1	0.0455166	0.05	6.83
24	5	0.2275831	0.23	7.06
25	1	0.0455166	0.05	7.10
26	1	0.0455166	0.05	7.15
27	1	0.0455166	0.05	7.19
28	1	0.0455166	0.05	7.24
29	11	0.5006827	0.50	7.74
30	9	0.4096495	0.41	8.15
31	6	0.2730997	0.27	8.42
32	25	1.1379153	1.14	9.56
33	11	0.5006827	0.50	10.06
34	16	0.7282658	0.73	10.79
35	8	0.3641329	0.36	11.15
36	8	0.3641329	0.36	11.52
37	5	0.2275831	0.23	11.74
38	8	0.3641329	0.36	12.11
39	6	0.2730997	0.27	12.38
40	1	0.0455166	0.05	12.43
41	5	0.2275831	0.23	12.65
42	4	0.1820665	0.18	12.84
43	4	0.1820665	0.18	13.02
44	1	0.0455166	0.05	13.06
45	16	0.7282658	0.73	13.79
46	30	1.3654984	1.37	15.16
47	30	1.3654984	1.37	16.52
48	18	0.8192990	0.82	17.34
49	11	0.5006827	0.50	17.84

(continued)

caseno	N	exactpercent	roundedpercent	cumulpercent
50	31	1.4110150	1.41	19.25
51	15	0.6827492	0.68	19.94
52	15	0.6827492	0.68	20.62
53	16	0.7282658	0.73	21.35
54	14	0.6372326	0.64	21.98
55	24	1.0923987	1.09	23.08
56	25	1.1379153	1.14	24.21
57	11	0.5006827	0.50	24.72
58	31	1.4110150	1.41	26.13
59	29	1.3199818	1.32	27.45
60	5	0.2275831	0.23	27.67
61	15	0.6827492	0.68	28.36
62	20	0.9103323	0.91	29.27
63	14	0.6372326	0.64	29.90
64	7	0.3186163	0.32	30.22
65	11	0.5006827	0.50	30.72
66	11	0.5006827	0.50	31.22
67	12	0.5461994	0.55	31.77
68	22	1.0013655	1.00	32.77
69	8	0.3641329	0.36	33.14
70	32	1.4565316	1.46	34.59
71	5	0.2275831	0.23	34.82
72	10	0.4551661	0.46	35.28
73	3	0.1365498	0.14	35.41
74	11	0.5006827	0.50	35.91
75	20	0.9103323	0.91	36.82
76	6	0.2730997	0.27	37.10
77	6	0.2730997	0.27	37.37
78	13	0.5917160	0.59	37.96

(continued)

caseno	N	exactpercent	roundedpercent	cumulpercent
79	9	0.4096495	0.41	38.37
80	11	0.5006827	0.50	38.87
81	5	0.2275831	0.23	39.10
82	15	0.6827492	0.68	39.78
83	8	0.3641329	0.36	40.15
84	10	0.4551661	0.46	40.60
85	1	0.0455166	0.05	40.65
86	7	0.3186163	0.32	40.96
87	29	1.3199818	1.32	42.28
88	28	1.2744652	1.27	43.56
89	25	1.1379153	1.14	44.70
90	32	1.4565316	1.46	46.15
91	43	1.9572144	1.96	48.11
92	16	0.7282658	0.73	48.84
93	9	0.4096495	0.41	49.25
94	34	1.5475649	1.55	50.80
95	16	0.7282658	0.73	51.52
96	13	0.5917160	0.59	52.12
97	8	0.3641329	0.36	52.48
98	12	0.5461994	0.55	53.03
99	7	0.3186163	0.32	53.35
100	6	0.2730997	0.27	53.62
101	4	0.1820665	0.18	53.80
102	14	0.6372326	0.64	54.44
103	24	1.0923987	1.09	55.53
104	11	0.5006827	0.50	56.03
105	21	0.9558489	0.96	56.99
106	21	0.9558489	0.96	57.94
107	19	0.8648157	0.86	58.81

(continued)

caseno	N	exactpercent	roundedpercent	cumulpercent
108	19	0.8648157	0.86	59.67
109	20	0.9103323	0.91	60.58
110	21	0.9558489	0.96	61.54
111	21	0.9558489	0.96	62.49
112	9	0.4096495	0.41	62.90
113	21	0.9558489	0.96	63.86
114	7	0.3186163	0.32	64.18
115	3	0.1365498	0.14	64.31
116	32	1.4565316	1.46	65.77
117	3	0.1365498	0.14	65.91
118	31	1.4110150	1.41	67.32
119	7	0.3186163	0.32	67.64
120	8	0.3641329	0.36	68.00
121	22	1.0013655	1.00	69.00
122	6	0.2730997	0.27	69.28
123	8	0.3641329	0.36	69.64
124	35	1.5930815	1.59	71.23
125	6	0.2730997	0.27	71.51
126	16	0.7282658	0.73	72.23
127	4	0.1820665	0.18	72.42
128	11	0.5006827	0.50	72.92
129	11	0.5006827	0.50	73.42
130	9	0.4096495	0.41	73.83
131	11	0.5006827	0.50	74.33
132	4	0.1820665	0.18	74.51
133	6	0.2730997	0.27	74.78
134	1	0.0455166	0.05	74.83
135	20	0.9103323	0.91	75.74
136	11	0.5006827	0.50	76.24

(continued)

caseno	N	exactpercent	roundedpercent	cumulpercent
137	13	0.5917160	0.59	76.83
138	4	0.1820665	0.18	77.01
139	8	0.3641329	0.36	77.38
140	16	0.7282658	0.73	78.11
141	11	0.5006827	0.50	78.61
142	8	0.3641329	0.36	78.97
143	15	0.6827492	0.68	79.65
144	17	0.7737824	0.77	80.43
145	3	0.1365498	0.14	80.56
146	5	0.2275831	0.23	80.79
147	1	0.0455166	0.05	80.84
148	17	0.7737824	0.77	81.61
149	8	0.3641329	0.36	81.98
150	40	1.8206645	1.82	83.80
151	14	0.6372326	0.64	84.43
152	16	0.7282658	0.73	85.16
153	14	0.6372326	0.64	85.80
154	13	0.5917160	0.59	86.39
155	18	0.8192990	0.82	87.21
156	10	0.4551661	0.46	87.66
157	13	0.5917160	0.59	88.26
158	17	0.7737824	0.77	89.03
159	17	0.7737824	0.77	89.80
160	17	0.7737824	0.77	90.58
161	17	0.7737824	0.77	91.35
162	4	0.1820665	0.18	91.53
163	26	1.1834320	1.18	92.72
164	12	0.5461994	0.55	93.26
165	10	0.4551661	0.46	93.72

(continued)

caseno	N	exactpercent	roundedpercent	cumulpercent
166	24	1.0923987	1.09	94.81
167	1	0.0455166	0.05	94.86
168	11	0.5006827	0.50	95.36
169	15	0.6827492	0.68	96.04
170	1	0.0455166	0.05	96.09
171	10	0.4551661	0.46	96.54
172	23	1.0468821	1.05	97.59
173	6	0.2730997	0.27	97.86
174	6	0.2730997	0.27	98.13
175	12	0.5461994	0.55	98.68
176	1	0.0455166	0.05	98.73
177	3	0.1365498	0.14	98.86
178	9	0.4096495	0.41	99.27
179	2	0.0910332	0.09	99.36
180	5	0.2275831	0.23	99.59
181	3	0.1365498	0.14	99.73
182	4	0.1820665	0.18	99.91
183	2	0.0910332	0.09	100.00
Total	2197	100.0000000	100.00	100.00

Frequency Table for Variable: shortname

148 unique value(s) detected.

shortname	N	exactpercent	roundedpercent	cumulpercent
1955AmityTreaty	12	0.5461994	0.55	0.55
ATILO-UNESCO	9	0.4096495	0.41	0.96
AccessPacificOcean	14	0.6372326	0.64	1.59
AdmissionUN	7	0.3186163	0.32	1.91

(continued)

shortname	N	exactpercent	roundedpercent	cumulpercent
AegeanSeaContinentalShelf	20	0.9103323	0.91	2.82
AerialHerbicideSpraying	4	0.1820665	0.18	3.00
AerialIncident1952	1	0.0455166	0.05	3.05
AerialIncident1953	1	0.0455166	0.05	3.10
AerialIncident1988	9	0.4096495	0.41	3.50
AerialIncident1999	7	0.3186163	0.32	3.82
AerialIncidentNov1954	1	0.0455166	0.05	3.87
AerialIncidentSept1954	1	0.0455166	0.05	3.91
AerialIndicent1955	21	0.9558489	0.96	4.87
Ambatielos	16	0.7282658	0.73	5.60
AngloIranianOil	12	0.5461994	0.55	6.14
Antarctica	2	0.0910332	0.09	6.24
ApplicationCERD	31	1.4110150	1.41	7.65
ApplicationGenocideConvention	87	3.9599454	3.96	11.61
ApplicationGenocideConvention-Revision	6	0.2730997	0.27	11.88
ArbitralAward1899	10	0.4551661	0.46	12.34
ArbitralAward1989	15	0.6827492	0.68	13.02
ArbitralAwardKingOfSpain	6	0.2730997	0.27	13.29
ArbitrationUNHQAgreement	6	0.2730997	0.27	13.56
ArmedActivities	38	1.7296313	1.73	15.29
ArmedActivitiesApp2002	16	0.7282658	0.73	16.02
ArrestWarrant	22	1.0013655	1.00	17.02
Asylum	9	0.4096495	0.41	17.43
Asylum-Interpretation	1	0.0455166	0.05	17.48
Avena	11	0.5006827	0.50	17.98
Avena-Interpretation	8	0.3641329	0.36	18.34
BarcelonaTraction1958	5	0.2275831	0.23	18.57
BarcelonaTraction1962	31	1.4110150	1.41	19.98

(continued)

shortname	N	exactpercent	roundedpercent	cumulpercent
CertainActivitiesBorderArea	40	1.8206645	1.82	21.80
CertainCriminalProceedings	11	0.5006827	0.50	22.30
CertainDocumentsSeizure	10	0.4551661	0.46	22.76
CertainExpensesUN	11	0.5006827	0.50	23.26
CertainPhosphateLands	11	0.5006827	0.50	23.76
CertainProperty	8	0.3641329	0.36	24.12
ChagosArchipelago	15	0.6827492	0.68	24.81
CompensationUNAT	6	0.2730997	0.27	25.08
CompetenceAdmissionGA	4	0.1820665	0.18	25.26
ConstitutionMaritimeSafetyCommittee	4	0.1820665	0.18	25.44
ConstructionWallOPT	11	0.5006827	0.50	25.94
ContinentalShelf	36	1.6385981	1.64	27.58
ContinentalShelf-InterpretationRevision	5	0.2275831	0.23	27.81
ConventionPrivilegesImmunitiesUN	5	0.2275831	0.23	28.04
ConventionTerrorismFinancingCERD	24	1.0923987	1.09	29.13
CorfuChannel	19	0.8648157	0.86	30.00
DelimitationContinentalShelf	13	0.5917160	0.59	30.59
Diallo	24	1.0923987	1.09	31.68
DiplomaticEnvoyUN	1	0.0455166	0.05	31.73
DiplomaticRelations	1	0.0455166	0.05	31.77
ELSI	6	0.2730997	0.27	32.04
EastTimor	10	0.4551661	0.46	32.50
ElectriciteBeyrouth	3	0.1365498	0.14	32.64
Fisheries	9	0.4096495	0.41	33.05
FisheriesJurisdiction	62	2.8220300	2.82	35.87
FrenchNationalsEgypt	1	0.0455166	0.05	35.91
FrontierDispute	22	1.0013655	1.00	36.91
GabcikovoNagymaros	16	0.7282658	0.73	37.64

(continued)

shortname	N	exactpercent	roundedpercent	cumulpercent
GuardianshipInfantsConvention	11	0.5006827	0.50	38.14
GuatemalaTerritorialInsularMaritimeClaim	3	0.1365498	0.14	38.28
GulfOfMaine	12	0.5461994	0.55	38.83
HayaDeLaTorre	2	0.0910332	0.09	38.92
ICAOCouncil	14	0.6372326	0.64	39.55
ICAOCouncil-CICA	6	0.2730997	0.27	39.83
ICAOCouncil-IASTA	6	0.2730997	0.27	40.10
ICERD	16	0.7282658	0.73	40.83
ImmunitiesCriminalProceedings	26	1.1834320	1.18	42.01
ImmunitySRCommHR	6	0.2730997	0.27	42.28
IndependenceDeclarationKosovo	11	0.5006827	0.50	42.79
Interhandel	16	0.7282658	0.73	43.51
InterimAccord1995	8	0.3641329	0.36	43.88
IranianAssets	12	0.5461994	0.55	44.42
IslaPortillos	10	0.4551661	0.46	44.88
Jadhav	11	0.5006827	0.50	45.38
Judgment2867ATILO-IFAD	5	0.2275831	0.23	45.61
JudgmentsCivilCommercialMatters	3	0.1365498	0.14	45.74
JurisdictionalImmunities2008	15	0.6827492	0.68	46.43
JurisdictionalImmunities2022	2	0.0910332	0.09	46.52
KasikiliSedudu	12	0.5461994	0.55	47.06
LaGrand	11	0.5006827	0.50	47.56
LandIslandMaritimeFrontier	20	0.9103323	0.91	48.48
LandIslandMaritimeFrontier-Revision	4	0.1820665	0.18	48.66
LandMaritimeBoundary	34	1.5475649	1.55	50.20
LandMaritimeBoundary-Interpretation	4	0.1820665	0.18	50.39
LandMaritimeDelimitationSovereigntyIslands	2	0.0910332	0.09	50.48

(continued)

shortname	N	exactpercent	roundedpercent	cumulpercent
LegalityNuclearWeaponsArmedConflict	9	0.4096495	0.41	50.89
LegalityThreatUseNuclearWeapons	16	0.7282658	0.73	51.62
Lockerbie	53	2.4123805	2.41	54.03
MaritimeDelimitation	30	1.3654984	1.37	55.39
MaritimeDelimitation-BlackSea	4	0.1820665	0.18	55.58
MaritimeDelimitation-CaribbeanPacific	13	0.5917160	0.59	56.17
MaritimeDelimitation-GreenlandJanMayen	13	0.5917160	0.59	56.76
MaritimeDelimitation-IndianOcean	17	0.7737824	0.77	57.53
MaritimeDispute	13	0.5917160	0.59	58.12
MilitaryParamilitaryActivitiesNicaragua	32	1.4565316	1.46	59.58
MinquiersEcrehos	7	0.3186163	0.32	59.90
MonetaryGold	7	0.3186163	0.32	60.22
MutualAssistanceCriminalMatters	11	0.5006827	0.50	60.72
Namibia	16	0.7282658	0.73	61.45
NavigationalRights	6	0.2730997	0.27	61.72
NorthSeaContinentalShelf	30	1.3654984	1.37	63.09
NorthernCameroons	18	0.8192990	0.82	63.91
NorwegianLoans	11	0.5006827	0.50	64.41
Nottebohm	11	0.5006827	0.50	64.91
NuclearDisarmament	51	2.3213473	2.32	67.23
NuclearTests	60	2.7309968	2.73	69.96
NuclearTests-ExaminationSituation	8	0.3641329	0.36	70.32
ObligationProsecuteExtradite	17	0.7737824	0.77	71.10
OilPlatforms	32	1.4565316	1.46	72.55
PassageGreatBelt	7	0.3186163	0.32	72.87
PassageIndianTerritory	25	1.1379153	1.14	74.01

(continued)

shortname	N	exactpercent	roundedpercent	cumulpercent
PeaceTreaties	10	0.4551661	0.46	74.47
PedraBranca	9	0.4096495	0.41	74.87
PedraBranca-Interpretation	1	0.0455166	0.05	74.92
PedraBranca-Revision	1	0.0455166	0.05	74.97
PetitionersComitteeSouthWestAfrica	6	0.2730997	0.27	75.24
PortBeyrouthSRO	4	0.1820665	0.18	75.42
PulpMills	20	0.9103323	0.91	76.33
RelocationEmbassyUSJerusalem	1	0.0455166	0.05	76.38
ReparationUN	7	0.3186163	0.32	76.70
ReservationsGenocideConvention	4	0.1820665	0.18	76.88
ReviewJudgment158UNAT	11	0.5006827	0.50	77.38
ReviewJudgment273UNAT	11	0.5006827	0.50	77.88
ReviewJudgment333UNAT	10	0.4551661	0.46	78.33
SanJuanRiver	16	0.7282658	0.73	79.06
SilalaWaters	4	0.1820665	0.18	79.24
SouthWestAfrica	60	2.7309968	2.73	81.98
SovereignRightsCaribbeanSea	18	0.8192990	0.82	82.79
SovereigntyFrontierLand	8	0.3641329	0.36	83.16
SovereigntyPulau	14	0.6372326	0.64	83.80
StatusSouthWestAfrica	7	0.3186163	0.32	84.11
TemplePreahVihear	16	0.7282658	0.73	84.84
TemplePreahVihear-Interpretation	14	0.6372326	0.64	85.48
TerritorialDispute	43	1.9572144	1.96	87.44
TerritorialDispute-CaribbeanSea	8	0.3641329	0.36	87.80
TransborderArmedActions	14	0.6372326	0.64	88.44
TreatmentAirCrew	2	0.0910332	0.09	88.53
TrialPakistaniPOW	5	0.2275831	0.23	88.76

(continued)

shortname	N	exactpercent	roundedpercent	cumulpercent
USDiplomaticStaffTehran	7	0.3186163	0.32	89.08
USNationalsMorocco	6	0.2730997	0.27	89.35
UseOfForce	179	8.1474738	8.15	97.50
ViennaConventionConsularRelations	7	0.3186163	0.32	97.82
VotingProcedureSouthWestAfrica	5	0.2275831	0.23	98.04
WHO-EgyptAgreement	11	0.5006827	0.50	98.54
WesternSahara	15	0.6827492	0.68	99.23
WhalingAntarctic	17	0.7737824	0.77	100.00
Total	2197	100.0000000	100.00	100.00

Frequency Table for Variable: applicant

81 unique value(s) detected.

applicant	N	exactpercent	roundedpercent	cumulpercent
ARG	20	0.9103323	0.91	0.91
ARM	5	0.2275831	0.23	1.14
AUS	48	2.1847975	2.18	3.32
AZE	3	0.1365498	0.14	3.46
BEL	64	2.9130633	2.91	6.37
BEN	6	0.2730997	0.27	6.65
BFA	16	0.7282658	0.73	7.37
BHR-EGY-ARE	6	0.2730997	0.27	7.65
BHR-EGY-SAU-ARE	6	0.2730997	0.27	7.92
BIH	49	2.2303141	2.23	10.15
BOL	14	0.6372326	0.64	10.79
BWA	12	0.5461994	0.55	11.33
CAN	12	0.5461994	0.55	11.88
CARAT	32	1.4565316	1.46	13.34

(continued)

applicant	N	exactpercent	roundedpercent	cumulpercent
CHE	16	0.7282658	0.73	14.06
CHL	4	0.1820665	0.18	14.25
CMR	56	2.5489304	2.55	16.80
COD	87	3.9599454	3.96	20.76
COL	12	0.5461994	0.55	21.30
CRI	69	3.1406463	3.14	24.44
DEU	83	3.7778789	3.78	28.22
DJI	11	0.5006827	0.50	28.72
DMA	1	0.0455166	0.05	28.77
DNK	13	0.5917160	0.59	29.36
ECOSOC	11	0.5006827	0.50	29.86
ECU	4	0.1820665	0.18	30.04
ESP	13	0.5917160	0.59	30.63
ETH	30	1.3654984	1.37	32.00
FIN	7	0.3186163	0.32	32.32
FRA	32	1.4565316	1.46	33.77
GAB	2	0.0910332	0.09	33.86
GBR	77	3.5047792	3.50	37.37
GEO	16	0.7282658	0.73	38.10
GIN	24	1.0923987	1.09	39.19
GMB	9	0.4096495	0.41	39.60
GNB	16	0.7282658	0.73	40.33
GNQ	26	1.1834320	1.18	41.51
GRC	36	1.6385981	1.64	43.15
GTM	3	0.1365498	0.14	43.29
GUY	10	0.4551661	0.46	43.74
HND	7	0.3186163	0.32	44.06
HRV	31	1.4110150	1.41	45.47
HUN	16	0.7282658	0.73	46.20

(continued)

applicant	N	exactpercent	roundedpercent	cumulpercent
IDN	14	0.6372326	0.64	46.84
IFAD	5	0.2275831	0.23	47.06
IMO	4	0.1820665	0.18	47.25
IND	25	1.1379153	1.14	48.38
IRN	65	2.9585799	2.96	51.34
ISR	8	0.3641329	0.36	51.71
ITA	7	0.3186163	0.32	52.03
KHM	30	1.3654984	1.37	53.39
LBR	30	1.3654984	1.37	54.76
LBY	83	3.7778789	3.78	58.53
LIE	19	0.8648157	0.86	59.40
MEX	19	0.8648157	0.86	60.26
MHL	51	2.3213473	2.32	62.59
MKD	8	0.3641329	0.36	62.95
MYS	11	0.5006827	0.50	63.45
NIC	136	6.1902594	6.19	69.64
NLD	11	0.5006827	0.50	70.14
NRU	11	0.5006827	0.50	70.64
NZL	37	1.6841147	1.68	72.33
PAK	12	0.5461994	0.55	72.87
PER	13	0.5917160	0.59	73.46
PRT	35	1.5930815	1.59	75.06
PRY	7	0.3186163	0.32	75.38
PSE	1	0.0455166	0.05	75.42
QAT	52	2.3668639	2.37	77.79
ROU	4	0.1820665	0.18	77.97
SCG	163	7.4192080	7.42	85.39
SLV	24	1.0923987	1.09	86.48
SOM	17	0.7737824	0.77	87.26

(continued)

applicant	N	exactpercent	roundedpercent	cumulpercent
TLS	10	0.4551661	0.46	87.71
TUN	19	0.8648157	0.86	88.58
UKR	28	1.2744652	1.27	89.85
UNESCO	9	0.4096495	0.41	90.26
UNGA	141	6.4178425	6.42	96.68
UNSC	16	0.7282658	0.73	97.41
USA	21	0.9558489	0.96	98.36
WHO	20	0.9103323	0.91	99.27
YUG	16	0.7282658	0.73	100.00
Total	2197	100.0000000	100.00	100.00

Frequency Table for Variable: respondent

74 unique value(s) detected.

respondent	N	exactpercent	roundedpercent	cumulpercent
NA	238	10.8329540	10.83	10.83
ALB	19	0.8648157	0.86	11.70
ARE	23	1.0468821	1.05	12.74
ARG	1	0.0455166	0.05	12.79
ARM	3	0.1365498	0.14	12.93
AUS	31	1.4110150	1.41	14.34
AZE	5	0.2275831	0.23	14.57
BDI	3	0.1365498	0.14	14.70
BEL	43	1.9572144	1.96	16.66
BGR	21	0.9558489	0.96	17.61
BHR	29	1.3199818	1.32	18.93
BLZ	3	0.1365498	0.14	19.07
BOL	4	0.1820665	0.18	19.25

(continued)

respondent	N	exactpercent	roundedpercent	cumulpercent
BRA	1	0.0455166	0.05	19.30
CAN	34	1.5475649	1.55	20.85
CHE	4	0.1820665	0.18	21.03
CHL	28	1.2744652	1.27	22.30
COD	24	1.0923987	1.09	23.40
COL	70	3.1861629	3.19	26.58
CRI	19	0.8648157	0.86	27.45
CSK	1	0.0455166	0.05	27.49
DEU	27	1.2289486	1.23	28.72
DNK	22	1.0013655	1.00	29.72
EGY	1	0.0455166	0.05	29.77
ESP	52	2.3668639	2.37	32.13
FRA	135	6.1447428	6.14	38.28
FRA-GBR-USA	7	0.3186163	0.32	38.60
GBR	107	4.8702777	4.87	43.47
GNQ	2	0.0910332	0.09	43.56
GRC	8	0.3641329	0.36	43.92
GTM	11	0.5006827	0.50	44.42
HND	43	1.9572144	1.96	46.38
HUN	1	0.0455166	0.05	46.43
IND	54	2.4578971	2.46	48.88
IRN	19	0.8648157	0.86	49.75
ISL	49	2.2303141	2.23	51.98
ITA	43	1.9572144	1.96	53.94
JPN	17	0.7737824	0.77	54.71
KEN	17	0.7737824	0.77	55.48
LBN	7	0.3186163	0.32	55.80
LBY	19	0.8648157	0.86	56.67
MLI	8	0.3641329	0.36	57.03

