Corpus of Decisions

International Court of Justice

(CD-ICJ-Source)

Compilation Report

Version 2021-11-23

DOI: 10.5281/zenodo.3977177

Title	Source Code for the 'Corpus of Decisions: International Court of Justice'
Abbreviation	CD-ICJ-Source
Author	Seán Fobbe
Version	2021-11-23
Download	https://doi.org/10.5281/zenodo.3977177
License	MIT No Attribution (MIT-0)

Citation

Seán Fobbe (2021). Source Code for the 'Corpus of Decisions: International Court of Justice' (CD-ICJ-Source). Version 2021-11-23. Zenodo. DOI: 10.5281/zenodo.3977177.

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 2021-11-23. 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

1 Introduction **11**
- 1.1 Overview . . . 11
- 1.2 Functionality . . . 11
- 1.3 System Requirements . . . 12
- 1.4 Compilation . . . 12

2 Preamble **13**
- 2.1 Datestamp . . . 13
- 2.2 Date and Time (Begin) . . . 13
- 2.3 Load Packages . . . 13
- 2.4 Load Additional Functions . . . 15

3 Parameters **16**
- 3.1 Read Configuration File . . . 16
- 3.2 Name of Data Set . . . 17
- 3.3 DOI of Data Set Concept . . . 17
- 3.4 DOI of Specific Version . . . 17
- 3.5 License . . . 17
- 3.6 Output Directory . . . 18
- 3.7 Scope: Case Numbers . . . 18
- 3.8 Debugging Mode . . . 18
- 3.9 DPI for OCR . . . 19
- 3.10 Frequency Tables: Ignored Variables . . . 19
- 3.11 Knitr Options . . . 20
  - 3.11.1 Image Output File Formats . . . 20
  - 3.11.2 DPI for Raster Graphics . . . 20
  - 3.11.3 Alignment of Diagrams in Report . . . 20
  - 3.11.4 Set Knitr Options . . . 20
- 3.12 LaTeX Configuration . . . 20
  - 3.12.1 Construct LaTeX Definitions . . . 20
  - 3.12.2 Write LaTeX Definitions . . . 22
- 3.13 Write Package Citations . . . 22

4 Parallelization **23**
- 4.1 Detect Number of Logical Cores . . . 23
- 4.2 Set Number of OCR Control Cores . . . 23
- 4.3 Data.table . . . 23
- 4.4 Quanteda . . . 23

5 Create Directories **24**
- 5.1 Define Set of Data Directories . . . 24
- 5.2 Create Data Directories . . . 24
- 5.3 Create Output Directory . . . 24
- 5.4 Create Directory for Unlabelled Files . . . 24

6 Visualize Corpus Creation Process **25**
- 6.1 Workflow Part 1 . . . 25
- 6.2 Workflow Part 2 . . . 26

7 Prepare Download **30**
7.1 Define Download Scope . . . 30
7.2 Debugging Mode — Reduced Scope . . . 30
7.3 Show Function: f.linkextract . . . 30
7.4 Show Function: f.selectpdflinks . . . 30
7.5 Prepare Empty Link List . . . 31
7.6 Acquire Download Links . . . 31
7.7 Clean Links . . . 35
7.8 Remove Specific Links . . . 35
7.9 Add Specific Links . . . 35

8 Labelling Module **37**
8.1 List Unlabelled Files . . . 37
8.2 Write to Disk . . . 37
8.3 Download Unlabelled Files . . . 38
8.3.1 Prepare . . . 38
8.3.2 Number of Unlabelled Files to Download . . . 38
8.3.3 Timestamp (Unlabelled Download Begin) . . . 38
8.3.4 Execute Download . . . 38
8.3.5 Timestamp (Unlabelled Download End) . . . 39
8.3.6 Duration (Download) . . . 39
8.4 Download Result . . . 39
8.4.1 Number of Files to Download . . . 39
8.4.2 Number of Files Successfully Downloaded . . . 39
8.4.3 Number of Missing Files . . . 39
8.4.4 Names of Missing Files . . . 40
8.5 Store Unlabelled Files . . . 40
8.6 Manual Coding . . . 40
8.7 Read in Corrected Labels . . . 40
8.8 Apply Correct Labels to Link List . . . 40
8.9 REGEX VALIDATION 1: Strictly Validate Links against ICJ Naming Scheme 40
8.9.1 Execute Validation . . . 41
8.9.2 Results of Validation . . . 41
8.9.3 Stop Script on Failure . . . 41
8.10 Detect Duplicate Filenames . . . 41
8.11 Detect Missing Counterparts for each Language Version . . . 41
8.12 Difference in Number of Files . . . 42
8.13 Show Missing French Documents . . . 42
8.14 Show Missing English Documents . . . 43