(continued)

respondent	N	exactpercent	roundedpercent	cumulpercent
MLT	22	1.0013655	1.00	58.03
MMR	9	0.4096495	0.41	58.44
MYS	14	0.6372326	0.64	59.08
NAM	12	0.5461994	0.55	59.63
NER	14	0.6372326	0.64	60.26
NGA	38	1.7296313	1.73	61.99
NIC	75	3.4137460	3.41	65.41
NLD	44	2.0027310	2.00	67.41
NOR	33	1.5020482	1.50	68.91
PAK	42	1.9116978	1.91	70.82
PER	12	0.5461994	0.55	71.37
PRT	21	0.9558489	0.96	72.33
QAT	12	0.5461994	0.55	72.87
RUS	44	2.0027310	2.00	74.87
RWA	19	0.8648157	0.86	75.74
SCG	43	1.9572144	1.96	77.70
SEN	33	1.5020482	1.50	79.20
SGP	11	0.5006827	0.50	79.70
SRB	31	1.4110150	1.41	81.11
SUN	4	0.1820665	0.18	81.29
SVK	16	0.7282658	0.73	82.02
SWE	11	0.5006827	0.50	82.52
TCD	8	0.3641329	0.36	82.89
THA	30	1.3654984	1.37	84.25
TUR	20	0.9103323	0.91	85.16
UGA	32	1.4565316	1.46	86.62
UKR	4	0.1820665	0.18	86.80
URY	20	0.9103323	0.91	87.71
USA	194	8.8302230	8.83	96.54

(continued)

respondent	N	exactpercent	roundedpercent	cumulpercent
VEN	10	0.4551661	0.46	97.00
YUG	6	0.2730997	0.27	97.27
ZAF	60	2.7309968	2.73	100.00
Total	2197	100.0000000	100.00	100.00

Frequency Table for Variable: doctype

3 unique value(s) detected.

doctype	N	exactpercent	roundedpercent	cumulpercent
ADV	194	8.830223	8.83	8.83
JUD	1051	47.837961	47.84	56.67
ORD	952	43.331816	43.33	100.00
Total	2197	100.000000	100.00	100.00

Frequency Table for Variable: collision

3 unique value(s) detected.

collision	N	exactpercent	roundedpercent	cumulpercent
1	2180	99.2262176	99.23	99.23
2	16	0.7282658	0.73	99.95
3	1	0.0455166	0.05	100.00
Total	2197	100.0000000	100.00	100.00

Frequency Table for Variable: stage

5 unique value(s) detected.

stage	N	exactpercent	roundedpercent	cumulpercent
NA	1161	52.844788	52.84	52.84
CO	33	1.502048	1.50	54.35
IN	36	1.638598	1.64	55.99
ME	543	24.715521	24.72	80.70
PO	424	19.299044	19.30	100.00
Total	2197	100.000000	100.00	100.00

Frequency Table for Variable: opinion

15 unique value(s) detected.

opinion	N	exactpercent	roundedpercent	cumulpercent
0	780	35.5029586	35.50	35.50
1	256	11.6522531	11.65	47.16
2	230	10.4688211	10.47	57.62
3	202	9.1943559	9.19	66.82
4	172	7.8288575	7.83	74.65
5	147	6.6909422	6.69	81.34
6	127	5.7806099	5.78	87.12
7	95	4.3240783	4.32	91.44
8	71	3.2316796	3.23	94.67
9	57	2.5944470	2.59	97.27
10	30	1.3654984	1.37	98.63
11	14	0.6372326	0.64	99.27
12	8	0.3641329	0.36	99.64
13	4	0.1820665	0.18	99.82
14	4	0.1820665	0.18	100.00
Total	2197	100.0000000	100.00	100.00

Frequency Table for Variable: language

1 unique value(s) detected.

language	N	exactpercent	roundedpercent	cumulpercent
FR	2197	100	100	100
Total	2197	100	100	100

Frequency Table for Variable: year

76 unique value(s) detected.

year	N	exactpercent	roundedpercent	cumulpercent
1947	3	0.1365498	0.14	0.14
1948	12	0.5461994	0.55	0.68
1949	24	1.0923987	1.09	1.78
1950	31	1.4110150	1.41	3.19
1951	21	0.9558489	0.96	4.14
1952	26	1.1834320	1.18	5.33
1953	11	0.5006827	0.50	5.83
1954	19	0.8648157	0.86	6.69
1955	11	0.5006827	0.50	7.19
1956	22	1.0013655	1.00	8.19
1957	22	1.0013655	1.00	9.19
1958	28	1.2744652	1.27	10.47
1959	32	1.4565316	1.46	11.93
1960	26	1.1834320	1.18	13.11
1961	17	0.7737824	0.77	13.88
1962	42	1.9116978	1.91	15.79
1963	17	0.7737824	0.77	16.57
1964	15	0.6827492	0.68	17.25
1965	5	0.2275831	0.23	17.48

(continued)

year	N	exactpercent	roundedpercent	cumulpercent
1966	24	1.0923987	1.09	18.57
1967	4	0.1820665	0.18	18.75
1968	5	0.2275831	0.23	18.98
1969	24	1.0923987	1.09	20.07
1970	14	0.6372326	0.64	20.71
1971	16	0.7282658	0.73	21.44
1972	23	1.0468821	1.05	22.49
1973	62	2.8220300	2.82	25.31
1974	52	2.3668639	2.37	27.67
1975	15	0.6827492	0.68	28.36
1976	11	0.5006827	0.50	28.86
1977	1	0.0455166	0.05	28.90
1978	8	0.3641329	0.36	29.27
1979	3	0.1365498	0.14	29.40
1980	16	0.7282658	0.73	30.13
1981	8	0.3641329	0.36	30.50
1982	24	1.0923987	1.09	31.59
1983	2	0.0910332	0.09	31.68
1984	35	1.5930815	1.59	33.27
1985	18	0.8192990	0.82	34.09
1986	17	0.7737824	0.77	34.87
1987	17	0.7737824	0.77	35.64
1988	14	0.6372326	0.64	36.28
1989	21	0.9558489	0.96	37.23
1990	15	0.6827492	0.68	37.92
1991	23	1.0468821	1.05	38.96
1992	43	1.9572144	1.96	40.92
1993	29	1.3199818	1.32	42.24
1994	14	0.6372326	0.64	42.88

(continued)

year	N	exactpercent	roundedpercent	cumulpercent
1995	29	1.3199818	1.32	44.20
1996	55	2.5034137	2.50	46.70
1997	19	0.8648157	0.86	47.56
1998	61	2.7765134	2.78	50.34
1999	134	6.0992262	6.10	56.44
2000	37	1.6841147	1.68	58.12
2001	44	2.0027310	2.00	60.13
2002	49	2.2303141	2.23	62.36
2003	35	1.5930815	1.59	63.95
2004	77	3.5047792	3.50	67.46
2005	21	0.9558489	0.96	68.41
2006	18	0.8192990	0.82	69.23
2007	39	1.7751479	1.78	71.01
2008	42	1.9116978	1.91	72.92
2009	20	0.9103323	0.91	73.83
2010	42	1.9116978	1.91	75.74
2011	57	2.5944470	2.59	78.33
2012	36	1.6385981	1.64	79.97
2013	33	1.5020482	1.50	81.47
2014	40	1.8206645	1.82	83.30
2015	51	2.3213473	2.32	85.62
2016	75	3.4137460	3.41	89.03
2017	35	1.5930815	1.59	90.62
2018	64	2.9130633	2.91	93.54
2019	54	2.4578971	2.46	95.99
2020	35	1.5930815	1.59	97.59
2021	31	1.4110150	1.41	99.00
2022	22	1.0013655	1.00	100.00
Total	2197	100.0000000	100.00	100.00

Frequency Table for Variable: minority

2 unique value(s) detected.

minority	N	exactpercent	roundedpercent	cumulpercent
0	780	35.50296	35.5	35.5
1	1417	64.49704	64.5	100.0
Total	2197	100.00000	100.0	100.0

Frequency Table for Variable: fullname

182 unique value(s) detected.

fullname	N	exactpercent	roundedpercent	cumulpercent
Accordance with international law of the unilateral declaration of independence in respect of Kosovo	11	0.5006827	0.50	0.50
Admissibility of Hearings of Petitioners by the Committee on South West Africa	6	0.2730997	0.27	0.77
Aegean Sea Continental Shelf (Greece v. Turkey)	20	0.9103323	0.91	1.68
Aerial Herbicide Spraying (Ecuador v. Colombia)	4	0.1820665	0.18	1.87
Aerial Incident of 10 August 1999 (Pakistan v. India)	7	0.3186163	0.32	2.18
Aerial Incident of 10 March 1953 (United States of America v. Czechoslovakia)	1	0.0455166	0.05	2.23
Aerial Incident of 27 July 1955 (Israel v. Bulgaria)	8	0.3641329	0.36	2.59
Aerial Incident of 27 July 1955 (United Kingdom v. Bulgaria)	5	0.2275831	0.23	2.82
Aerial Incident of 27 July 1955 (United States of America v. Bulgaria)	8	0.3641329	0.36	3.19

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Aerial Incident of 3 July 1988 (Islamic Republic of Iran v. United States of America)	9	0.4096495	0.41	3.60
Aerial Incident of 4 September 1954 (United States of America v. Union of Soviet Socialist Republics)	1	0.0455166	0.05	3.64
Aerial Incident of 7 November 1954 (United States of America v. Union of Soviet Socialist Republics)	1	0.0455166	0.05	3.69
Aerial Incident of 7 October 1952 (United States of America v. Union of Soviet Socialist Republics)	1	0.0455166	0.05	3.73
Ahmadou Sadio Diallo (Republic of Guinea v. Democratic Republic of the Congo)	24	1.0923987	1.09	4.82
Allegations of Genocide under the Convention on the Prevention and Punishment of the Crime of Genocide (Ukraine v. Russian Federation)	4	0.1820665	0.18	5.01
Alleged Violations of Sovereign Rights and Maritime Spaces in the Caribbean Sea (Nicaragua v. Colombia)	18	0.8192990	0.82	5.83
Alleged violations of the 1955 Treaty of Amity, Economic Relations, and Consular Rights (Islamic Republic of Iran v. United States of America)	12	0.5461994	0.55	6.37
Ambatielos (Greece v. United Kingdom)	16	0.7282658	0.73	7.10
Anglo-Iranian Oil Co. (United Kingdom v. Iran)	12	0.5461994	0.55	7.65
Antarctica (United Kingdom v. Argentina)	1	0.0455166	0.05	7.69

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Antarctica (United Kingdom v. Chile)	1	0.0455166	0.05	7.74
Appeal Relating to the Jurisdiction of the ICAO Council (India v. Pakistan)	14	0.6372326	0.64	8.38
Appeal Relating to the Jurisdiction of the ICAO Council under Article 84 of the Convention on International Civil Aviation (Bahrain, Egypt, Saudi Arabia and United Arab Emirates v. Qatar)	6	0.2730997	0.27	8.65
Appeal Relating to the Jurisdiction of the ICAO Council under Article II, Section 2, of the 1944 International Air Services Transit Agreement (Bahrain, Egypt and United Arab Emirates v. Qatar)	6	0.2730997	0.27	8.92
Applicability of Article VI, Section 22, of the Convention on the Privileges and Immunities of the United Nations	5	0.2275831	0.23	9.15
Applicability of the Obligation to Arbitrate under Section 21 of the United Nations Headquarters Agreement of 26 June 1947	6	0.2730997	0.27	9.42
Application for Review of Judgment No. 158 of the United Nations Administrative Tribunal	11	0.5006827	0.50	9.92
Application for Review of Judgment No. 273 of the United Nations Administrative Tribunal	11	0.5006827	0.50	10.42
Application for Review of Judgment No. 333 of the United Nations Administrative Tribunal	10	0.4551661	0.46	10.88

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Application for Revision and Interpretation of the Judgment of 24 February 1982 in the Case concerning the Continental Shelf (Tunisia/Libyan Arab Jamahiriya) (Tunisia v. Libyan Arab Jamahiriya)	5	0.2275831	0.23	11.11
Application for Revision of the Judgment of 11 July 1996 in the Case concerning Application of the Convention on the Prevention and Punishment of the Crime of Genocide (Bosnia and Herzegovina v. Yugoslavia), Preliminary Objections (Yugoslavia v. Bosnia and Herzegovina)	6	0.2730997	0.27	11.38
Application for Revision of the Judgment of 11 September 1992 in the Case concerning the Land, Island and Maritime Frontier Dispute (El Salvador/Honduras: Nicaragua intervening) (El Salvador v. Honduras)	4	0.1820665	0.18	11.56
Application for revision of the Judgment of 23 May 2008 in the case concerning Sovereignty over Pedra Branca/Pulau Batu Puteh, Middle Rocks and South Ledge (Malaysia/Singapore) (Malaysia v. Singapore)	1	0.0455166	0.05	11.61
Application of the Convention of 1902 Governing the Guardianship of Infants (Netherlands v. Sweden)	11	0.5006827	0.50	12.11
Application of the Convention on the Prevention and Punishment of the Crime of Genocide (Bosnia and Herzegovina v. Serbia and Montenegro)	43	1.9572144	1.96	14.06

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Application of the Convention on the Prevention and Punishment of the Crime of Genocide (Croatia v. Serbia)	31	1.4110150	1.41	15.48
Application of the Convention on the Prevention and Punishment of the Crime of Genocide (The Gambia v. Myanmar)	9	0.4096495	0.41	15.89
Application of the Interim Accord of 13 September 1995 (the former Yugoslav Republic of Macedonia v. Greece)	8	0.3641329	0.36	16.25
Application of the International Convention for the Suppression of the Financing of Terrorism and of the International Convention on the Elimination of All Forms of Racial Discrimination (Ukraine v. Russian Federation)	24	1.0923987	1.09	17.34
Application of the International Convention on the Elimination of All Forms of Racial Discrimination (Armenia v. Azerbaijan)	5	0.2275831	0.23	17.57
Application of the International Convention on the Elimination of All Forms of Racial Discrimination (Azerbaijan v. Armenia)	3	0.1365498	0.14	17.71
Application of the International Convention on the Elimination of All Forms of Racial Discrimination (Georgia v. Russian Federation)	16	0.7282658	0.73	18.43
Application of the International Convention on the Elimination of All Forms of Racial Discrimination (Qatar v. United Arab Emirates)	23	1.0468821	1.05	19.48

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Arbitral Award Made by the King of Spain on 23 December 1906 (Honduras v. Nicaragua)	6	0.2730997	0.27	19.75
Arbitral Award of 3 October 1899 (Guyana v. Venezuela)	10	0.4551661	0.46	20.21
Arbitral Award of 31 July 1989 (Guinea-Bissau v. Senegal)	15	0.6827492	0.68	20.89
Armed Activities on the Territory of the Congo (Democratic Republic of the Congo v. Burundi)	3	0.1365498	0.14	21.03
Armed Activities on the Territory of the Congo (Democratic Republic of the Congo v. Rwanda)	3	0.1365498	0.14	21.17
Armed Activities on the Territory of the Congo (Democratic Republic of the Congo v. Uganda)	32	1.4565316	1.46	22.62
Armed Activities on the Territory of the Congo (New Application: 2002) (Democratic Republic of the Congo v. Rwanda)	16	0.7282658	0.73	23.35
Arrest Warrant of 11 April 2000 (Democratic Republic of the Congo v. Belgium)	22	1.0013655	1.00	24.35
Asylum (Colombia v. Peru)	9	0.4096495	0.41	24.76
Avena and Other Mexican Nationals (Mexico v. United States of America)	11	0.5006827	0.50	25.26
Barcelona Traction, Light and Power Company, Limited (Belgium v. Spain)	5	0.2275831	0.23	25.49
Barcelona Traction, Light and Power Company, Limited (Belgium v. Spain) (New Application: 1962)	31	1.4110150	1.41	26.90

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Border and Transborder Armed Actions (Nicaragua v. Costa Rica)	3	0.1365498	0.14	27.04
Border and Transborder Armed Actions (Nicaragua v. Honduras)	11	0.5006827	0.50	27.54
Certain Activities Carried Out by Nicaragua in the Border Area (Costa Rica v. Nicaragua)	40	1.8206645	1.82	29.36
Certain Criminal Proceedings in France (Republic of the Congo v. France)	11	0.5006827	0.50	29.86
Certain Expenses of the United Nations (Article 17, paragraph 2, of the Charter)	11	0.5006827	0.50	30.36
Certain Iranian Assets (Islamic Republic of Iran v. United States of America)	12	0.5461994	0.55	30.91
Certain Norwegian Loans (France v. Norway)	11	0.5006827	0.50	31.41
Certain Phosphate Lands in Nauru (Nauru v. Australia)	11	0.5006827	0.50	31.91
Certain Property (Liechtenstein v. Germany)	8	0.3641329	0.36	32.27
Certain Questions concerning Diplomatic Relations (Honduras v. Brazil)	1	0.0455166	0.05	32.32
Certain Questions of Mutual Assistance in Criminal Matters (Djibouti v. France)	11	0.5006827	0.50	32.82
Compagnie du Port, des Quais et des Entrepôts de Beyrouth and Société Radio-Orient (France v. Lebanon)	4	0.1820665	0.18	33.00
Competence of the General Assembly for the Admission of a State to the United Nations	4	0.1820665	0.18	33.18

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Conditions of Admission of a State to Membership in the United Nations (Article 4 of the Charter)	7	0.3186163	0.32	33.50
Constitution of the Maritime Safety Committee of the Inter-Governmental Maritime Consultative Organization	4	0.1820665	0.18	33.68
Construction of a Road in Costa Rica along the San Juan River (Nicaragua v. Costa Rica)	16	0.7282658	0.73	34.41
Continental Shelf (Libyan Arab Jamahiriya/Malta)	22	1.0013655	1.00	35.41
Continental Shelf (Tunisia/Libyan Arab Jamahiriya)	14	0.6372326	0.64	36.05
Corfu Channel (United Kingdom of Great Britain and Northern Ireland v. Albania)	19	0.8648157	0.86	36.91
Delimitation of the Maritime Boundary in the Gulf of Maine Area (Canada/United States of America)	12	0.5461994	0.55	37.46
Difference Relating to Immunity from Legal Process of a Special Rapporteur of the Commission on Human Rights	6	0.2730997	0.27	37.73
Dispute over the Status and Use of the Waters of the Silala (Chile v. Bolivia)	4	0.1820665	0.18	37.92
Dispute regarding Navigational and Related Rights (Costa Rica v. Nicaragua)	6	0.2730997	0.27	38.19
East Timor (Portugal v. Australia)	10	0.4551661	0.46	38.64
Effect of Awards of Compensation Made by the United Nations Administrative Tribunal	6	0.2730997	0.27	38.92

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Electricité de Beyrouth Company (France v. Lebanon)	3	0.1365498	0.14	39.05
Elettronica Sicula S.p.A. (ELSI) (United States of America v. Italy)	6	0.2730997	0.27	39.33
Fisheries (United Kingdom v. Norway)	9	0.4096495	0.41	39.74
Fisheries Jurisdiction (Federal Republic of Germany v. Iceland)	25	1.1379153	1.14	40.87
Fisheries Jurisdiction (Spain v. Canada)	13	0.5917160	0.59	41.47
Fisheries Jurisdiction (United Kingdom v. Iceland)	24	1.0923987	1.09	42.56
Frontier Dispute (Benin/Niger)	6	0.2730997	0.27	42.83
Frontier Dispute (Burkina Faso/Niger)	8	0.3641329	0.36	43.20
Frontier Dispute (Burkina Faso/Republic of Mali)	8	0.3641329	0.36	43.56
Gabčíkovo-Nagymaros Project (Hungary/Slovakia)	16	0.7282658	0.73	44.29
Guatemala's Territorial, Insular and Maritime Claim (Guatemala/Belize)	3	0.1365498	0.14	44.42
Haya de la Torre (Colombia v. Peru)	2	0.0910332	0.09	44.52
Immunities and Criminal Proceedings (Equatorial Guinea v. France)	26	1.1834320	1.18	45.70
Interhandel (Switzerland v. United States of America)	16	0.7282658	0.73	46.43
International Status of South West Africa	7	0.3186163	0.32	46.75

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Interpretation of Peace Treaties with Bulgaria, Hungary and Romania	10	0.4551661	0.46	47.20
Interpretation of the Agreement of 25 March 1951 between the WHO and Egypt	11	0.5006827	0.50	47.70
Jadhav (India v. Pakistan)	11	0.5006827	0.50	48.20
Judgment No.2867 of the Administrative Tribunal of the International Labour Organization upon a Complaint Filed against the International Fund for Agricultural Development	5	0.2275831	0.23	48.43
Judgments of the Administrative Tribunal of the ILO upon Complaints Made against UNESCO	9	0.4096495	0.41	48.84
Jurisdiction and Enforcement of Judgments in Civil and Commercial Matters (Belgium v. Switzerland)	3	0.1365498	0.14	48.98
Jurisdictional Immunities of the State (Germany v. Italy: Greece intervening)	15	0.6827492	0.68	49.66
Kasikili/Sedudu Island (Botswana/Namibia)	12	0.5461994	0.55	50.20
LaGrand (Germany v. United States of America)	11	0.5006827	0.50	50.71
Land Boundary in the Northern Part of Isla Portillos (Costa Rica v. Nicaragua)	10	0.4551661	0.46	51.16
Land and Maritime Boundary between Cameroon and Nigeria (Cameroon v. Nigeria: Equatorial Guinea intervening)	34	1.5475649	1.55	52.71

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Land and Maritime Delimitation and Sovereignty over Islands (Gabon/Equatorial Guinea)	2	0.0910332	0.09	52.80
Land, Island and Maritime Frontier Dispute (El Salvador/Honduras: Nicaragua intervening)	20	0.9103323	0.91	53.71
Legal Consequences for States of the Continued Presence of South Africa in Namibia (South West Africa) notwithstanding Security Council Resolution 276 (1970)	16	0.7282658	0.73	54.44
Legal Consequences of the Construction of a Wall in the Occupied Palestinian Territory	11	0.5006827	0.50	54.94
Legal Consequences of the Separation of the Chagos Archipelago from Mauritius in 1965	15	0.6827492	0.68	55.62
Legality of Use of Force (Serbia and Montenegro v. Belgium)	21	0.9558489	0.96	56.58
Legality of Use of Force (Serbia and Montenegro v. Canada)	21	0.9558489	0.96	57.53
Legality of Use of Force (Serbia and Montenegro v. France)	19	0.8648157	0.86	58.40
Legality of Use of Force (Serbia and Montenegro v. Germany)	19	0.8648157	0.86	59.26
Legality of Use of Force (Serbia and Montenegro v. Italy)	20	0.9103323	0.91	60.17
Legality of Use of Force (Serbia and Montenegro v. Netherlands)	21	0.9558489	0.96	61.13

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Legality of Use of Force (Serbia and Montenegro v. Portugal)	21	0.9558489	0.96	62.08
Legality of Use of Force (Serbia and Montenegro v. United Kingdom)	21	0.9558489	0.96	63.04
Legality of Use of Force (Yugoslavia v. Spain)	9	0.4096495	0.41	63.45
Legality of Use of Force (Yugoslavia v. United States of America)	7	0.3186163	0.32	63.77
Legality of the Threat or Use of Nuclear Weapons	16	0.7282658	0.73	64.50
Legality of the Use by a State of Nuclear Weapons in Armed Conflict	9	0.4096495	0.41	64.91
Maritime Delimitation and Territorial Questions between Qatar and Bahrain (Qatar v. Bahrain)	29	1.3199818	1.32	66.23
Maritime Delimitation between Guinea-Bissau and Senegal (Guinea-Bissau v. Senegal)	1	0.0455166	0.05	66.27
Maritime Delimitation in the Area between Greenland and Jan Mayen (Denmark v. Norway)	13	0.5917160	0.59	66.86
Maritime Delimitation in the Black Sea (Romania v. Ukraine)	4	0.1820665	0.18	67.05
Maritime Delimitation in the Caribbean Sea and the Pacific Ocean (Costa Rica v. Nicaragua)	13	0.5917160	0.59	67.64
Maritime Delimitation in the Indian Ocean (Somalia v. Kenya)	17	0.7737824	0.77	68.41

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Maritime Dispute (Peru v. Chile)	13	0.5917160	0.59	69.00
Military and Paramilitary Activities in and against Nicaragua (Nicaragua v. United States of America)	32	1.4565316	1.46	70.46
Minquiers and Ecrehos (France/United Kingdom)	7	0.3186163	0.32	70.78
Monetary Gold Removed from Rome in 1943 (Italy v. France, United Kingdom of Great Britain and Northern Ireland and United States of America)	7	0.3186163	0.32	71.10
North Sea Continental Shelf (Federal Republic of Germany/Denmark)	15	0.6827492	0.68	71.78
North Sea Continental Shelf (Federal Republic of Germany/Netherlands)	15	0.6827492	0.68	72.46
Northern Cameroons (Cameroon v. United Kingdom)	18	0.8192990	0.82	73.28
Nottebohm (Liechtenstein v. Guatemala)	11	0.5006827	0.50	73.78
Nuclear Tests (Australia v. France)	31	1.4110150	1.41	75.19
Nuclear Tests (New Zealand v. France)	29	1.3199818	1.32	76.51
Obligation to Negotiate Access to the Pacific Ocean (Bolivia v. Chile)	14	0.6372326	0.64	77.15
Obligations concerning Negotiations relating to Cessation of the Nuclear Arms Race and to Nuclear Disarmament (Marshall Islands v. India)	17	0.7737824	0.77	77.92

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Obligations concerning Negotiations relating to Cessation of the Nuclear Arms Race and to Nuclear Disarmament (Marshall Islands v. Pakistan)	17	0.7737824	0.77	78.70
Obligations concerning Negotiations relating to Cessation of the Nuclear Arms Race and to Nuclear Disarmament (Marshall Islands v. United Kingdom)	17	0.7737824	0.77	79.47
Oil Platforms (Islamic Republic of Iran v. United States of America)	32	1.4565316	1.46	80.93
Passage through the Great Belt (Finland v. Denmark)	7	0.3186163	0.32	81.25
Protection of French Nationals and Protected Persons in Egypt (France v. Egypt)	1	0.0455166	0.05	81.29
Pulp Mills on the River Uruguay (Argentina v. Uruguay)	20	0.9103323	0.91	82.20
Question of the Delimitation of the Continental Shelf between Nicaragua and Colombia beyond 200 nautical miles from the Nicaraguan Coast (Nicaragua v. Colombia)	13	0.5917160	0.59	82.79
Questions of Interpretation and Application of the 1971 Montreal Convention arising from the Aerial Incident at Lockerbie (Libyan Arab Jamahiriya v. United Kingdom)	28	1.2744652	1.27	84.07

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Questions of Interpretation and Application of the 1971 Montreal Convention arising from the Aerial Incident at Lockerbie (Libyan Arab Jamahiriya v. United States of America)	25	1.1379153	1.14	85.21
Questions of jurisdictional immunities of the State and measures of constraint against State-owned property (Germany v. Italy)	2	0.0910332	0.09	85.30
Questions relating to the Obligation to Prosecute or Extradite (Belgium v. Senegal)	17	0.7737824	0.77	86.07
Questions relating to the Seizure and Detention of Certain Documents and Data (Timor-Leste v. Australia)	10	0.4551661	0.46	86.53
Relocation of the United States Embassy to Jerusalem (Palestine v. United States of America)	1	0.0455166	0.05	86.57
Reparation for Injuries Suffered in the Service of the United Nations	7	0.3186163	0.32	86.89
Request for Interpretation of the Judgment of 11 June 1998 in the Case concerning the Land and Maritime Boundary between Cameroon and Nigeria (Cameroon v. Nigeria), Preliminary Objections (Nigeria v. Cameroon)	4	0.1820665	0.18	87.07
Request for Interpretation of the Judgment of 15 June 1962 in the Case concerning the Temple of Preah Vihear (Cambodia v. Thailand) (Cambodia v. Thailand)	14	0.6372326	0.64	87.71

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Request for Interpretation of the Judgment of 20 November 1950 in the Asylum Case (Colombia v. Peru)	1	0.0455166	0.05	87.76
Request for Interpretation of the Judgment of 23 May 2008 in the case concerning Sovereignty over Pedra Branca/Pulau Batu Puteh, Middle Rocks and South Ledge (Malaysia/Singapore) (Malaysia v. Singapore)	1	0.0455166	0.05	87.80
Request for Interpretation of the Judgment of 31 March 2004 in the Case concerning Avena and Other Mexican Nationals (Mexico v. United States of America) (Mexico v. United States of America)	8	0.3641329	0.36	88.17
Request for an Examination of the Situation in Accordance with Paragraph 63 of the Court's Judgment of 20 December 1974 in the Nuclear Tests (New Zealand v. France) Case	8	0.3641329	0.36	88.53
Reservations to the Convention on the Prevention and Punishment of the Crime of Genocide	4	0.1820665	0.18	88.71
Right of Passage over Indian Territory (Portugal v. India)	25	1.1379153	1.14	89.85
Rights of Nationals of the United States of America in Morocco (France v. United States of America)	6	0.2730997	0.27	90.12
South West Africa (Ethiopia v. South Africa)	30	1.3654984	1.37	91.49
South West Africa (Liberia v. South Africa)	30	1.3654984	1.37	92.85

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Sovereignty over Certain Frontier Land (Belgium/Netherlands)	8	0.3641329	0.36	93.22
Sovereignty over Pedra Branca/Pulau Batu Puteh, Middle Rocks and South Ledge (Malaysia/Singapore)	9	0.4096495	0.41	93.63
Sovereignty over Pulau Ligitan and Pulau Sipadan (Indonesia/Malaysia)	14	0.6372326	0.64	94.26
Status vis-à-vis the Host State of a Diplomatic Envoy to the United Nations (Commonwealth of Dominica v. Switzerland)	1	0.0455166	0.05	94.31
Temple of Preah Vihear (Cambodia v. Thailand)	16	0.7282658	0.73	95.04
Territorial Dispute (Libyan Arab Jamahiriya/Chad)	8	0.3641329	0.36	95.40
Territorial and Maritime Dispute (Nicaragua v. Colombia)	35	1.5930815	1.59	97.00
Territorial and Maritime Dispute between Nicaragua and Honduras in the Caribbean Sea (Nicaragua v. Honduras)	8	0.3641329	0.36	97.36
Treatment in Hungary of Aircraft and Crew of United States of America (United States of America v. Hungarian People's Republic)	1	0.0455166	0.05	97.41
Treatment in Hungary of Aircraft and Crew of United States of America (United States of America v. Union of Soviet Socialist Republics)	1	0.0455166	0.05	97.45
Trial of Pakistani Prisoners of War (Pakistan v. India)	5	0.2275831	0.23	97.68

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
United States Diplomatic and Consular Staff in Tehran (United States of America v. Iran)	7	0.3186163	0.32	98.00
Vienna Convention on Consular Relations (Paraguay v. United States of America)	7	0.3186163	0.32	98.32
Voting Procedure on Questions relating to Reports and Petitions concerning the Territory of South West Africa	5	0.2275831	0.23	98.54
Western Sahara	15	0.6827492	0.68	99.23
Whaling in the Antarctic (Australia v. Japan: New Zealand intervening)	17	0.7737824	0.77	100.00
Total	2197	100.0000000	100.00	100.00

Frequency Table for Variable: applicant_region

8 unique value(s) detected.

applicant_region	N	exactpercent	roundedpercent	cumulpercent
Africa	444	20.2093764	20.21	20.21
Americas	376	17.1142467	17.11	37.32
Asia	252	11.4701866	11.47	48.79
Asia|Africa|Asia	6	0.2730997	0.27	49.07
Asia|Africa|Asia|Asia	6	0.2730997	0.27	49.34
Europe	728	33.1360947	33.14	82.48
NA	238	10.8329540	10.83	93.31
Oceania	147	6.6909422	6.69	100.00
Total	2197	100.0000000	100.00	100.00

Frequency Table for Variable: respondent_region

7 unique value(s) detected.

respondent_region	N	exactpercent	roundedpercent	cumulpercent
NA	238	10.8329540	10.83	10.83
Africa	290	13.1998179	13.20	24.03
Americas	525	23.8962221	23.90	47.93
Asia	295	13.4274010	13.43	61.36
Europe	811	36.9139736	36.91	98.27
Europe|Europe|Americas	7	0.3186163	0.32	98.59
Oceania	31	1.4110150	1.41	100.00
Total	2197	100.0000000	100.00	100.00

Frequency Table for Variable: applicant_subregion

17 unique value(s) detected.

applicant_subregion	N	exactpercent	roundedpercent	cumulpercent
Australia and New Zealand	85	3.8689122	3.87	3.87
Eastern Europe	48	2.1847975	2.18	6.05
Latin America and the Caribbean	343	15.6121985	15.61	21.67
Micronesia	62	2.8220300	2.82	24.49
NA	229	10.4233045	10.42	34.91
Northern Africa	102	4.6426946	4.64	39.55
Northern America	33	1.5020482	1.50	41.06
Northern Europe	97	4.4151115	4.42	45.47
South-eastern Asia	65	2.9585799	2.96	48.43
Southern Asia	102	4.6426946	4.64	53.07
Southern Europe	358	16.2949477	16.29	69.37
Sub-Saharan Africa	342	15.5666818	15.57	84.93

(continued)

applicant_subregion	N	exactpercent	roundedpercent	cumulpercent
UNESCO	9	0.4096495	0.41	85.34
Western Asia	85	3.8689122	3.87	89.21
Western Asia|Northern Africa|Western Asia	6	0.2730997	0.27	89.49
Western Asia|Northern Africa|Western Asia|Western Asia	6	0.2730997	0.27	89.76
Western Europe	225	10.2412381	10.24	100.00
Total	2197	100.0000000	100.00	100.00

Frequency Table for Variable: respondent_subregion

15 unique value(s) detected.

respondent_subregion	N	exactpercent	roundedpercent	cumulpercent
NA	238	10.8329540	10.83	10.83
Australia and New Zealand	31	1.4110150	1.41	12.24
Eastern Asia	17	0.7737824	0.77	13.02
Eastern Europe	91	4.1420118	4.14	17.16
Latin America and the Caribbean	297	13.5184342	13.52	30.68
Northern Africa	20	0.9103323	0.91	31.59
Northern America	228	10.3777879	10.38	41.97
Northern Europe	222	10.1046882	10.10	52.07
South-eastern Asia	64	2.9130633	2.91	54.98
Southern Asia	115	5.2344106	5.23	60.22
Southern Europe	245	11.1515703	11.15	71.37
Sub-Saharan Africa	270	12.2894857	12.29	83.66
Western Asia	99	4.5061447	4.51	88.17
Western Europe	253	11.5157032	11.52	99.68

(continued)

respondent_subregion	N	exactpercent	roundedpercent	cumulpercent
Western Europe|Northern Europe|Northern America	7	0.3186163	0.32	100.00
Total	2197	100.0000000	100.00	100.00

Frequency Table for Variable: doi_concept

1 unique value(s) detected.

doi_concept	N	exactpercent	roundedpercent	cumulpercent
10.5281/zenodo.3826444	2197	100	100	100
Total	2197	100	100	100

Frequency Table for Variable: doi_version

1 unique value(s) detected.

doi_version	N	exactpercent	roundedpercent	cumulpercent
10.5281/zenodo.7051929	2197	100	100	100
Total	2197	100	100	100

Frequency Table for Variable: version

1 unique value(s) detected.

version	N	exactpercent	roundedpercent	cumulpercent
2022-09-07	2197	100	100	100
Total	2197	100	100	100

Frequency Table for Variable: license

1 unique value(s) detected.

license	N	exactpercent	roundedpercent	cumulpercent
Creative Commons Zero 1.0 Universal	2197	100	100	100
Total	2197	100	100	100

23 Visualize Frequency Tables

23.1 Load Tables

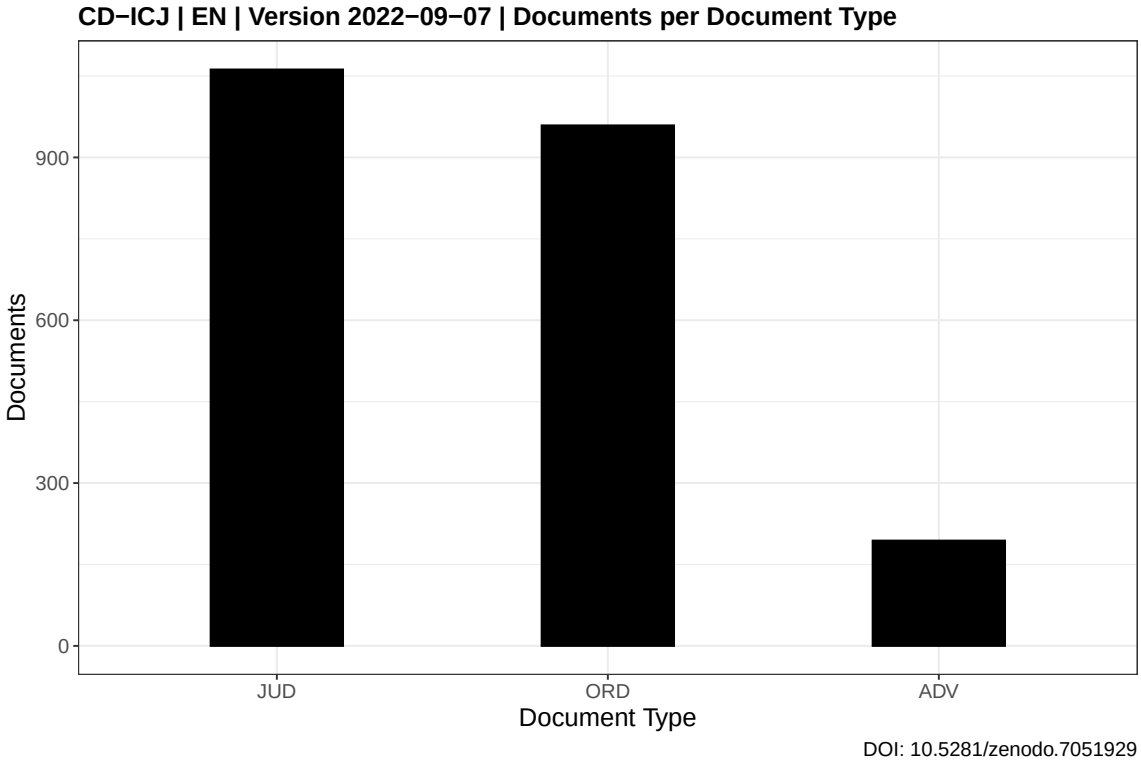
```
prefix.en <- paste0("ANALYSIS/",
                    datashort,
                    "_EN_01_FrequencyTable_var-")

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"))
```

23.2 Doctype

23.2.1 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",
                      datestamp,
                      "| 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)
    )
```

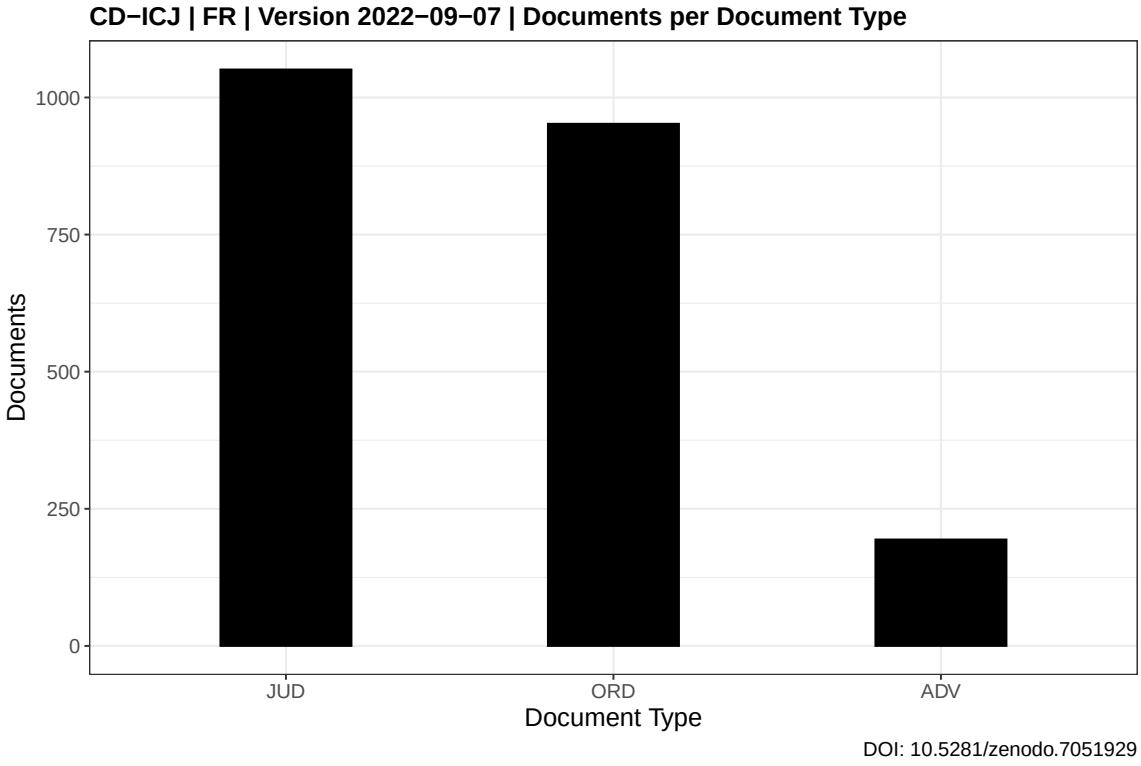


DOI: 10.5281/zenodo.7051929

23.2.2 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",
                      datestamp,
                      "| 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)
    )
```



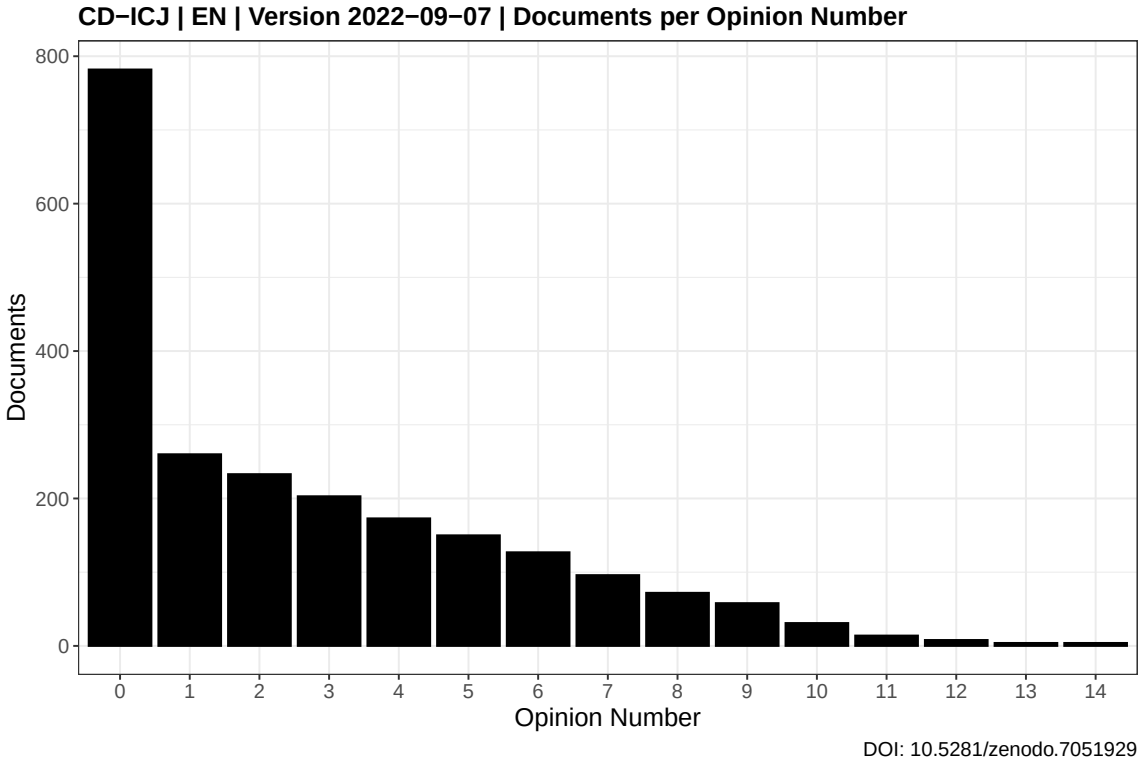
DOI: 10.5281/zenodo.7051929

23.3 Opinion

23.3.1 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",
                      datestamp,
                      "| 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)
    )
```

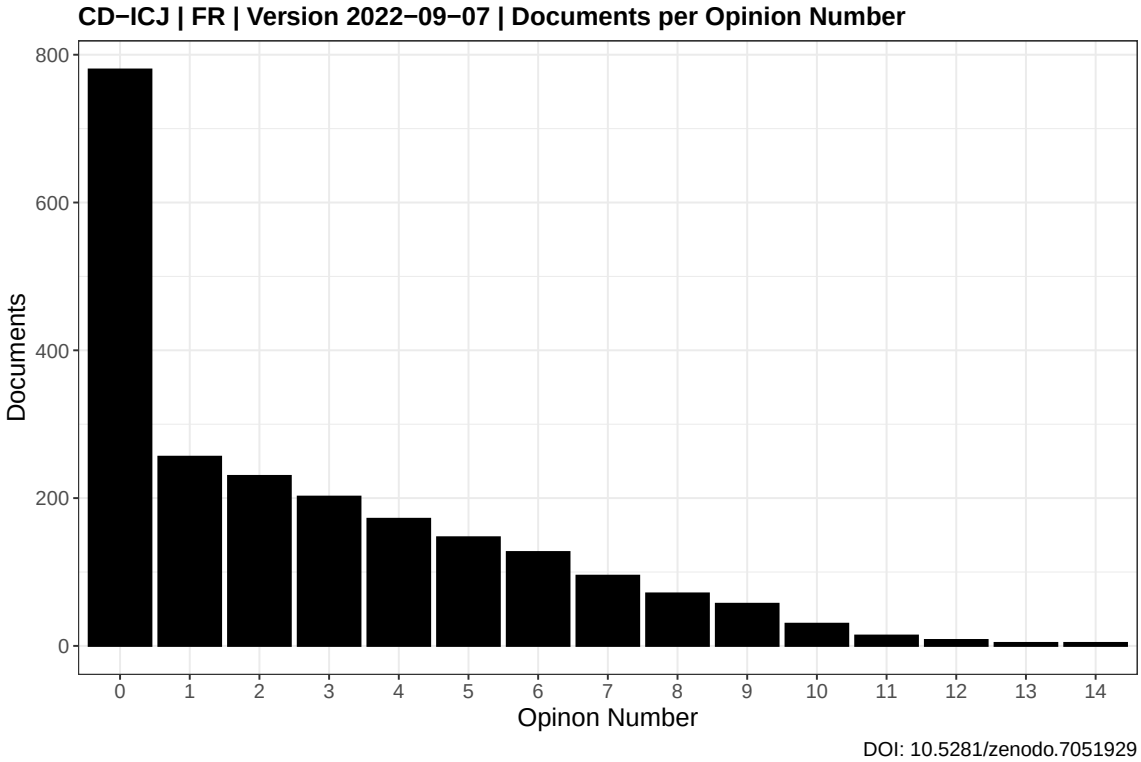


DOI: 10.5281/zenodo.7051929

23.3.2 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",
                      datestamp,
                      "| Documents per Opinion Number"),
        caption = paste("DOI:",
                        doi.version),
        x = "Opinon 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)
    )
```