9 Download Module **44**
9.1 Prepare Download Table . . . 44
9.2 Timestamp (Download Begin) . . . 44
9.3 Execute Download (All Files) . . . 44
9.4 Timestamp (Download End) . . . 44
9.5 Duration (Download) . . . 44
9.6 Debugging Mode — Delete Random Files . . . 45
9.7 Download Result . . . 45
9.7.1 Number of Files to Download . . . 45
9.7.2 Number of Files Successfully Downloaded . . . 45

9.7.3 Number of Missing Files . . . 45
9.7.4 Names of Missing Files . . . 46
9.8 Timestamp (Retry Download Begin) . . . 46
9.9 Retry Download . . . 46
9.10 Timestamp (Retry Download End) . . . 47
9.11 Duration (Retry Download) . . . 47
9.12 Retry Result . . . 47
9.12.1 Successful during Retry . . . 47
9.12.2 Missing after Retry . . . 47
9.13 Final Download Result . . . 48
9.13.1 Number of Files to Download . . . 48
9.13.2 Number of Files Successfully Downloaded . . . 48
9.13.3 Number of Missing Files . . . 48
9.13.4 Names of Missing Files . . . 48

10 File Split Module **49**
10.1 Armed Activities Order . . . 49
10.2 Case 146 . . . 49
10.2.1 English on Even Pages . . . 49
10.2.2 English on Odd Pages . . . 50
10.2.3 Delete Bilingual Files . . . 50

11 Filename Enhancement Module **51**
11.1 Enhance Syntax . . . 51
11.2 Manual Coding . . . 51
11.3 Read Hand Coded Data . . . 51
11.4 Add Hand Coded Data to Filenames . . . 52
11.5 Add Stage of Proceedings . . . 52
11.6 REGEX VALIDATION 2: Strictly Validate Naming Scheme against Codebook Schema . . . 53
11.6.1 Execute Validation . . . 53
11.6.2 Results of Validation . . . 53
11.6.3 Stop Script on Failure . . . 53
11.7 Execute Rename . . . 54

12 Detect Missing Counterparts for each Language Variant **55**
12.1 Difference between French and English File Lists . . . 55
12.2 Show Missing French Documents . . . 55
12.3 Show Missing English Documents . . . 56

13 Text Extraction Module **57**
13.1 Define Set of Files to Process . . . 57
13.2 Number of Files to Process . . . 57
13.3 Show Function: f.dopar.pagenums . . . 57
13.4 Count Pages . . . 58
13.5 Show Function: f.dopar.pdfextract . . . 58
13.6 Extract Text . . . 59
13.7 Copy and Move EXTRACTED TXT Files . . . 59

14 Tesseract OCR Module **60**

14.1 Mark Files for OCR . . . 60
14.2 Copy and Move Born-Digital Files . . . 60
14.3 Show Function: f.dopar.pdfocr . . . 60
14.4 English . . . 61
14.4.1 Number of English Documents to Process . . . 61
14.4.2 Number of English Pages to Process . . . 62
14.4.3 Run OCR on English Documents . . . 62
14.5 French . . . 62
14.5.1 Number of French Documents to Process . . . 62
14.5.2 Number of French Pages to Process . . . 63
14.5.3 Run OCR on French Documents . . . 63
14.6 Rename Files . . . 63
14.7 Copy and Move TXT Files . . . 64
14.8 Copy and Move PDF Files . . . 64

15 Create Majority-Only Variant 65

16 Read in TXT Files 66
16.1 Define Variable Names . . . 66
16.2 BEST Variants . . . 66
16.2.1 English . . . 66
16.2.2 French . . . 66
16.3 EXTRACTED Variants . . . 66
16.3.1 English . . . 66
16.3.2 French . . . 67
16.4 Convert to Data Tables . . . 67

17 Clean Texts 68
17.1 Remove Hyphenation across Linebreaks . . . 68
17.1.1 Show Function: f.hyphen.remove . . . 68
17.1.2 Execute Function . . . 68
17.2 Replace Special Characters . . . 69
17.2.1 Show Function: f.special.replace . . . 69
17.2.2 Execute Function . . . 69