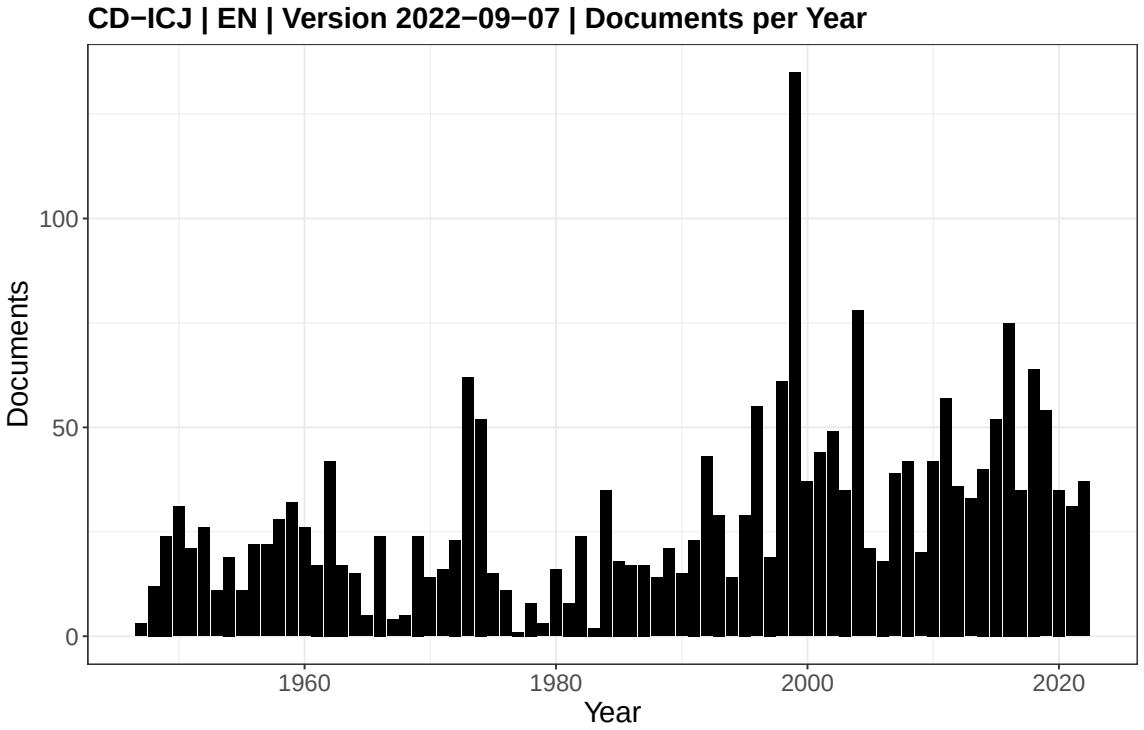
DOI: 10.5281/zenodo.7051929

23.4 Year

23.4.1 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",
                      datestamp,
                      "| 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)
    )
```

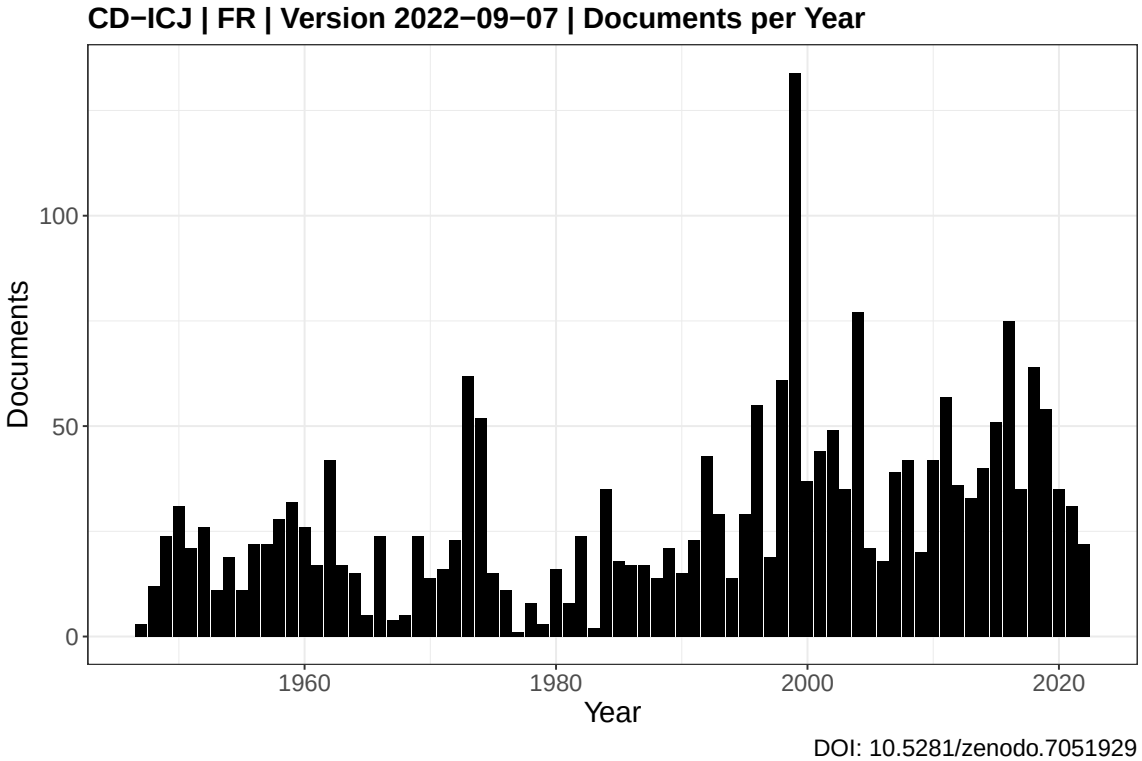


DOI: 10.5281/zenodo.7051929

23.4.2 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",
                      datestamp,
                      "| 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.7051929

24 Summary Statistics

24.1 Linguistic Metrics

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

24.1.1 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)
```

}

24.1.2 Calculate Linguistic Metrics

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

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

```
## [1] "Parallel processing using 16 threads. Begin at 2022-09-07 09:02:48.
    Processing 2215 documents with a total length of 86773117 characters."
```

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

```
## [1] "Runtime was 12.27 secs. Ended at 2022-09-07 09:03:00."
```

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

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

```
## [1] "Parallel processing using 16 threads. Begin at 2022-09-07 09:03:00.
    Processing 2197 documents with a total length of 91759481 characters."
```

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

```
## [1] "Runtime was 14.61 secs. Ended at 2022-09-07 09:03:15."
```

24.1.3 Add Linguistic Metrics to Full Corpora

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

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

24.1.4 Create Metadata-only Variants

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

24.1.5 Calculate Summaries: English

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


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

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

tokens.temp <- tokens(corpus(data.best.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"))
```

24.1.6 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	86,773,117	374	4,474.5	16,322	39,175.2221	45,038.5	744,410
ntokens	15,421,131	71	762.0	2,868	6,962.1359	8,064.5	142,587
ntypes	89,730	53	290.0	712	1,048.6889	1,401.0	9,965
nsentences	523,985	1	20.0	94	236.5621	268.5	5,622

24.1.7 Write Summaries to Disk: English

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

24.1.8 Calculate Summaries: French

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


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


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

tokens.temp <- tokens(corpus(data.best.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"))
```

24.1.9 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	91,759,481	396	4,577	17,125	41,765.8084	47,738	817,704
ntokens	15,775,848	69	785	2,949	7,180.6318	8,450	148,560
ntypes	112,419	55	319	840	1,239.0660	1,679	12,070
nsentences	516,650	1	23	95	235.1616	264	5,555

24.1.10 Write Summaries to Disk: French

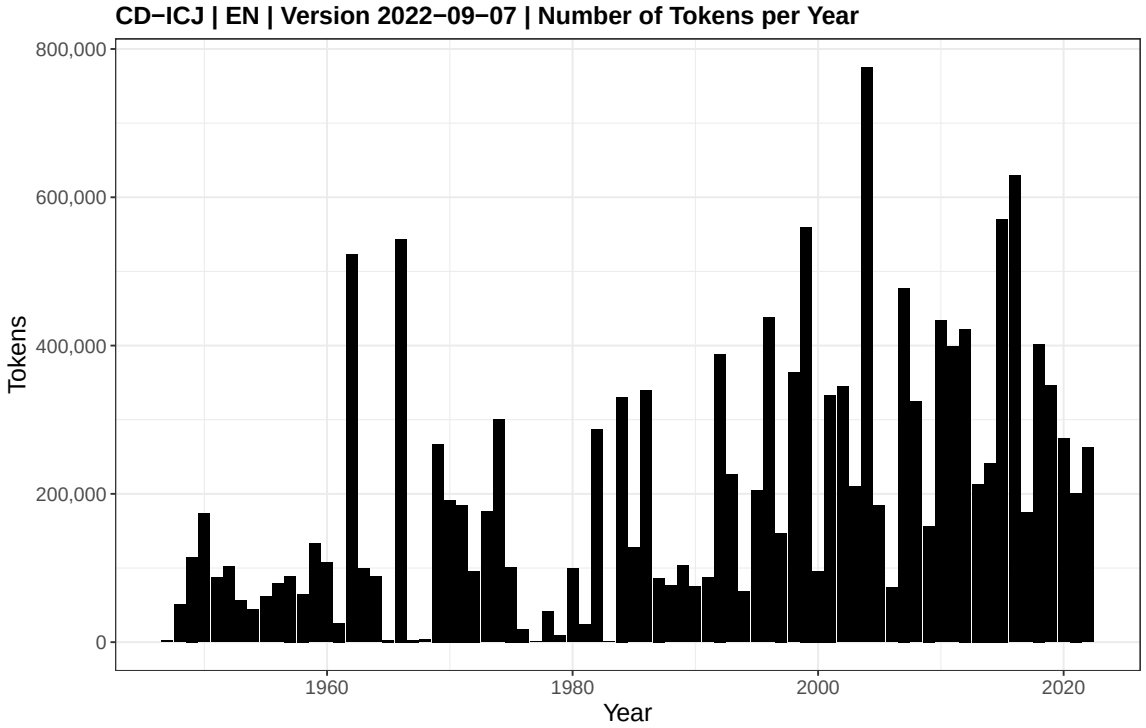
```
fwrite(dt.stats.ling,
       paste0(outputdir,
              datashort,
              "_FR_00_CorpusStatistics_Summaries_Linguistic.csv"),
       na = "NA")
```

24.2 Distributions

24.2.1 Tokens per Year: English

```
tokens.year.en <- meta.best.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",
                      datestamp,
                      "| 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")
    )
)
```

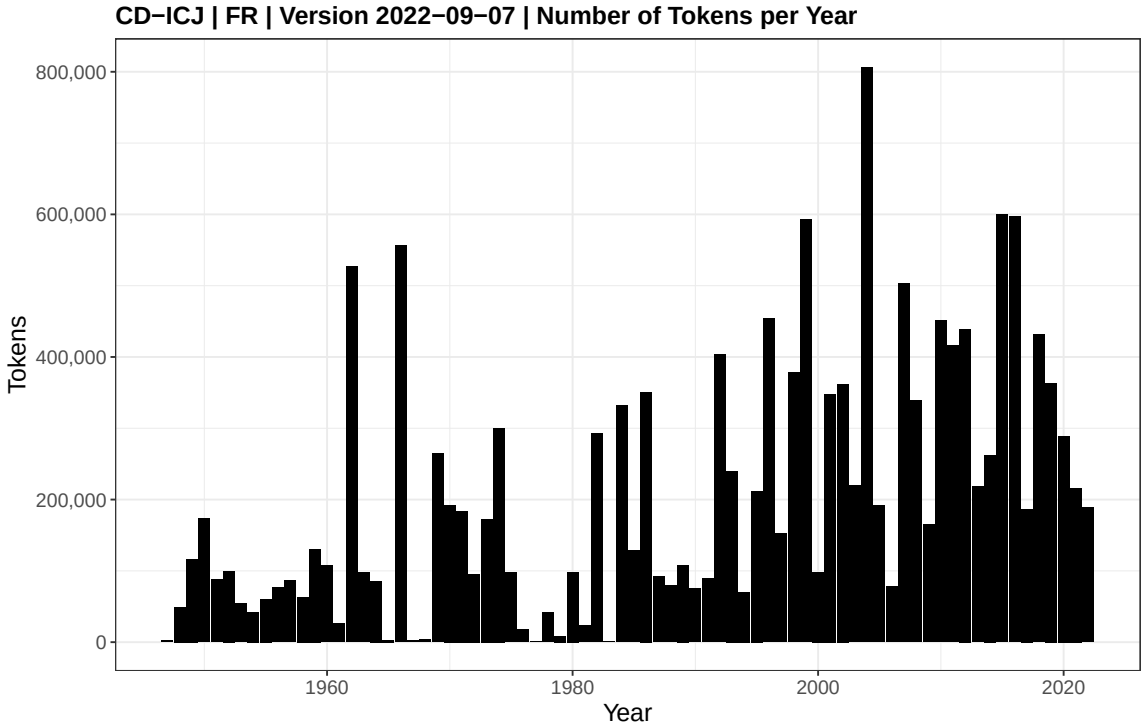


DOI: 10.5281/zenodo.7051929

24.2.2 Tokens per Year: French

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

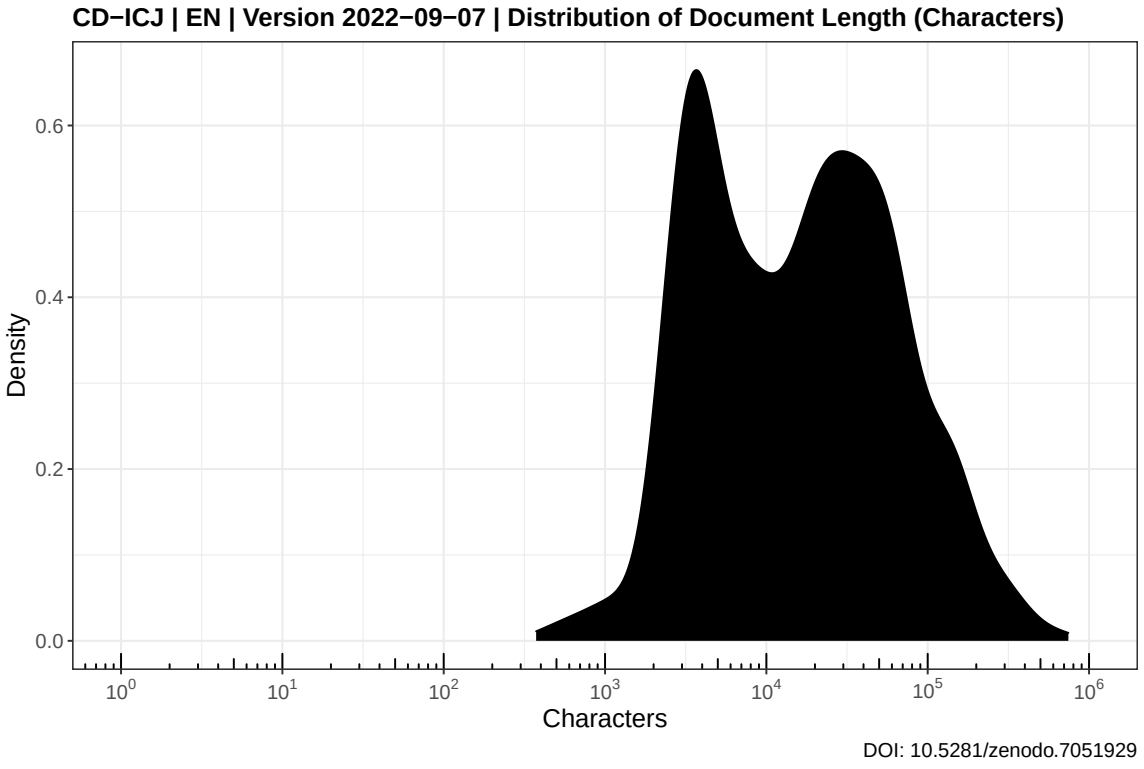
```
print(
    ggplot(data = tokens.year.fr,
           aes(x = year,
               y = V1))+
    geom_bar(stat = "identity",
             fill = "black")+
    scale_y_continuous(labels = comma)+
    theme_bw()+
    labs(
        title = paste(datashort,
                      "| FR | Version",
                      datestamp,
                      "| 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")
    )
)
```



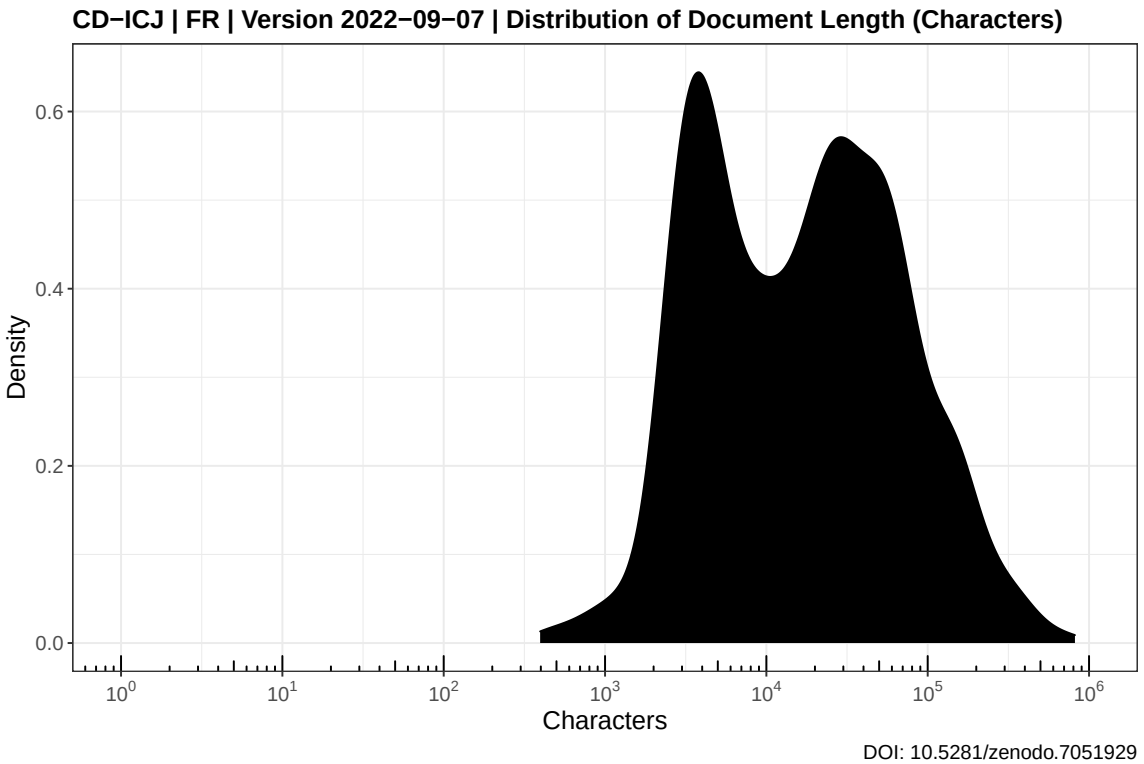
DOI: 10.5281/zenodo.7051929

24.2.3 Density: Characters

```
ggplot(data = meta.best.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",
                      datestamp,
                      "| 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)
    )
```

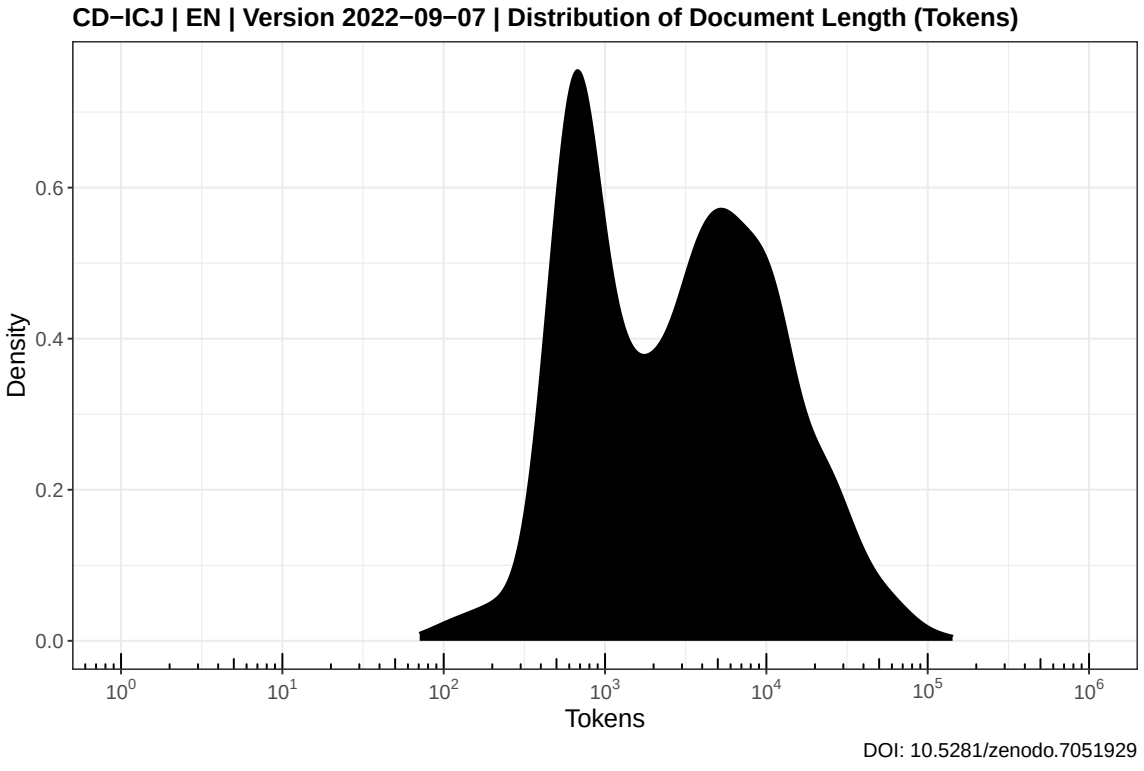


```
ggplot(data = meta.best.fr) +
    geom_density(aes(x = nchars),
                 fill = "black") +
    scale_x_log10(breaks = trans_breaks("log10", function(x) 10^x),
                  labels = trans_format("log10", math_format(10^.x)))+
    annotation_logticks(sides = "b")+
    coord_cartesian(xlim = c(1, 10^6))+
    theme_bw() +
    labs(
        title = paste(datashort,
                      "| FR | Version",
                      datestamp,
                      "| 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)
    )
```

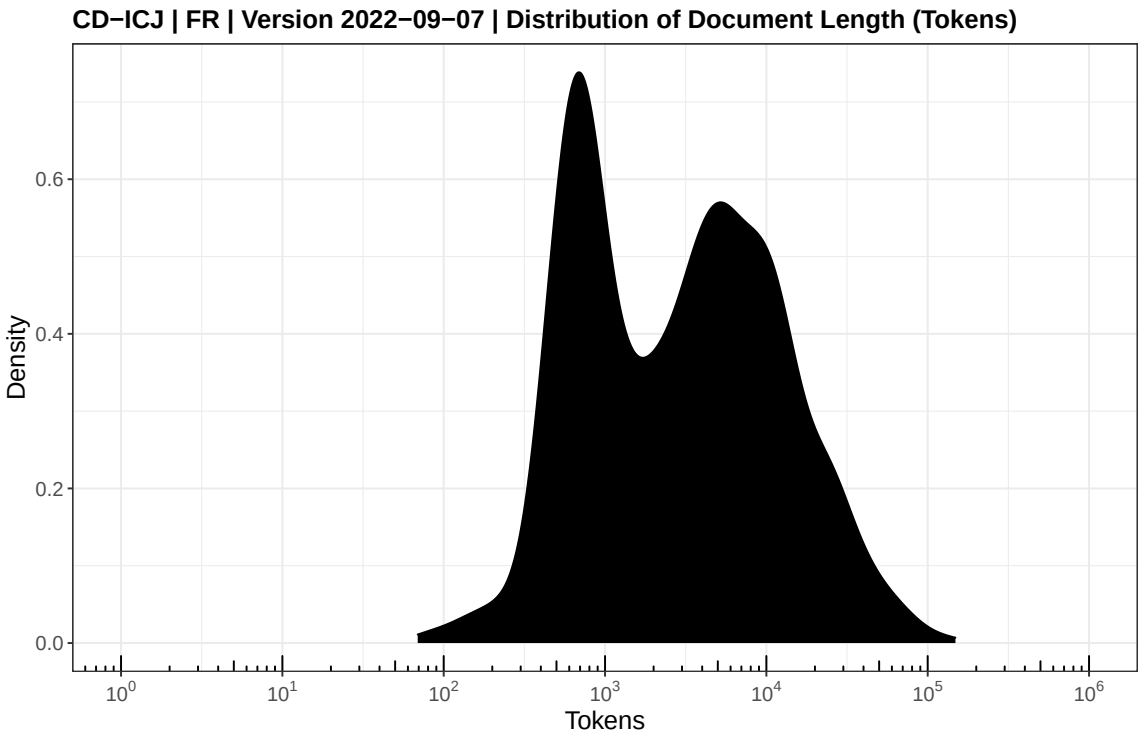


24.2.4 Density: Tokens

```
ggplot(data = meta.best.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",
                      datestamp,
                      "| 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)
    )
```

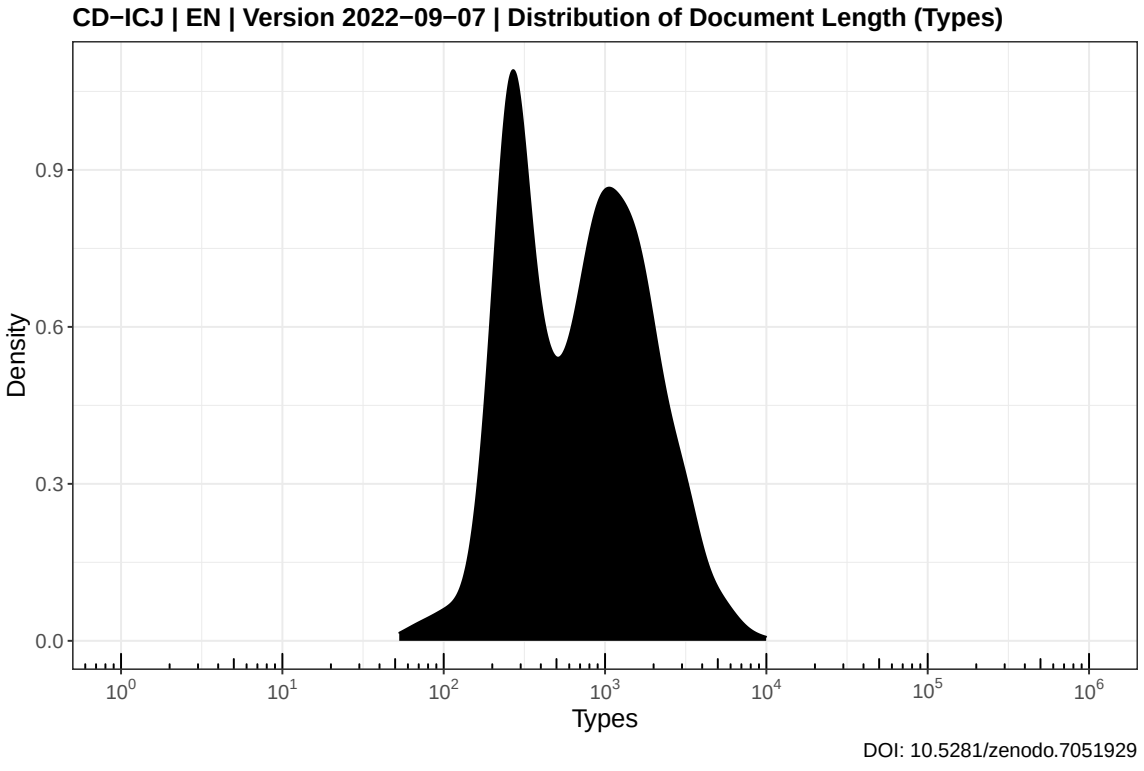


```
ggplot(data = meta.best.fr) +
    geom_density(aes(x = ntokens),
                 fill = "black") +
    scale_x_log10(breaks = trans_breaks("log10", function(x) 10^x),
                  labels = trans_format("log10", math_format(10^.x)))+
    annotation_logticks(sides = "b")+
    coord_cartesian(xlim = c(1, 10^6))+
    theme_bw() +
    labs(
        title = paste(datashort,
                      "| FR | Version",
                      datestamp,
                      "| 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)
    )
```

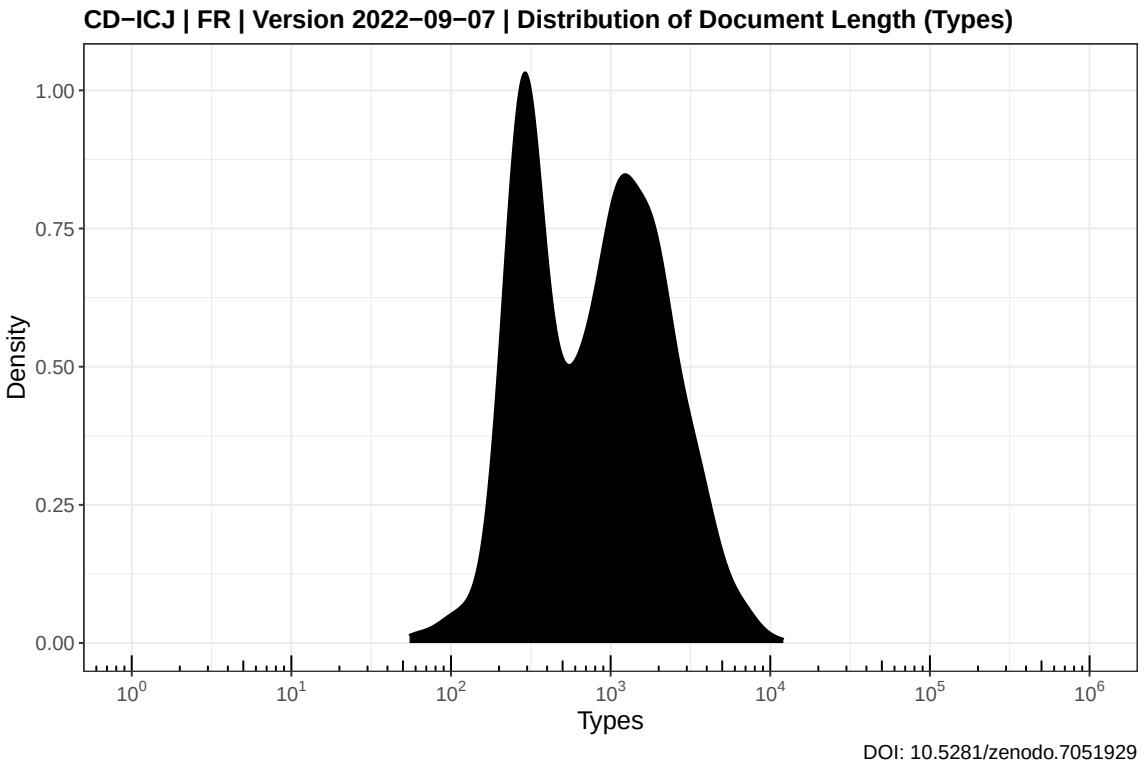


24.2.5 Density: Types

```
ggplot(data = meta.best.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",
                      datestamp,
                      "| 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)
    )
```

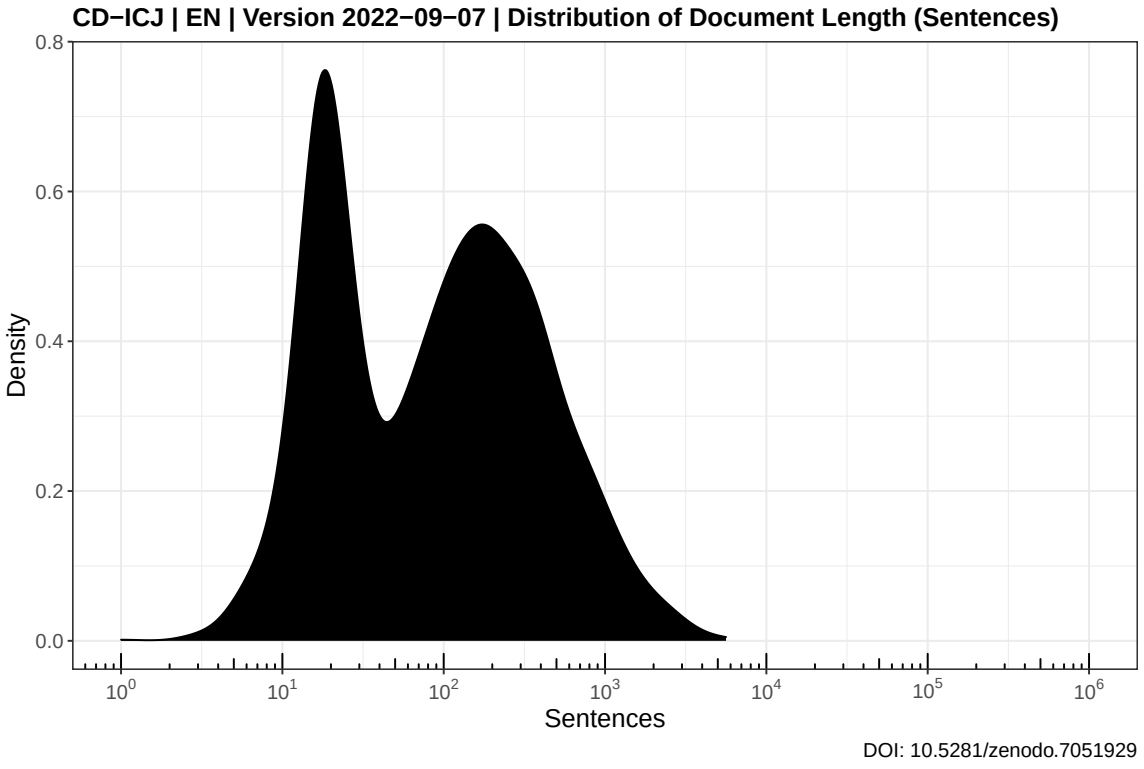


```
ggplot(data = meta.best.fr) +
    geom_density(aes(x = ntypes),
                 fill = "black") +
    scale_x_log10(breaks = trans_breaks("log10", function(x) 10^x),
                  labels = trans_format("log10", math_format(10^.x)))+
    annotation_logticks(sides = "b")+
    coord_cartesian(xlim = c(1, 10^6))+
    theme_bw() +
    labs(
        title = paste(datashort,
                      "| FR | Version",
                      datestamp,
                      "| 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)
    )
```

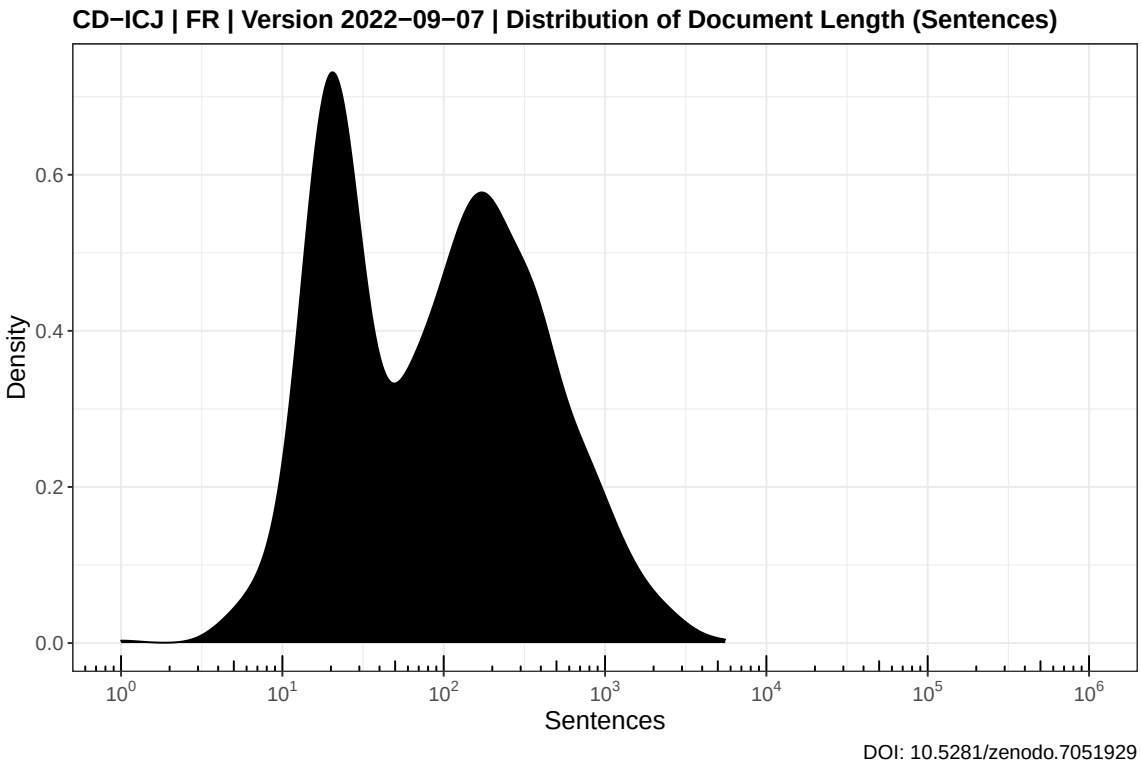


24.2.6 Density: Sentences

```
ggplot(data = meta.best.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",
                      datestamp,
                      "| 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)
    )
```



```
ggplot(data = meta.best.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",
                      datestamp,
                      "| 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)
    )
```



DOI: 10.5281/zenodo.7051929

24.2.7 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)
```

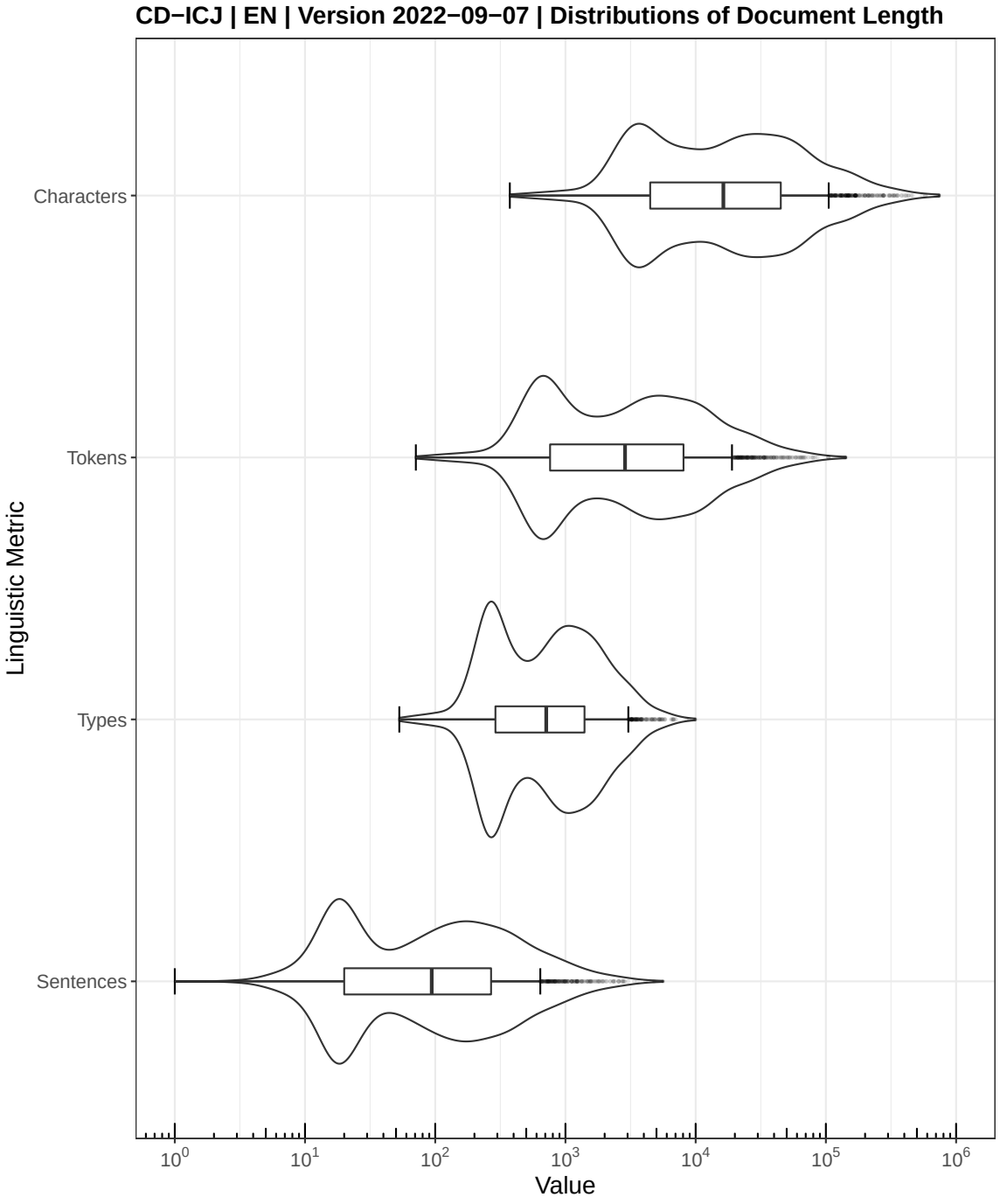
```
## function(x) {
##
##     body = log10(boxplot.stats(10^x)[["stats"]])
##
##     names(body) = c("ymin",
##                     "lower",
##                     "middle",
##                     "upper",
##                     "ymax")
##
##     return(body)
##
## }
```

```
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")))
```

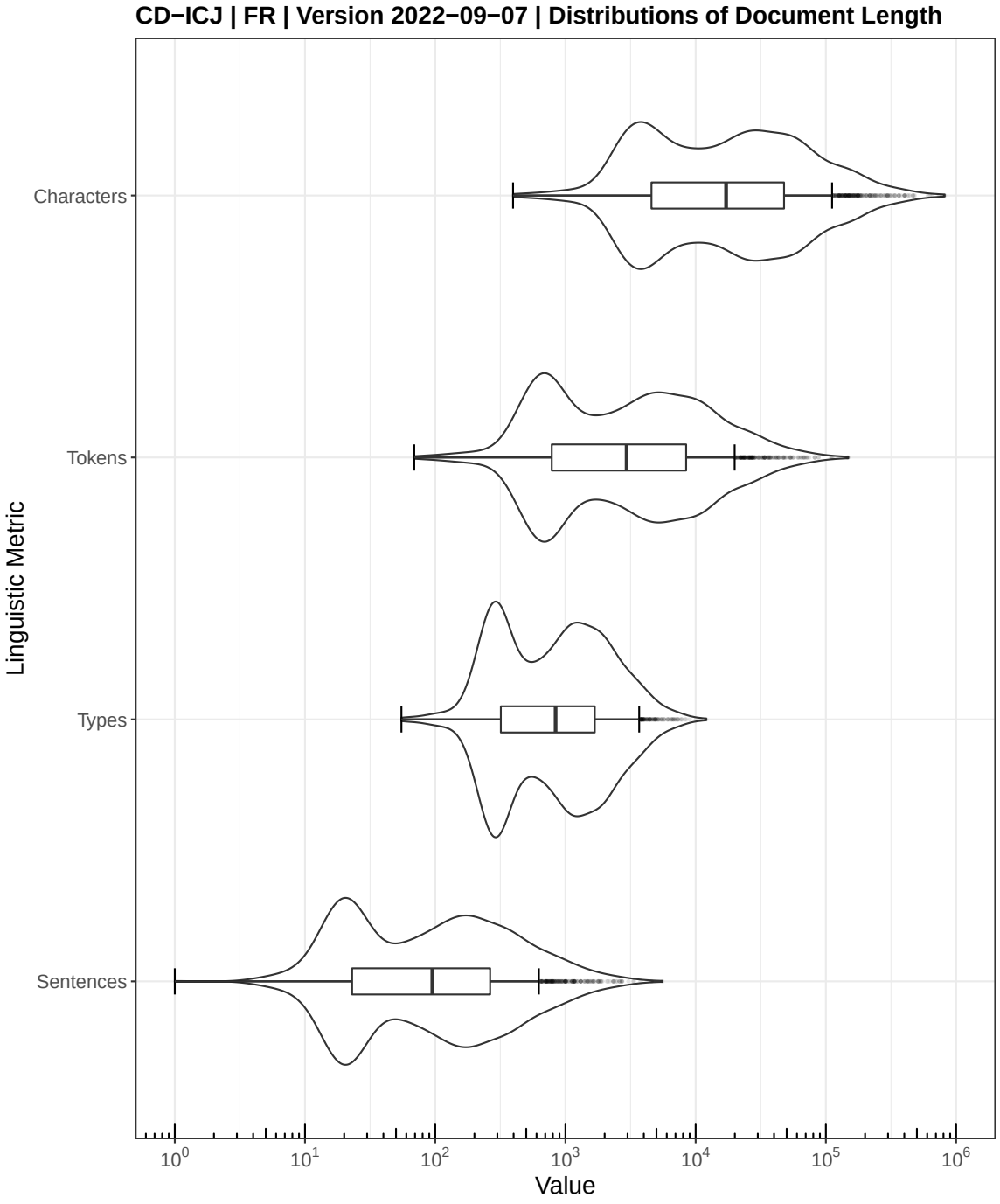
```
ggplot(dt.allmetrics.en, aes(x = value,
                             y = variable))+
    geom_violin()+
    stat_summary(fun.data = f.boxplot.body,
                 geom = "errorbar",
                 width = 0.1) +
    stat_summary(fun.data = f.boxplot.body,
                 geom = "boxplot",
                 width = 0.1) +
    stat_summary(fun.data = f.boxplot.outliers,
                 geom = "point",
                 size =  0.5,
                 alpha = 0.1)+
    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",
                      datestamp,
                      "| 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.7051929

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

```
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 = 0.1)+
    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",
                      datestamp,
                      "| 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.7051929

24.3 Number of Majority Opinions

24.3.1 English

```
dt.maj.disaggregated <- meta.best.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	28
JUD	147
ORD	607
Total	782

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

24.3.2 French

```
dt.maj.disaggregated <- meta.best.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	28
JUD	147
ORD	605
Total	780

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

24.4 Number of Minority Opinions

24.4.1 English

```
dt.min.disaggregated <- meta.best.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	166
JUD	915
ORD	352
Total	1433

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

24.4.2 French

```
dt.min.disaggregated <- meta.best.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	166
JUD	904
ORD	347
Total	1417

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

24.5 Year Range

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1947    1973    1999    1992    2010    2022
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1947    1973    1998    1992    2010    2022
```

24.6 Date Range

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

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

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## "1947-07-31" "1973-09-17" "1999-03-03" "1992-08-27" "2010-11-30" "2022-07-22"
```

```
summary(meta.best.fr$date) # French
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## "1947-07-31" "1973-07-13" "1998-12-04" "1992-06-07" "2010-07-22" "2022-07-22"
```

25 Test and Sort Variable Names

25.1 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.

25.1.1 Sort Variables: Full Data Set

```
setcolorder(data.best.en, # English
            c("doc_id",
              "text",
              "court",
              "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"))
```

```
setcolorder(data.best.fr, # French
            c("doc_id",
              "text",
              "court",
              "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"))
```

25.1.2 Sort Variables: Metadata

```
setcolorder(meta.best.en, # English
            c("doc_id",
              "court",
              "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"))
```

```
setcolorder(meta.best.fr, # French
            c("doc_id",
              "court",
              "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"))
```

25.2 Number of Variables: Full Data Set

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

```
## [1] 28
```

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

```
## [1] 28
```

25.3 Number of Variables: Metadata

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

```
## [1] 27
```

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

```
## [1] 27
```

25.4 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.best.en) # English
```

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

```
names(data.best.fr) # French
```

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

25.5 List All Variables: Metadata

```
names(meta.best.en) # English
```

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

```
names(meta.best.fr) # French
```

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

26 Calculate Detailed Token Frequencies

26.1 Create Corpora

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

26.2 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)
```

26.3 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)
```

26.4 Most Frequent Tokens | TF Weighting | Tables

26.4.1 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	119388	1	2201	all
article	50393	2	2056	all
international	49449	3	2021	all
case	45548	4	2030	all
states	42554	5	1654	all
law	41571	6	1512	all
united	41371	7	1582	all
judgment	36023	8	1533	all
jurisdiction	32054	9	1560	all
state	31856	10	1536	all
parties	30667	11	1869	all
convention	29698	12	1213	all
p	27737	13	1846	all
paragraph	25818	14	1769	all
application	25731	15	1716	all
nations	24652	16	1321	all
may	24645	17	1835	all
dispute	24584	18	1543	all
legal	23105	19	1477	all
one	21410	20	2013	all
question	21261	21	1575	all
general	20907	22	1849	all
rights	20401	23	1285	all
order	20222	24	1986	all
para	20143	25	1158	all
present	19908	26	1829	all
opinion	18407	27	1537	all
treaty	18363	28	1066	all
also	17538	29	1507	all
government	16716	30	1481	all

(continued)

Feature	Frequency	Rank	Docfreq	Group
reports	16662	31	1779	all
republic	16534	32	1310	all
statute	16205	33	1723	all
two	15880	34	1671	all
proceedings	15581	35	1664	all
op	15495	36	851	all
whether	15396	37	1471	all
right	15031	38	1269	all
measures	14529	39	1054	all
agreement	14003	40	1275	all
must	13981	41	1391	all
nicaragua	13935	42	611	all
v	13606	43	1530	all
view	13533	44	1522	all
made	13411	45	1567	all
part	13280	46	1413	all
fact	12698	47	1391	all
court's	12513	48	1474	all
decision	12415	49	1790	all
upon	12367	50	1446	all
justice	12173	51	1653	all
first	11927	52	1451	all
within	11906	53	1618	all
respect	11792	54	1474	all
rules	11720	55	1667	all
territory	11565	56	1004	all
can	11535	57	1372	all
council	11506	58	734	all
force	11319	59	1118	all

(continued)