18 OCR Quality Control Module 70
18.1 Create Corpora . . . 70
18.2 Subset to 2004 and earlier . . . 70
18.3 Show Function: f.token.processor . . . 70
18.4 Tokenize . . . 71
18.5 Create Document-Feature-Matrices . . . 71
18.6 Features Reduction . . . 71

19 Language Purity Module 73
19.1 Limit Detection to English and French . . . 73
19.2 Automatic Language Detection . . . 73
19.3 Detected Languages . . . 73
19.4 Show Mismatches . . . 73
19.5 Final Note: Human Review of Mismatches . . . 74

20 Add and Delete Variables **75**
20.1 Delete Textcat Classifications . . . 75
20.2 Add Variable “year” . . . 75
20.3 Add Variable “minority” . . . 75
20.4 Add Variable “fullname” . . . 75
20.4.1 Read Hand Coded Data . . . 75
20.4.2 Create Variable . . . 75
20.5 Add Variable “applicant_region” . . . 75
20.5.1 Read Hand Coded Data . . . 75
20.5.2 Merge Regions for English Version . . . 76
20.5.3 Merge Regions for French Version . . . 76
20.6 Add Variable “respondent_region” . . . 76
20.6.1 Read Hand Coded Data . . . 76
20.6.2 Merge Regions for English Version . . . 76
20.6.3 Merge Regions for French Version . . . 77
20.7 Add Variable “applicant_subregion” . . . 77
20.7.1 Read Hand Coded Data . . . 77
20.7.2 Merge Subregions for English Version . . . 77
20.7.3 Merge Subregions for French Version . . . 78
20.8 Add Variable “respondent_subregion” . . . 78
20.8.1 Read Hand Coded Data . . . 78
20.8.2 Merge Subregions for English Version . . . 78
20.8.3 Merge Subregions for French Version . . . 78
20.9 Add Variable “doi_concept” . . . 79
20.10 Add Variable “doi_version” . . . 79
20.11 Add Variable “version” . . . 79
20.12 Add Variable “license” . . . 79

21 Frequency Tables **80**
21.1 Show Function: f.fast.freqtable . . . 80
21.2 English Corpus . . . 81
21.2.1 Variables to Ignore . . . 81
21.2.2 Variables to Analyze . . . 81
21.2.3 Construct Frequency Tables . . . 82
21.3 French Corpus . . . 127
21.3.1 Variables to Ignore . . . 127
21.3.2 Variables to Analyze . . . 127
21.3.3 Construct Frequency Tables . . . 127

22 Visualize Frequency Tables **172**
22.1 Load Tables . . . 172
22.2 Doctype . . . 173
22.2.1 English . . . 173
22.2.2 French . . . 175
22.3 Opinion . . . 177
22.3.1 English . . . 177
22.3.2 French . . . 179
22.4 Year . . . 181
22.4.1 English . . . 181
22.4.2 French . . . 183

23 Summary Statistics **184**
23.1 Linguistic Metrics . . . 184
23.1.1 Show Function: f.lingsummarize.iterator . . . 184
23.1.2 Calculate Linguistic Metrics . . . 186
23.1.3 Add Linguistic Metrics to Full Corpora . . . 186
23.1.4 Create Metadata-only Variants . . . 186
23.1.5 Calculate Summaries: English . . . 187
23.1.6 Show Summaries: English . . . 188
23.1.7 Write Summaries to Disk: English . . . 188
23.1.8 Calculate Summaries: French . . . 189
23.1.9 Show Summaries: French . . . 190
23.1.10 Write Summaries to Disk: French . . . 190
23.2 Distributions . . . 191
23.2.1 Tokens per Year: English . . . 191
23.2.2 Tokens per Year: French . . . 193
23.2.3 Density: Characters . . . 195
23.2.4 Density: Tokens . . . 197
23.2.5 Density: Types . . . 199
23.2.6 Density: Sentences . . . 201
23.2.7 All Distributions of Linguistic Metrics . . . 203
23.3 Number of Majority Opinions . . . 208
23.3.1 English . . . 208
23.3.2 French . . . 209
23.4 Number of Minority Opinions . . . 210
23.4.1 English . . . 210
23.4.2 French . . . 211
23.5 Year Range . . . 211
23.6 Date Range . . . 212