Feature	Frequency	Rank	Docfreq	Group
interpretation	11234	60	1179	all
time	11158	61	1433	all
however	11054	62	1448	all
basis	11008	63	1371	all
concerning	10928	64	1666	all
genocide	10915	65	474	all
regard	10726	66	1812	all
assembly	10704	67	696	all
declaration	10641	68	1223	all
thus	10579	69	1439	all
obligation	10570	70	1095	all
request	10544	71	1370	all
principle	10542	72	1114	all
line	10402	73	596	all
mr	10356	74	685	all
maritime	10354	75	649	all
use	10324	76	1041	all
party	10255	77	1325	all
delimitation	10168	78	443	all
certain	10099	79	1394	all
claim	10017	80	1085	all
point	9985	81	1229	all
boundary	9806	82	521	all
pp	9760	83	999	all
obligations	9649	84	1041	all
even	9555	85	1323	all
provisional	9496	86	790	all
nuclear	9488	87	389	all
preliminary	9466	88	1065	all

(continued)

Feature	Frequency	Rank	Docfreq	Group
new	9250	89	1102	all
cases	9203	90	1170	all
particular	9134	91	1334	all
continental	9129	92	455	all
shelf	8984	93	432	all
resolution	8934	94	842	all
yugoslavia	8916	95	432	all
provisions	8909	96	1176	all
judge	8886	97	1612	all
effect	8786	98	1251	all
whereas	8750	99	1097	all
without	8658	100	1367	all

26.4.2 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	119675	1	2187	all
droit	56493	2	1651	all
l'article	43361	3	2004	all

(continued)

Feature	Frequency	Rank	Docfreq	Group
comme	38314	4	1869	all
p	36650	5	1865	all
fait	34913	6	2099	all
parties	32400	7	1878	all
être	31585	8	1739	all
d'une	30572	9	1773	all
si	30392	10	1633	all
convention	29839	11	1208	all
entre	29699	12	1682	all
international	28176	13	1842	all
plus	28049	14	1585	all
question	27998	15	1642	all
d'un	27780	16	1799	all
qu'il	27484	17	1722	all
nations	27391	18	1308	all
paragraphe	26305	19	1722	all
etats	23520	20	1405	all
compétence	23331	21	1444	all
différend	21032	22	1456	all
arrêt	21023	23	1258	all
deux	20906	24	1725	all
peut	20180	25	1513	all
statut	20177	26	1821	all
droits	19938	27	1272	all
ainsi	19713	28	1840	all
unies	19419	29	1254	all
tout	18131	30	1558	all
demande	18036	31	1524	all
non	17983	32	1474	all

(continued)

Feature	Frequency	Rank	Docfreq	Group
partie	17488	33	1512	all
dont	17130	34	1950	all
qu'elle	17070	35	1590	all
gouvernement	17065	36	1467	all
république	16522	37	1314	all
traité	16407	38	1046	all
l'affaire	16174	39	1604	all
mesures	16021	40	1141	all
op	15669	41	863	all
internationale	15661	42	1805	all
n'a	15632	43	1498	all
nicaragua	15611	44	599	all
selon	15511	45	1417	all
n'est	15442	46	1474	all
recueil	15253	47	1748	all
cas	15224	48	1433	all
point	14978	49	1395	all
juridique	14725	50	1317	all
doit	13800	51	1570	all
laquelle	13798	52	1790	all
etats-unis	13642	53	905	all
conseil	13636	54	1187	all
procédure	13570	55	1715	all
requête	13563	56	1428	all
bien	12886	57	1433	all
etat	12819	58	1213	all
déclaration	12758	59	1316	all
justice	12604	60	1685	all
aussi	12237	61	1388	all

(continued)

Feature	Frequency	Rank	Docfreq	Group
contre	12227	62	1544	all
générale	12206	63	1120	all
présente	12142	64	1568	all
territoire	11960	65	1006	all
l'arrêt	11926	66	1259	all
décision	11768	67	1393	all
c'est	11720	68	1382	all
où	11547	69	1403	all
voir	11500	70	1177	all
règlement	11405	71	1618	all
général	11061	72	1639	all
l'etat	11008	73	1016	all
principe	10891	74	1144	all
faire	10854	75	1490	all
donc	10739	76	1361	all
savoir	10687	77	1459	all
société	10667	78	535	all
délimitation	10537	79	448	all
autre	10434	80	1326	all
questions	10390	81	1364	all
ni	10340	82	1310	all
devant	10290	83	1337	all
génocide	10158	84	457	all
affaire	10150	85	1652	all
l'organisation	10147	86	891	all
autres	10101	87	1755	all
dispositions	10081	88	1211	all
toute	9979	89	1376	all
l'assemblée	9945	90	686	all

(continued)

Feature	Frequency	Rank	Docfreq	Group
sens	9913	91	1294	all
avis	9744	92	1683	all
ligne	9678	93	507	all
tant	9447	94	1238	all
titre	9392	95	1218	all
fond	9327	96	1308	all
termes	9288	97	1314	all
mandat	9244	98	351	all
obligations	9207	99	1037	all
yougoslavie	9131	100	409	all

26.5 Most Frequent Tokens | TFIDF Weighting | Tables

26.5.1 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
nicaragua	7794.299	1	611	all
convention	7766.411	2	1213	all
genocide	7308.634	3	474	all

(continued)

Feature	Weight	Rank	Docfreq	Group
nuclear	7167.464	4	389	all
delimitation	7107.127	5	443	all
league	7088.020	6	317	all
mandate	6960.937	7	254	all
law	6893.384	8	1512	all
op	6437.307	9	851	all
diss	6431.822	10	376	all
shelf	6377.652	11	432	all
yugoslavia	6329.379	12	432	all
continental	6274.931	13	455	all
boundary	6163.424	14	521	all
costa	6093.700	15	249	all
united	6047.085	16	1582	all
line	5930.466	17	596	all
treaty	5832.392	18	1066	all
judgment	5757.613	19	1533	all
para	5673.582	20	1158	all
nations	5533.657	21	1321	all
maritime	5520.018	22	649	all
council	5519.171	23	734	all
africa	5517.099	24	440	all
states	5397.474	25	1654	all
assembly	5381.591	26	696	all
mr	5278.279	27	685	all
islands	5137.690	28	287	all
weapons	5097.432	29	244	all
state	5064.547	30	1536	all
south	4959.496	31	559	all
jurisdiction	4880.194	32	1560	all

(continued)

Feature	Weight	Rank	Docfreq	Group
rights	4824.237	33	1285	all
sep	4730.678	34	492	all
rica	4704.905	35	242	all
measures	4686.084	36	1054	all
human	4577.435	37	582	all
qatar	4566.231	38	129	all
federal	4489.099	39	565	all
river	4443.410	40	271	all
el	4286.106	41	229	all
provisional	4251.802	42	790	all
mandates	4224.899	43	110	all
legal	4066.324	44	1477	all
sea	4051.036	45	617	all
territory	3974.197	46	1004	all
indb	3900.694	47	309	all
bahrain	3897.829	48	97	all
honduras	3889.074	49	260	all
dispute	3859.880	50	1543	all
area	3829.942	51	663	all
charter	3820.268	52	799	all
equidistance	3797.835	53	106	all
drc	3794.246	54	60	all
chamber	3784.528	55	268	all
republic	3771.446	56	1310	all
salvador	3765.279	57	143	all
resolution	3752.831	58	842	all
serbia	3728.239	59	294	all
right	3636.181	60	1269	all
cerd	3582.633	61	97	all

(continued)

Feature	Weight	Rank	Docfreq	Group
territorial	3554.403	62	719	all
tribunal	3509.188	63	608	all
security	3415.524	64	799	all
respondent	3388.912	65	768	all
use	3385.477	66	1041	all
pp	3375.088	67	999	all
force	3360.972	68	1118	all
agreement	3358.809	69	1275	all
colombia	3353.221	70	262	all
disarmament	3290.579	71	98	all
obligation	3233.993	72	1095	all
fry	3215.376	73	64	all
coast	3168.218	74	250	all
obligations	3164.129	75	1041	all
question	3148.609	76	1575	all
principle	3146.666	77	1114	all
cançado	3129.866	78	196	all
trindade	3118.043	79	197	all
claim	3104.709	80	1085	all
zone	3080.339	81	297	all
interpretation	3076.542	82	1179	all
ibid	3059.865	83	795	all
member	3057.939	84	719	all
trusteeship	3033.357	85	107	all
evidence	3027.757	86	851	all
nicaragua's	3011.905	87	202	all
preliminary	3010.416	88	1065	all
west	3009.968	89	465	all
kingdom	3003.075	90	848	all

(continued)

Feature	Weight	Rank	Docfreq	Group
committee	2986.401	91	570	all
treaties	2981.377	92	786	all
island	2981.003	93	260	all
also	2933.414	94	1507	all
government	2922.269	95	1481	all
commission	2921.868	96	693	all
opinion	2921.188	97	1537	all
protection	2888.953	98	819	all
congo	2872.384	99	286	all
bosnia	2869.607	100	267	all

26.5.2 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
nicaragua	8810.899	1	599	all
convention	7751.072	2	1208	all
mandat	7363.058	3	351	all
délimitation	7276.347	4	448	all

(continued)

Feature	Weight	Rank	Docfreq	Group
droit	7009.820	5	1651	all
génocide	6926.881	6	457	all
yougoslavie	6666.605	7	409	all
société	6543.951	8	535	all
plateau	6371.713	9	430	all
op	6358.782	10	863	all
diss	6287.724	11	377	all
nations	6169.064	12	1308	all
ligne	6163.164	13	507	all
continental	6128.540	14	439	all
frontière	5475.480	15	544	all
traité	5287.949	16	1046	all
nucléaires	5264.797	17	386	all
etats-unis	5254.646	18	905	all
rica	5162.764	19	233	all
costa	5096.556	20	239	all
arrêt	5090.707	21	1258	all
qatar	5043.423	22	128	all
l'assemblée	5027.257	23	686	all
zone	5019.014	24	503	all
droits	4732.144	25	1272	all
unies	4729.158	26	1254	all
mer	4679.950	27	618	all
ind	4588.263	28	468	all
etats	4566.496	29	1405	all
mesures	4558.686	30	1141	all
conservatoires	4487.181	31	657	all
maritime	4430.696	32	616	all
compétence	4252.375	33	1444	all

(continued)

Feature	Weight	Rank	Docfreq	Group
mandats	4222.769	34	125	all
bahreïn	4103.369	35	94	all
salvador	4074.420	36	143	all
honduras	4072.390	37	257	all
résolution	4072.186	38	713	all
territoire	4057.216	39	1006	all
plus	3977.370	40	1585	all
l'organisation	3977.141	41	891	all
armes	3949.308	42	272	all
si	3915.823	43	1633	all
rdc	3886.026	44	58	all
indb	3883.274	45	308	all
îles	3861.861	46	240	all
charte	3778.178	47	787	all
africain	3774.646	48	362	all
différend	3757.760	49	1456	all
tribunal	3710.209	50	696	all
république	3688.284	51	1314	all
l'etat	3686.979	52	1016	all
conseil	3645.985	53	1187	all
fédérale	3624.888	54	460	all
colombie	3606.426	55	258	all
tutelle	3576.398	56	161	all
générale	3571.622	57	1120	all
question	3540.540	58	1642	all
commission	3530.294	59	779	all
entre	3445.205	60	1682	all
défendeur	3433.140	61	749	all
rfy	3414.096	62	79	all

(continued)

Feature	Weight	Rank	Docfreq	Group
mandataire	3341.099	63	111	all
l'homme	3332.620	64	436	all
etat	3306.908	65	1213	all
côte	3276.604	66	232	all
juridique	3272.547	67	1317	all
peut	3268.981	68	1513	all
sécurité	3251.164	69	831	all
être	3206.840	70	1739	all
milles	3176.405	71	206	all
traités	3125.622	72	826	all
ciedr	3122.941	73	83	all
voir	3117.116	74	1177	all
non	3117.040	75	1474	all
l'ouganda	3106.284	76	46	all
fleuve	3103.607	77	192	all
principe	3086.553	78	1144	all
sud-ouest	3085.670	79	399	all
membres	3079.771	80	900	all
pacte	3015.041	81	345	all
juridiction	3014.674	82	987	all
obligations	3001.954	83	1037	all
gouvernement	2993.200	84	1467	all
ibid	2985.251	85	776	all
l'emploi	2983.982	86	577	all
cameroun	2981.343	87	362	all
libye	2976.161	88	153	all
congo	2962.460	89	277	all
cançado	2958.658	90	194	all
point	2954.498	91	1395	all

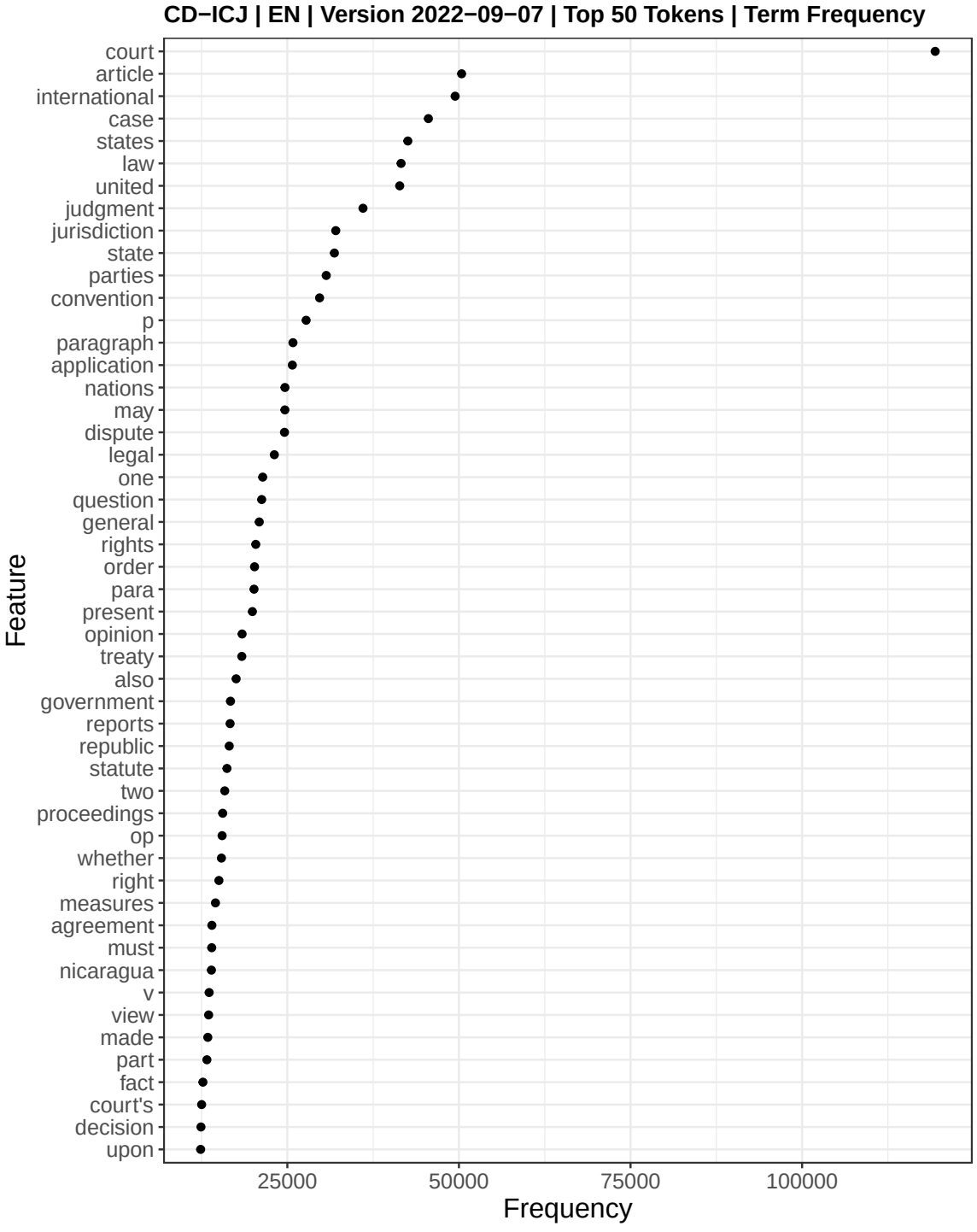
(continued)

Feature	Weight	Rank	Docfreq	Group
selon	2954.228	92	1417	all
trindade	2946.009	93	194	all
royaume-uni	2934.037	94	689	all
membre	2928.575	95	675	all
force	2914.320	96	874	all
qu'il	2907.722	97	1722	all
bosnie-herzégovine	2886.178	98	252	all
l'arrêt	2883.758	99	1259	all
demande	2864.930	100	1524	all

26.6 Most Frequent Tokens | TF Weighting | Scatterplots

26.6.1 English

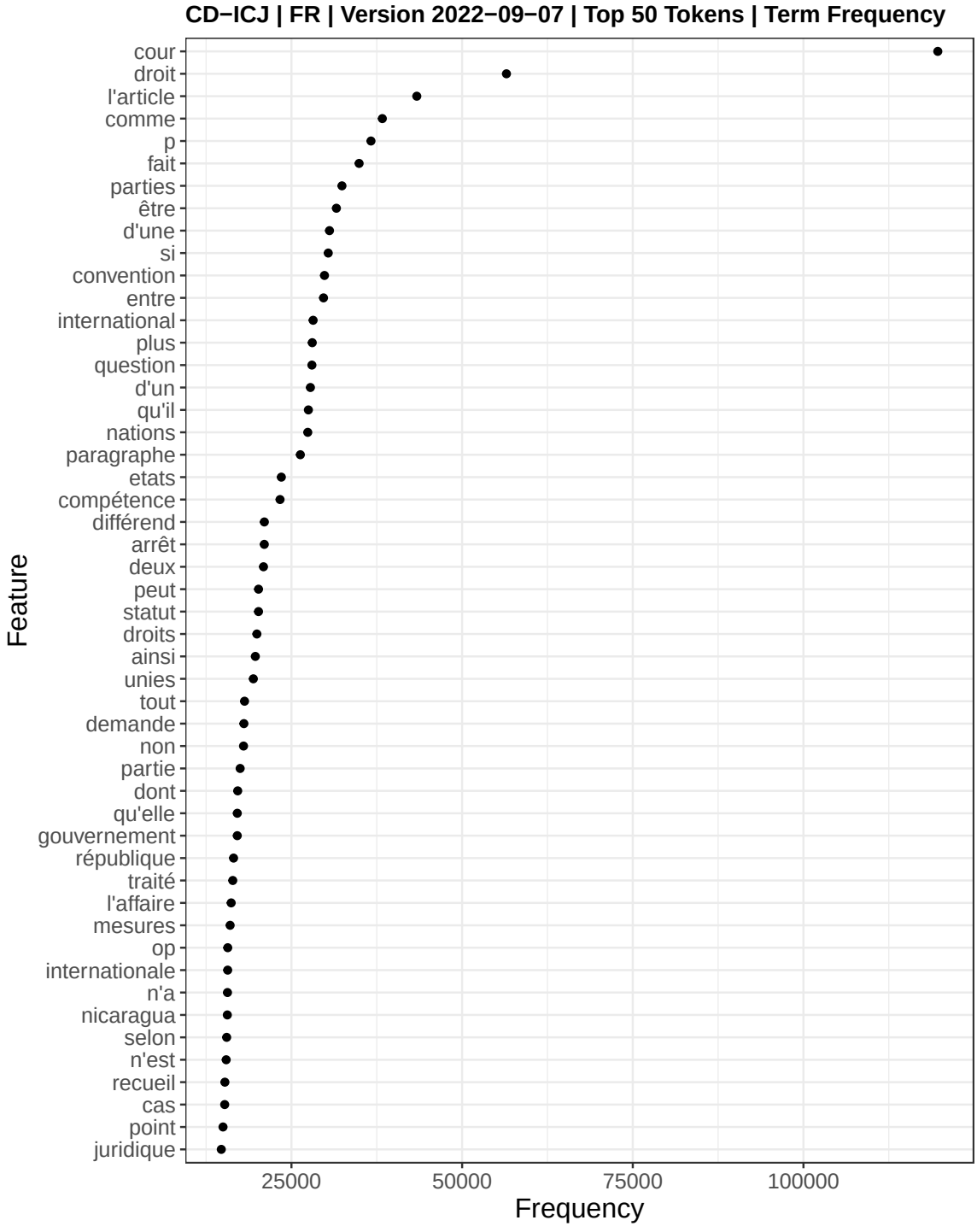
```
print(
    ggplot(data = tstat.en[1:50, ],
           aes(x = reorder(feature,
                           frequency),
               y = frequency))+
    geom_point()+
    coord_flip()+
    labs(
        title = paste(datashort,
                      "| EN | Version",
                      datestamp,
                      "| 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.7051929

26.6.2 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",
                      datestamp,
                      "| 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")
    )
)
```

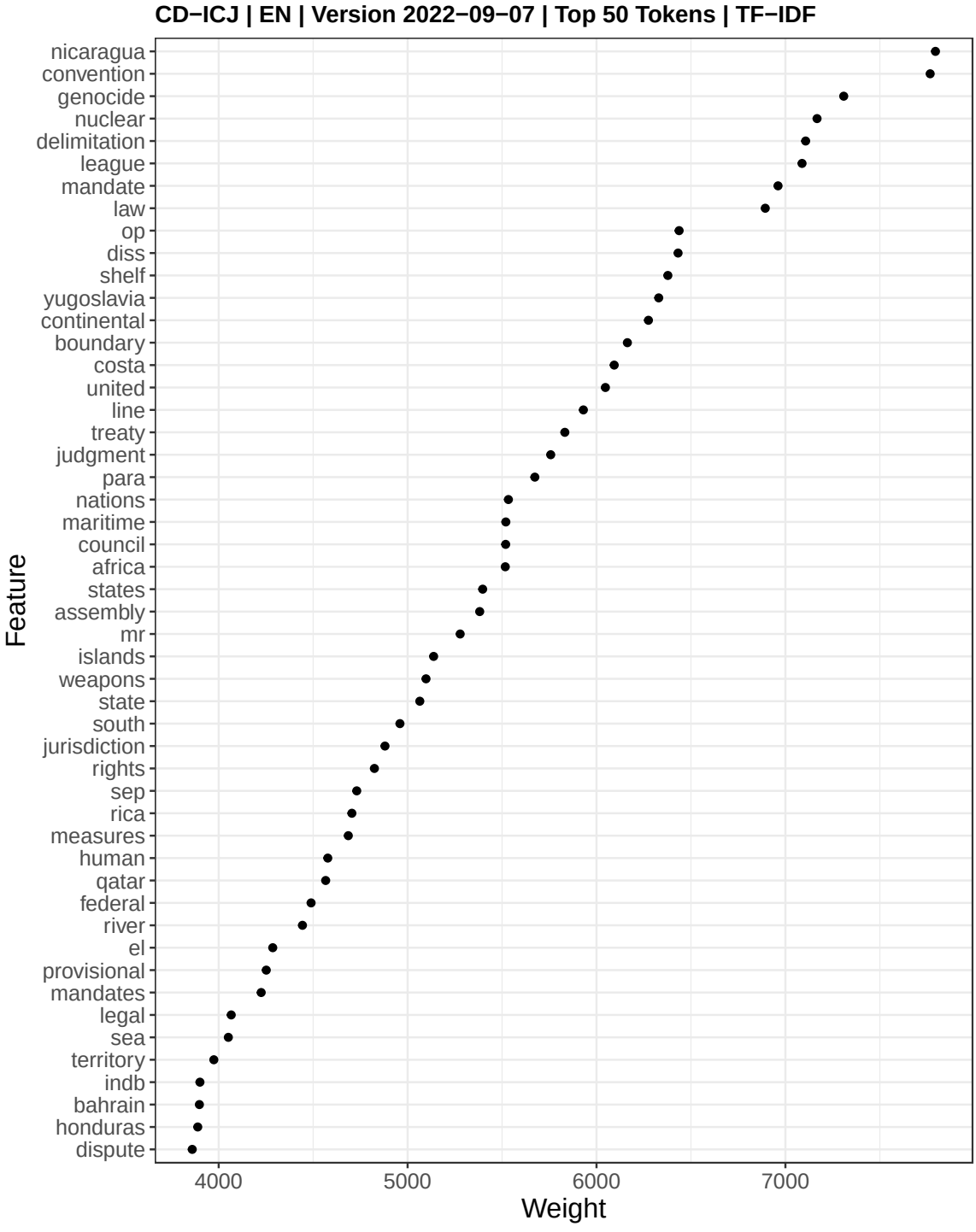


DOI: 10.5281/zenodo.7051929

26.7 Most Frequent Tokens | TFIDF Weighting | Scatterplots

26.7.1 English

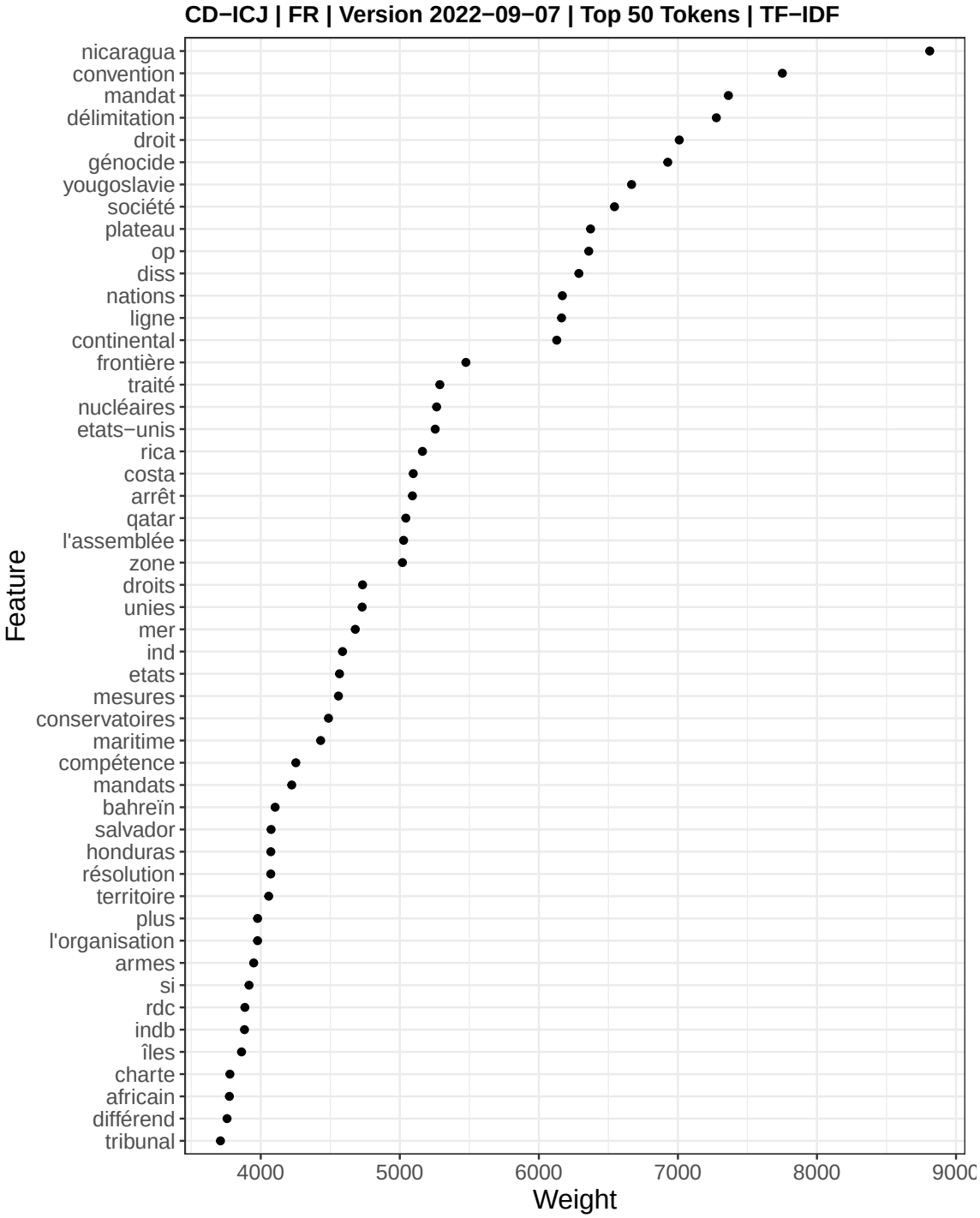
```
print(
    ggplot(data = tstat.tfidf.en[1:50, ],
           aes(x = reorder(feature,
                           frequency),
               y = frequency))+
    geom_point()+
    coord_flip()+
    theme_bw()+
    labs(
        title = paste(datashort,
                      "| EN | Version",
                      datestamp,
                      "| 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.7051929

26.7.2 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",
                      datestamp,
                      "| 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")
    )
)
```

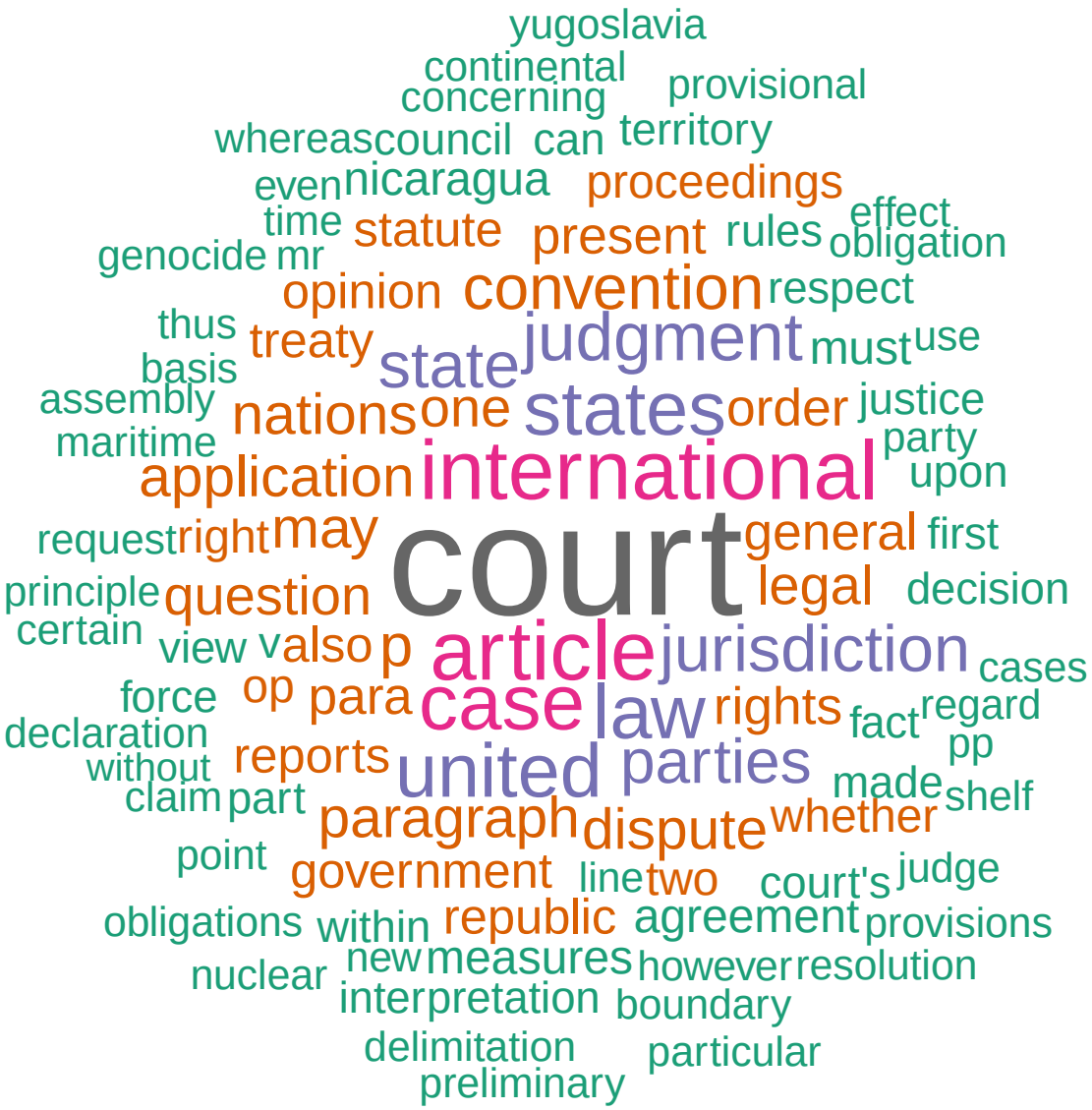


DOI: 10.5281/zenodo.7051929

26.8 Most Frequent Tokens | TF Weighting | Wordclouds

26.8.1 English

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



26.8.2 French

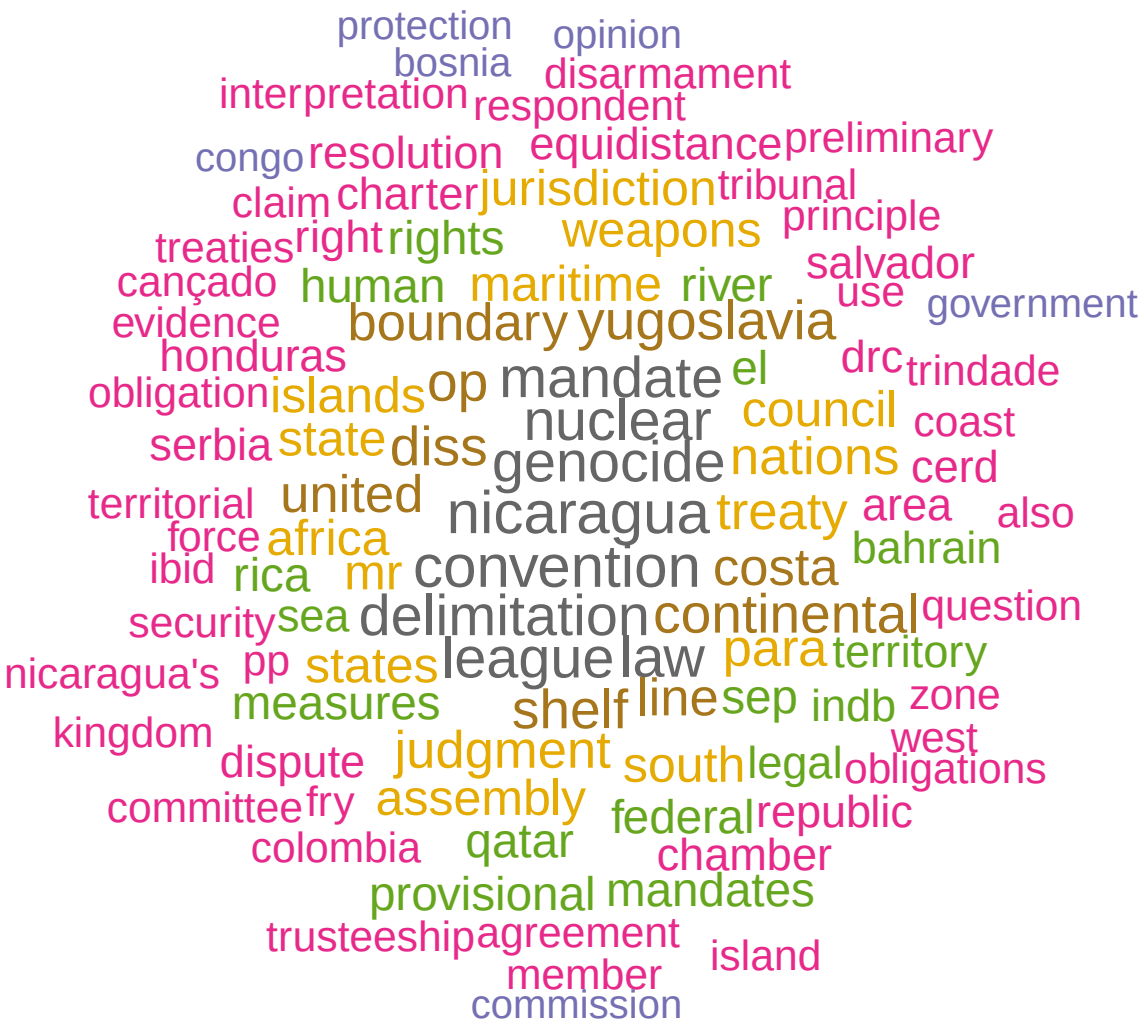
```
textplot_wordcloud(dfm.fr,
                   max_words = 100,
                   min_size = 1,
                   max_size = 5,
                   random_order = FALSE,
                   rotation = 0,
                   color = brewer.pal(8, "Dark2"))
```

26.9 Most Frequent Tokens | TFIDF Weighting | Wordclouds

26.9.1 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"))
```

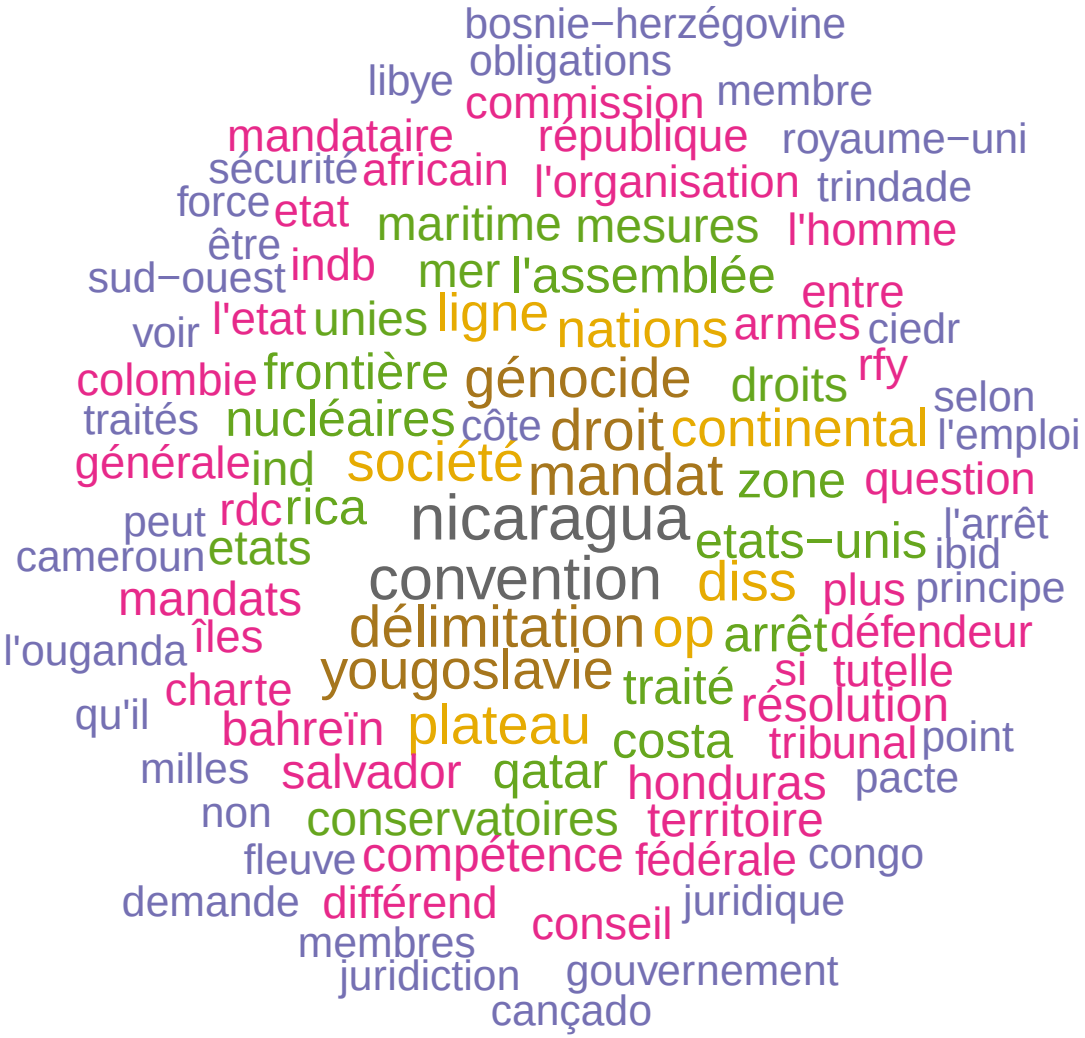
```
## Warning in dfm_trim.dfm(x, min_termfreq = min_count): dfm has been previously
## weighted
```



26.9.2 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
```



27 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.

27.1 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 <- 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
```

27.2 English

27.2.1 Calculate Similarity

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

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

27.2.2 Create Empty Lists

```
list.ndrop <- vector("list",
                     threshold.N)

list.drop.ids <- vector("list",
                        threshold.N)

list.pair.ids <- vector("list",
                        threshold.N)
```

27.2.3 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)
```

27.2.4 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"))
}
```

27.2.5 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"))
}
```

27.2.6 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")
```

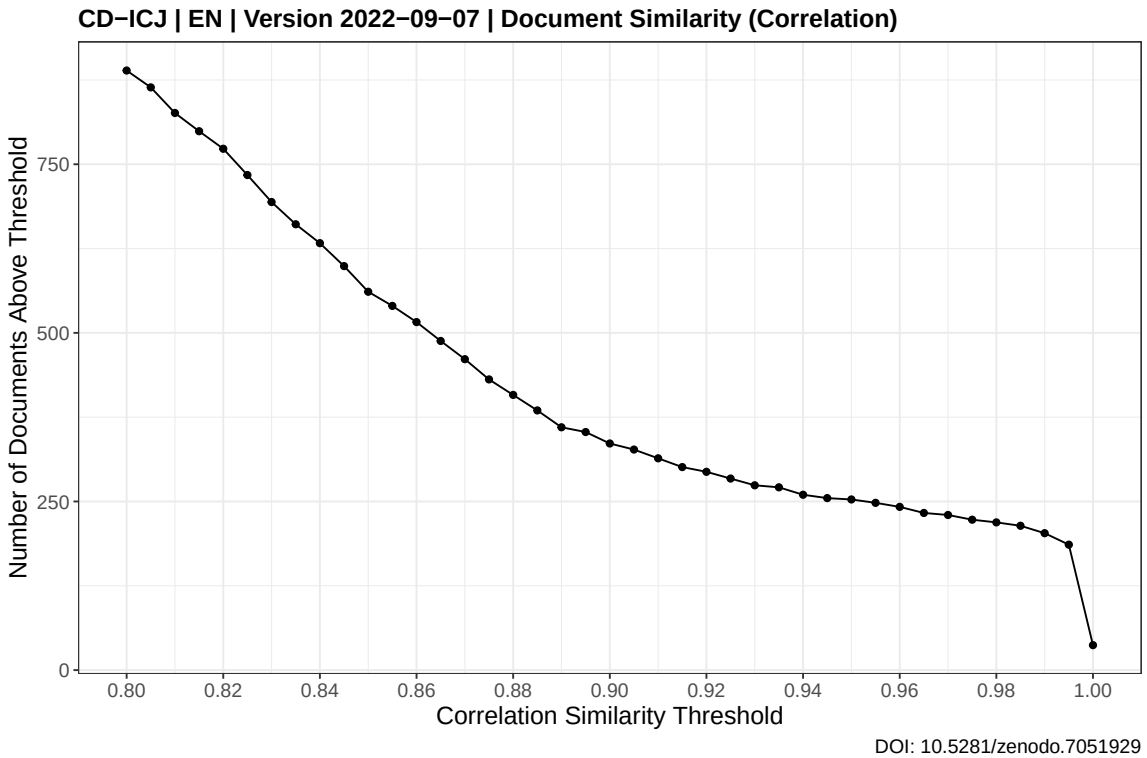
Threshold	Number to Drop
0.800	889
0.805	864
0.810	826
0.815	799
0.820	773
0.825	734
0.830	694
0.835	661
0.840	633
0.845	599
0.850	561
0.855	540
0.860	516
0.865	488
0.870	461
0.875	431
0.880	408
0.885	385
0.890	360
0.895	353
0.900	336
0.905	327
0.910	314
0.915	301
0.920	294
0.925	284
0.930	274
0.935	271
0.940	260
0.945	255

(continued)

Threshold	Number to Drop
0.950	253
0.955	248
0.960	242
0.965	233
0.970	230
0.975	223
0.980	219
0.985	214
0.990	203
0.995	186
1.000	37

```
fwrite(dt.ndrop,
       paste0(outputdir,
              datashort,
              "_EN_18_DocumentSimilarity_Correlation_Table.csv"))
```

```
print(
    ggplot(data = dt.ndrop,
           aes(x = threshold,
               y = ndrop))+
    geom_line()+
    geom_point()+
    labs(
        title = paste(datashort,
                      "| EN | Version",
                      datestamp,
                      "| 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"
    )
)
```



27.3 French

27.3.1 Calculate Similarity

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

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

27.3.2 Create Empty Lists

```
list.ndrop <- vector("list",
                     threshold.N)

list.drop.ids <- vector("list",
                        threshold.N)

list.pair.ids <- vector("list",
                        threshold.N)
```

27.3.3 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)
```

27.3.4 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"))
}
```

27.3.5 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"))

}
```

27.3.6 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")
```

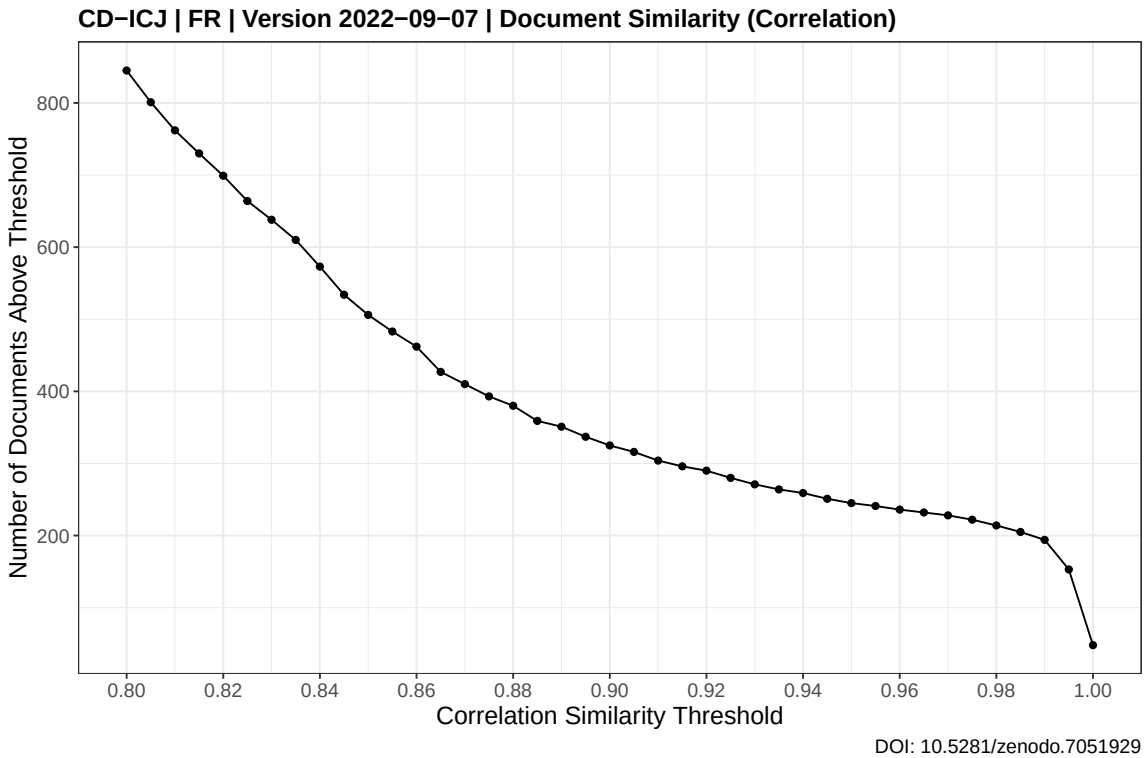
Threshold	Number to Drop
0.800	845
0.805	801
0.810	762
0.815	730
0.820	699
0.825	664
0.830	638
0.835	610
0.840	573
0.845	534
0.850	506
0.855	483
0.860	462
0.865	427
0.870	410
0.875	393
0.880	380
0.885	359
0.890	351
0.895	337
0.900	325
0.905	316
0.910	304
0.915	296
0.920	290
0.925	280
0.930	271
0.935	264
0.940	259
0.945	251

(continued)

Threshold	Number to Drop
0.950	245
0.955	241
0.960	236
0.965	232
0.970	228
0.975	222
0.980	214
0.985	205
0.990	194
0.995	153
1.000	48

```
fwrite(dt.ndrop,
       paste0(outputdir,
              datashort,
              "_FR_18_DocumentSimilarity_Correlation_Table.csv"))
```

```
print(
    ggplot(data = dt.ndrop,
           aes(x = threshold,
               y = ndrop))+
    geom_line()+
    geom_point()+
    labs(
        title = paste(datashort,
                      "| FR | Version",
                      datestamp,
                      "| 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"
    )
)
```



28 Create CSV Files

28.1 Full Data Set

```
csvname.full.en <- paste(datashort,
                         datestamp,
                         "EN_CSV_BEST_FULL.csv",
                         sep = "_")

csvname.full.fr <- paste(datashort,
                         datestamp,
                         "FR_CSV_BEST_FULL.csv",
                         sep = "_")


fwrite(data.best.en,
       csvname.full.en,
       na = "NA")

fwrite(data.best.fr,
       csvname.full.fr,
       na = "NA")
```

28.2 Metadata Only

These files are the same as the full data set, minus the "text" variable.