24 Test and Sort Variable Names **213**
24.1 Semantic Sorting of Variable Names . . . 213
24.1.1 Sort Variables: Full Data Set . . . 213
24.1.2 Sort Variables: Metadata . . . 215
24.2 Number of Variables: Full Data Set . . . 217
24.3 Number of Variables: Metadata . . . 217
24.4 List All Variables: Full Data Set . . . 217
24.5 List All Variables: Metadata . . . 218

25 Calculate Detailed Token Frequencies **219**
25.1 Create Corpora . . . 219
25.2 Process Tokens . . . 219
25.3 Construct Document-Feature-Matrices . . . 219
25.4 Most Frequent Tokens | TF Weighting | Tables . . . 219
25.4.1 English . . . 219
25.4.2 French . . . 223
25.5 Most Frequent Tokens | TFIDF Weighting | Tables . . . 227
25.5.1 English . . . 227
25.5.2 French . . . 231
25.6 Most Frequent Tokens | TF Weighting | Scatterplots . . . 236
25.6.1 English . . . 236

25.6.2 French . . . 238
25.7 Most Frequent Tokens | TFIDF Weighting | Scatterplots . . . 240
25.7.1 English . . . 240
25.7.2 French . . . 242
25.8 Most Frequent Tokens | TF Weighting | Wordclouds . . . 244
25.8.1 English . . . 244
25.8.2 French . . . 245
25.9 Most Frequent Tokens | TFIDF Weighting | Wordclouds . . . 246
25.9.1 English . . . 246
25.9.2 French . . . 247

26 Document Similarity 248
26.1 Set Ranges . . . 248
26.2 English . . . 249
26.2.1 Calculate Similarity . . . 249
26.2.2 Create Empty Lists . . . 249
26.2.3 Build Tables . . . 249
26.2.4 IDs of Paired Documents Above Threshold . . . 249
26.2.5 IDs of Duplicate Documents per Threshold . . . 250
26.2.6 Count of Duplicate Documents per Threshold . . . 250
26.3 French . . . 254
26.3.1 Calculate Similarity . . . 254
26.3.2 Create Empty Lists . . . 254
26.3.3 Build Tables . . . 254
26.3.4 IDs of Paired Documents Above Threshold . . . 254
26.3.5 IDs of Duplicate Documents per Threshold . . . 255
26.3.6 Count of Duplicate Documents per Threshold . . . 255

27 Create CSV Files 259
27.1 Full Data Set . . . 259
27.2 Metadata Only . . . 259

28 Final File Count per Folder 260

29 File Size Distribution 261
29.1 English . . . 261
29.1.1 Corpus Object in RAM . . . 261
29.1.2 Create Data Table of Filenames . . . 261
29.1.3 Total Size Comparison . . . 261
29.1.4 Analyze Files Larger than 10 MB . . . 262
29.1.5 Plot Density Distribution for Files 10MB or Less . . . 265
29.2 French . . . 267
29.2.1 Corpus Object in RAM . . . 267
29.2.2 Create Data Table of filenames . . . 267
29.2.3 Total Size Comparison . . . 267
29.2.4 Analyze Files Larger than 10 MB . . . 268
29.2.5 Plot Density Distribution for Files 10MB or Less . . . 271

30 Create ZIP Archives 273
30.1 ZIP CSV Files . . . 273

30.2 ZIP Data Directories . . . 273
30.3 ZIP ANALYSIS Directory . . . 274
30.4 ZIP Unlabelled Files Directory . . . 274
30.5 ZIP Source Files . . . 274

31 Delete CSV and Directories **275**
31.1 Delete CSVs . . . 275
31.2 Delete Data Directories . . . 275

32 Cryptography Module **276**
32.1 Create Set of ZIP Archives . . . 276
32.2 Show Function: f.dopar.multihashes . . . 276
32.3 Compute Hashes . . . 277
32.4 Convert to Data Table . . . 277
32.5 Add Index . . . 277
32.6 Save to Disk . . . 278
32.7 Add Whitespace to Enable Automatic Linebreak . . . 278
32.8 Print to Report . . . 279

33 Finalize **283**
33.1 Datestamp . . . 283
33.2 Date and Time (Begin) . . . 283
33.3 Date and Time (End) . . . 283
33.4 Script Runtime . . . 283
33.5 Warnings . . . 283