```
csvname.meta.en <- paste(datashort,
                     datestamp,
                     "EN_CSV_BEST_META.csv",
                     sep = "_")

csvname.meta.fr <- paste(datashort,
                     datestamp,
                     "FR_CSV_BEST_META.csv",
                     sep = "_")


fwrite(meta.best.en,
       csvname.meta.en,
       na = "NA")

fwrite(meta.best.fr,
       csvname.meta.fr,
       na = "NA")
```

29 Final File Count per Folder

```
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
EN_PDF_ORIGINAL_FULL	2215
FR_PDF_ORIGINAL_FULL	2197
EN_PDF_ENHANCED_max2004	1484
FR_PDF_ENHANCED_max2004	1482
EN_PDF_BEST_FULL	2215
FR_PDF_BEST_FULL	2197
EN_PDF_BEST_MajorityOpinions	782
FR_PDF_BEST_MajorityOpinions	780
EN_TXT_BEST_FULL	2215
FR_TXT_BEST_FULL	2197
EN_TXT_TESSERACT_max2004	1484
FR_TXT_TESSERACT_max2004	1482
EN_TXT_EXTRACTED_FULL	2215
FR_TXT_EXTRACTED_FULL	2197

30 File Size Distribution

30.1 English

30.1.1 Corpus Object in RAM

```
print(object.size(data.best.en),
      humanReadable = TRUE,
      units = "MB")
```

```
## 82.1 Mb
```

30.1.2 Create Data Table of Filenames

```
best <- list.files("EN_PDF_BEST_FULL",
                   full.names = TRUE)

original <- list.files("EN_PDF_ORIGINAL_FULL",
                       full.names = TRUE)

MB <- file.size(best) / 10^6

dt1 <- data.table(MB,
                  rep("BEST",
                      length(MB)))


MB <- file.size(original) / 10^6

dt2 <- data.table(MB, rep("ORIGINAL",
                          length(MB)))


dt <- rbind(dt1,
            dt2)

setnames(dt,
         "V2",
         "variant")
```

30.1.3 Total Size Comparison

```
kable(dt[,
         .(MB_total = sum(MB)),
         keyby = variant],
      format = "latex",
      align = "r",
      booktabs = TRUE,
      longtable = TRUE)
```

variant	MB_total
BEST	2749.033
ORIGINAL	1362.009

30.1.4 Analyze Files Larger than 10 MB

```
# Summarize
summary(dt[MB > 10]$MB)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.02   11.75   13.69   17.16   18.60   63.10
```

```
# Space required by large files

kable(dt[MB > 10,
         .(total = sum(MB)),
         keyby = variant],
      format = "latex",
      align = "r",
      booktabs = TRUE,
      longtable = TRUE)
```

variant	total
BEST	638.3742
ORIGINAL	151.1618

```
# Show Individual Large File Sizes

kable(dt[MB > 10][order(MB)],
      format = "latex",
      align = "r",
      booktabs = TRUE,
      longtable = TRUE)
```

MB	variant
10.01953	ORIGINAL
10.03045	ORIGINAL

10.17802	BEST
10.18249	BEST
10.27246	BEST
10.75333	ORIGINAL
10.96772	BEST
10.96772	ORIGINAL
10.96773	BEST
10.96773	ORIGINAL
11.61704	ORIGINAL
11.74503	BEST
11.74503	BEST
11.84531	ORIGINAL
11.87470	BEST
12.03525	BEST
12.35245	BEST
12.74963	BEST
12.75576	BEST
13.06405	BEST
13.12759	BEST
13.12759	ORIGINAL
13.29246	BEST
14.08516	BEST
14.25152	BEST
14.63674	ORIGINAL
14.90139	BEST
15.05370	ORIGINAL
15.49849	BEST
15.49975	BEST
15.91648	ORIGINAL
16.22107	BEST
16.22615	ORIGINAL

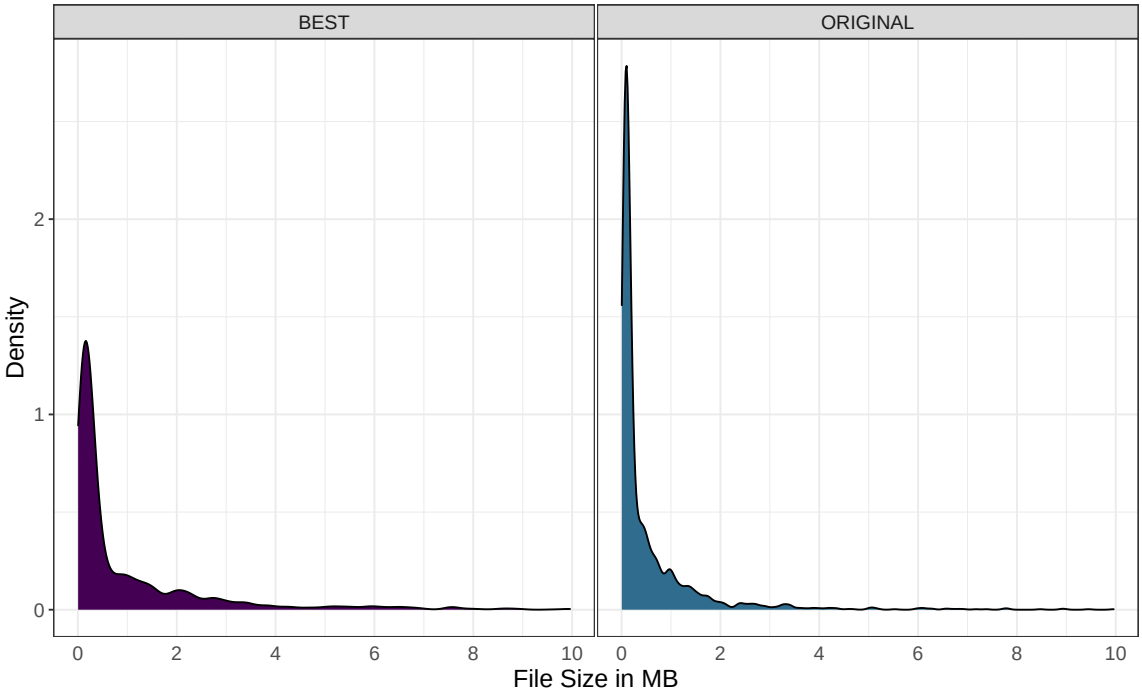
18.04857	BEST
18.77929	BEST
19.32415	BEST
19.79244	BEST
19.79421	BEST
20.80212	BEST
24.11553	BEST
24.76594	BEST
31.08607	BEST
34.08729	BEST
34.24754	BEST
42.66818	BEST
63.09513	BEST

30.1.5 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",
                      datestamp,
                      "| 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−ICJ | EN | Version 2022−09−07 | Distribution of File Sizes up to 10 MB



DOI: 10.5281/zenodo.7051929

30.2 French

30.2.1 Corpus Object in RAM

```
print(object.size(data.best.en),
      humanReadable = TRUE,
      units = "MB")
```

```
## 82.1 Mb
```

30.2.2 Create Data Table of filenames

```
best <- list.files("FR_PDF_BEST_FULL",
                   full.names = TRUE)

original <- list.files("FR_PDF_ORIGINAL_FULL",
                       full.names = TRUE)


MB <- file.size(best) / 10^6

dt1 <- data.table(MB,
                  rep("BEST",
                      length(MB)))



MB <- file.size(original) / 10^6

dt2 <- data.table(MB,
                  rep("ORIGINAL",
                      length(MB)))

dt <- rbind(dt1,
            dt2)

setnames(dt,
         "V2",
         "variant")
```

30.2.3 Total Size Comparison

```
kable(dt[,
         .(MB_total = sum(MB)),
         keyby = variant],
      format = "latex",
      align = "r",
      booktabs = TRUE,
      longtable = TRUE)
```

variant	MB_total
BEST	2960.974
ORIGINAL	1427.689

30.2.4 Analyze Files Larger than 10 MB

```
summary(dt[MB > 10]$MB)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.17   11.35   14.97   17.92   19.90   69.14
```

```
# Space required by large files

kable(dt[MB > 10,
         .(total = sum(MB)),
         keyby = variant],
      format = "latex",
      align = "r",
      booktabs = TRUE,
      longtable = TRUE)
```

variant	total
BEST	837.7203
ORIGINAL	166.0686

```
# Show Individual Large File Sizes

kable(dt[MB > 10][order(MB)],
      format = "latex",
      align = "r",
      booktabs = TRUE,
      longtable = TRUE)
```

MB	variant
10.16579	ORIGINAL
10.17264	ORIGINAL
10.27243	BEST

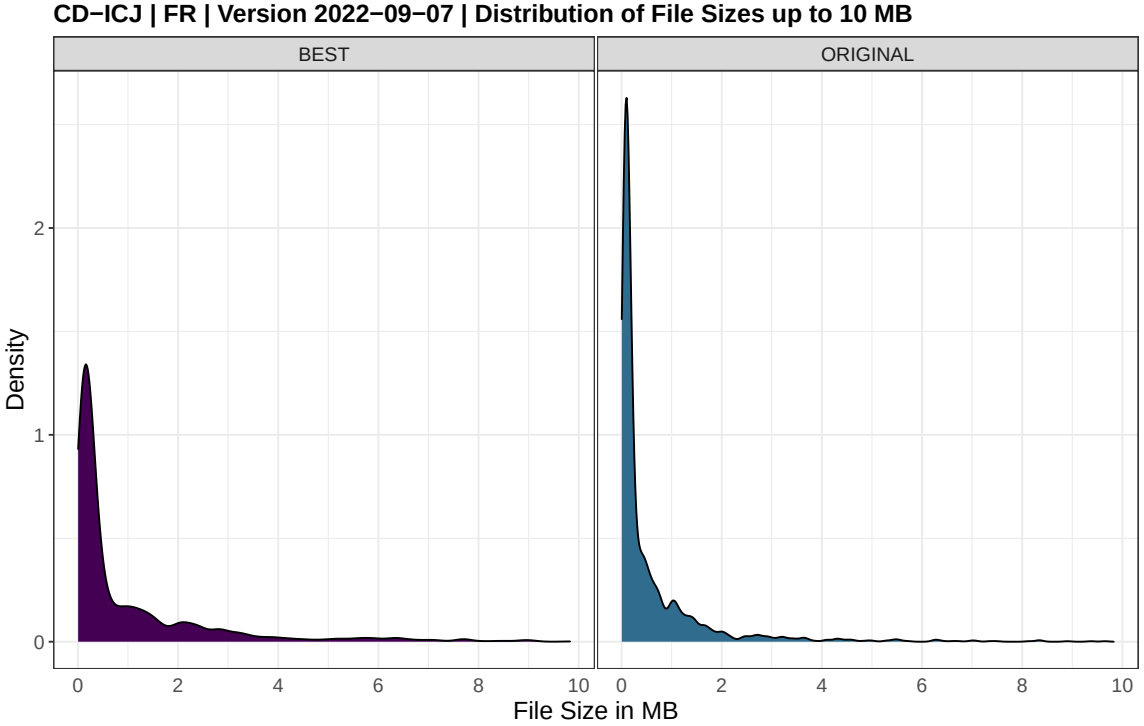
10.37758	BEST
10.37758	BEST
10.62218	ORIGINAL
10.62229	ORIGINAL
10.65307	BEST
10.66941	BEST
11.00723	ORIGINAL
11.11991	BEST
11.32830	BEST
11.32830	ORIGINAL
11.34529	BEST
11.34529	ORIGINAL
11.79755	BEST
11.91450	ORIGINAL
12.17423	BEST
12.17423	BEST
12.81937	BEST
13.01561	BEST
13.06957	ORIGINAL
13.22076	BEST
13.51862	BEST
13.51862	BEST
13.66332	BEST
13.92358	BEST
14.76233	ORIGINAL
15.18048	BEST
15.30020	BEST
15.69768	BEST
16.01056	ORIGINAL
16.44028	BEST
16.44029	BEST

16.69984	BEST
17.43610	ORIGINAL
17.47657	BEST
17.61179	ORIGINAL
18.52545	BEST
18.61401	BEST
19.10715	BEST
19.86228	BEST
20.01589	BEST
20.71804	BEST
20.71804	BEST
21.13851	BEST
21.41228	BEST
23.48641	BEST
23.85617	BEST
23.85633	BEST
26.41957	BEST
36.38981	BEST
36.89126	BEST
44.06520	BEST
44.30212	BEST
69.13705	BEST

30.2.5 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",
                      datestamp,
                      "| 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()
    )
)
```



DOI: 10.5281/zenodo.7051929

31 Create ZIP Archives

31.1 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)
```

31.2 ZIP Data Directories

Note: Vector of Directories was created at the beginning of the script.

```
for (dir in dirset){
    zip(paste(datashort,
              datestamp,
              dir,
              sep = "_"),
        dir)
}
```

31.3 ZIP ANALYSIS Directory

```
zip(paste(datashort,
          datestamp,
          "EN-FR",
          basename(outputdir),
          sep = "_"),
    basename(outputdir))
```

31.4 ZIP Unlabelled Files Directory

```
zip(dir.unlabelled,
    dir.unlabelled)
```

31.5 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,
          datestamp,
          "Source_Files.zip",
          sep = "_"),
    files.source)
```

32 Delete CSV and Directories

The metadata CSV files are retained for Codebook generation.

32.1 Delete CSVs

```
unlink(csvname.full.fr)
unlink(csvname.full.en)
unlink(csvname.meta.fr)
unlink(csvname.meta.en)
```

32.2 Delete Data Directories

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

unlink(dir.unlabelled,
       recursive = TRUE)
```

33 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.

33.1 Create Set of ZIP Archives

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

33.2 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)
```

}

33.3 Compute Hashes

```
multihashes <- f.dopar.multihashes(files.zip)
```

```
## [1] "Parallel processing using 16 threads."
## [1] "Processed 21 files. Runtime was 13.41 secs."
```

33.4 Convert to Data Table

```
setDT(multihashes)
```

33.5 Add Index

```
multihashes$index <- seq_len(multihashes[,.N])
```

33.6 Save to Disk

```
fwrite(multihashes,
       paste(datashort,
             datestamp,
             "CryptographicHashes.csv",
             sep = "_"),
       na = "NA")
```

33.7 Add Whitespace to Enable Automatic Linebreak

This is only used for display and will be discarded after printing to the Compilation Report.

```
multihashes$sha3.512 <- paste(substr(multihashes$sha3.512, 1, 64),
                              substr(multihashes$sha3.512, 65, 128))
```

33.8 Print to Report

```
kable(multihashes[,.(index,filename)],
      format = "latex",
      align = c("p{1cm}",
                "p{13cm}"),
      booktabs = TRUE,
      longtable = TRUE)
```

index	filename
1	CD-ICJ_2022-09-07_EN_CSV_BEST_FULL.zip
2	CD-ICJ_2022-09-07_EN_CSV_BEST_META.zip
3	CD-ICJ_2022-09-07_EN_PDF_BEST_FULL.zip
4	CD-ICJ_2022-09-07_EN_PDF_BEST_MajorityOpinions.zip
5	CD-ICJ_2022-09-07_EN_PDF_ENHANCED_max2004.zip
6	CD-ICJ_2022-09-07_EN_PDF_ORIGINAL_FULL.zip
7	CD-ICJ_2022-09-07_EN_TXT_BEST_FULL.zip
8	CD-ICJ_2022-09-07_EN_TXT_EXTRACTED_FULL.zip
9	CD-ICJ_2022-09-07_EN_TXT_TESSERACT_max2004.zip
10	CD-ICJ_2022-09-07_EN-FR_ANALYSIS.zip
11	CD-ICJ_2022-09-07_FR_CSV_BEST_FULL.zip
12	CD-ICJ_2022-09-07_FR_CSV_BEST_META.zip
13	CD-ICJ_2022-09-07_FR_PDF_BEST_FULL.zip
14	CD-ICJ_2022-09-07_FR_PDF_BEST_MajorityOpinions.zip
15	CD-ICJ_2022-09-07_FR_PDF_ENHANCED_max2004.zip
16	CD-ICJ_2022-09-07_FR_PDF_ORIGINAL_FULL.zip
17	CD-ICJ_2022-09-07_FR_TXT_BEST_FULL.zip
18	CD-ICJ_2022-09-07_FR_TXT_EXTRACTED_FULL.zip
19	CD-ICJ_2022-09-07_FR_TXT_TESSERACT_max2004.zip
20	CD-ICJ_2022-09-07_Source_Files.zip
21	CD-ICJ_2022-09-07_UnlabelledFiles.zip

```
kable(multihashes[,.(index,sha2.256)],
      format = "latex",
      align = c("c",
                "p{13cm}"),
      booktabs = TRUE,
      longtable = TRUE)
```

index	sha2.256
1	132275f5de8301e3fe8b5d7a139f9915c339a7b17c267d40fb9d2640a2e5ac03
2	39040d749106eb4a48c4ffc7140c1f07ebcd7fc1ff3dc6cb7635187766e8f3a6
3	506d330e365b223135d45ce3441934d5480cf9f8ac29e3a3433ec5c859a73b70
4	cb8d47784d5b785bc62ea818079448cdbc8dc6e40d8769d27e411eccbc68bb89
5	ca285e899a7c419e1c28fa409a7427a32d18b3f23cc8ee8c73164a7411c1826f
6	3d7e1821e265ab8d064b6c327f1d38e7f35dabaf3419bd2317c9e63810d8d752
7	03c7d41a1ad52f56ac406d5a66c27d5c33d900c8376c1a1f4f4ad8db0502628e
8	d444af5cebb43407f340d03081ddd9cbe953cd115d62a2fade9e0f26c7882aca
9	90016ed7d3895d6c178074d629ecb070280f2df092460fd375e354b17a515352
10	52302980597421ccc21c5d6fbce46dc7df9d5afa0383b980742e1f1f6c8f6072
11	0d018cbd3486c0e89fa2e1c467efd0d045df2fbf7bf2636958b919db614003fb
12	a432c93dfc8dd6fcf3edf12e15c3da652174da3a8fb95de17466cbcf9066dd13
13	466a5717c0374d865d619f970e48ffa0c7c7633d496f564e27ef4e4f75577fa3
14	ad7b4c068d2b9ac25ceddb46d12a2b11eba897ba940d54a7f99a45a71f5aa992
15	875f5c2ae772f4d2a25d56fe3c773f64de7527f4a71fce2d6653296a2cffbc8d
16	c975c85e9227ee26cdd677234d065cc884993fe9979de5fe03d677d2653f25b7
17	3e689c705baecbd8b3022e798314e8dbe57754c3bc2481f02ff3b1afacdbb5fd
18	dc4996a1ef75807a7585f066dc226007619dac0b42a5bb1216c822a81f7909ff
19	7abc8afe87da6cc91349429534e66a8a9e5d0066d9bc4fadb104fa146815740a
20	d136f0dc688b7322e870ef4399e76b346603e2e078ab24ca0e960febf44b0b65
21	5c94a33bcd3549676a8d1aaf9df8248207211db4b60f13b3cedcb306a4671907

```
kable(multihashes[,.(index,sha3.512)],
      format = "latex",
      align = c("c",
                "p{13cm}"),
      booktabs = TRUE,
      longtable = TRUE)
```

index	sha3.512
1	91aa9717650346dd652b3545d30efaaaae96d67f59ddc59cbbc01d181e5afdf4 27d70191b5f2c962f33e67ae04288023721a04845a008881305028588c0c4d40
2	876fc166e8011769a913ffcf1cf5790262a9e0e3e975cf2c9cc38fe5e67e65c2 6fb292479ab866bec0eb6df6e405ba2321ec49c6fae67b2860ad78407348a86b
3	6c1a8226bf81338255b133ab88a842520e60154495d4933a4b6180e0505f574c a9546a14bcd5ead339c824bc93805be02f727a03b744e1603150946bf55ceb1d
4	86f957bd4edd5ab8a3ca1323a76ac8fe7fb3b11d7d62cf57ae82ad88c54aa2e4 98190d899790d9694d0cac206c3f16647885e3afe5831f61a768b6ec20f83852
5	977f1f4ece9d7bc4e4743f6ac022199f4a3c805bed6493abd515f650ecb207fe 5f58cbcac3199f8f5c39b227503f52167b4e56c9ac1d6c058395d4962c7df312
6	91134175076bc8b99ead82b52fd05f54f4b6dac50b792cf0ae68dc63f9538472 db7141eaf3cba9bf0582f896cd8443318ed8d09d09e3fa65eeced185ce4540c1
7	0473c0856395edf63bbe708bf19687b9737c196273ecefd0d213bd823360e40f ef3e4524d449dcdb5febf18d3e61e6f3fe16e5a51c9d633aea9a304e1a7dbae4
8	ea398fd1103d2bc7cdf3fc6c24e45487c14d3d7081a9c33b29d3684117a3aff3 9f311b590a60cfe917e9dcb20f43452989f16ac5ebd61c5eef9e540d09179757
9	88c76f7aeba99ebcafa12449c3414182f020534583e3fcee8c1c27bf1f09e553 a051003b50ceb2819e340a83b25ff5652024c2c1fa5adc35c2a20d92ed6fe3c4
10	fa4a1914de57313f8bc4d8e8e96ceb3f4a513e9ddc8aeec7696c38cc2bdd5620 3b6a9c091172e45977a1ac734b21f06f70195ddf948d81dc517f27362ebac320
11	de5646b13d9b9f4f74c55f2da0e6000a49e39a3020c710463f84f9caa4d69a0c bd8d6268531169477f8d6e069dd5e580bc9fdc56c7db8fadba799f3ad7628ee4
12	2d373dbe2baee615c4a79631e45a0160d0cfa94ec52df17379bbf5cea6e71899 ddfa4d3ddd010abdc7ea2e4d2fbb6ce24d5b715d7e356be94ef0c2cda77ecd42
13	844d7fc4bd6632c5feba33aa1c9afe2bb8a4eae44c287fab0f24ca0ff9e4d40f 4a322221248243af8e21bf2bbf5ac2c3c384c1f3b306698aafa6675f4087b069
14	0146c788ff6eafd9d4dbf44f99e62e1a66e62174d063009e93e06b9d95e5b643 9aea841d5cd5d77b0ceb72bd040ff6829f7bc9f7befec73369a35d54d0411c16
15	69f48340c35568d4ec5f3da9a8c70f820be23f6d21666bf41fac1fef9a86fac9 349fbca00feae3d01aa19886ef3d81656f8e5f83b580d163f9678d3bb96228eb

16	2f8ec9795dfb072e9964e7d13027e8b1ac1751a41a6ece00083865eae47d464e a03993a0700efca0ea8a6c62a8a543a8d8a6add5fdc17b04a87dc537bbdf1418
17	625ce28bd3c831a62105cf0dd2f74e921a0bb29c5ee306bdf1f4f81d5b81c9e5 7db26fdd0586f6749a887c736f5d975239ad2ff6148812641e8fb45aa5ce54ca
18	f0a89f9b364b4489b14324f7c05116b63669084d5b7a10dfc94aa905e7bc899e 04aecd8d386e8417adef93419f8869336270aa152f044a7b4efddfd2d1e4b19d
19	31b115976de8566fe9cf102fcdafa371966e794e5e13c765fc838dd16d6f9d22 8dea37a1c52277de2ce92318931996ffa51454d1cb3a86141a96617b5e04aeea
20	e9a4c864cdbdc567ebae4190086d53e5f079dbae5cb6366122d70a2fc9426475 10bcdbe71eb1b4945c1b65837d72a869e40b6c77e6ea9568fe2eedbd1ee2ee1d
21	8d928381fabc81f61d440ae8705d57b329960c6e7f956384adc67868895f0d14 99e24702086ab69219addc961ce2d991cffd110f1819099f69c792343a0071b1

34 Finalize

34.1 Datestamp

```
print(datestamp)
```

```
## [1] "2022-09-07"
```

34.2 Date and Time (Begin)

```
print(begin.script)
```

```
## [1] "2022-09-07 01:10:29 CEST"
```

34.3 Date and Time (End)

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

```
## [1] "2022-09-07 09:14:45 CEST"
```

34.4 Script Runtime

```
print(end.script - begin.script)
```

```
## Time difference of 8.071213 hours
```

34.5 Warnings

```
warnings()
```

35 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
##
## 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      rmarkdown_2.16       gtable_0.3.1
## [49] codetools_0.2-18    curl_4.3.2           R6_2.5.1
## [52] gridExtra_2.3       dplyr_1.0.10         fastmap_1.1.0
## [55] utf8_1.2.2          fastmatch_1.1-3      stringi_1.7.8
## [58] Rcpp_1.0.9          vctrs_0.4.1          tidyselect_1.1.2
## [61] 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

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.org/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/.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with Kable and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.