34 Strict Replication Parameters **284**

References **288**

1 Introduction

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 its website (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 **R** 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("CD-ICJ_Source_FullCompilation.R")
```

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] "2021-11-23"
```

2.2 Date and Time (Begin)

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

```
## [1] "2021-11-23 03:59:58 CET"
```

2.3 Load Packages

```
library(httr)           # HTTP Tools
library(rvest)          # Web Scraping
library(mgsub)          # Vectorized Gsub
library(stringr)        # String Manipulation
library(pdftools)       # PDF utilities
```

```
## Using poppler version 21.01.0
```

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

```
## Linking to ImageMagick 6.9.11.27
## Enabled features: cairo, fontconfig, freetype, fftw, ghostscript, lcms, pango,
     raw, rsvg, webp, x11
## Disabled features: heic
```

```
## Using 16 threads
```

```
library(DiagrammeR)    # Graph/Network Visualization
library(DiagrammeRsvg) # Export DiagrammeR Graphs as SVG
library(rsvg)          # Render SVG to PDF
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: 3.1.0
## Unicode version: 13.0
## ICU version: 67.1
```

```
## 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 1.14.0 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 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!

The configuration is read, printed, re-written to a temporary file and re-read to achieve transposition with correct column classes, something fread() cannot do directly. This procedure allows for a source CSV file that is easier to edit and easier to access within R.

```
config <- fread("CD-ICJ_Source_Config.csv")

kable(config,
      format = "latex",
      align = c("p{5cm}",
                "p{9cm}"),
      booktabs = TRUE,
      col.names = c("Key",
                    "Value"))
```

Key	Value
datatitle	Corpus of Decisions: International Court of Justice
datashort	CD-ICJ
doi.data.concept	10.5281/zenodo.3826444
doi.data.version	10.5281/zenodo.3826445
doi.software.concept	10.5281/zenodo.3977176
doi.software.version	10.5281/zenodo.3977177
license	Creative Commons Zero 1.0 Universal
caseno.begin	1
caseno.end	181
caseno.exclude	2
mode.debug.toggle	FALSE
mode.debug.sample	3
ocr.dpi	300
plot.format	pdf png
plot.dpi	300
fig.align	center
freq.var.ignore	date doc_id text

```
temp <- transpose(config,
                  make.names = "key")

fwrite(temp,
       "temp.csv")

config <- fread("temp.csv")

unlink("temp.csv")
```

3.2 Name of Data Set

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

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

3.3 DOI of Data Set Concept

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

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

3.4 DOI of Specific Version

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

```
## [1] "10.5281/zenodo.3826445"
```

3.5 License

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

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

3.6 Output Directory

The directory name must include a terminating slash!

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

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] 181
```

```
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$mode.debug.toggle
mode.debug.sample <- config$mode.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 <- unlist(tstrsplit(config$freq.var.ignore,
                                    split = " "))

print(freq.var.ignore)
```

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

3.11 Knitr Options

3.11.1 Image Output File Formats

```
plot.format <- unlist(tstrsplit(config$plot.format,
                                split = " "))

print(plot.format)
```

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

3.11.2 DPI for Raster Graphics

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

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

3.11.3 Alignment of Diagrams in Report

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

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

3.11.4 Set Knitr Options

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

3.12 LaTeX Configuration

3.12.1 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$datatitle,
                   "}"),
            paste0("\\newcommand{\\datashort}{",
                   config$datashort,
                   "}"),
            paste0("\\newcommand{\\softwaretitle}{Source Code for the \\
enquote{",
                   config$datatitle,
                   "}}"),
            paste0("\\newcommand{\\softwareshort}{",
                   config$datashort,
                   "-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,
                   "}"))
```

3.12.2 Write LaTeX Definitions

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

3.13 Write Package Citations

```
write_bib(c(.packages()),
          "packages.bib")
```

```
## tweaking foreach
```

4 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.

4.1 Detect Number of Logical Cores

This will detect the maximum number of threads (= logical cores) available on the system.

```
fullCores <- detectCores()
print(fullCores)
```

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

4.2 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
```

4.3 Data.table

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

4.4 Quanteda

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

5 Create Directories

5.1 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")
```

5.2 Create Data Directories

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

5.3 Create Output Directory

```
dir.create(outputdir)
```

5.4 Create Directory for Unlabelled Files

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

dir.create(dir.unlabelled)
```

6 Visualize Corpus Creation Process

6.1 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")
```

6.2 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']
```

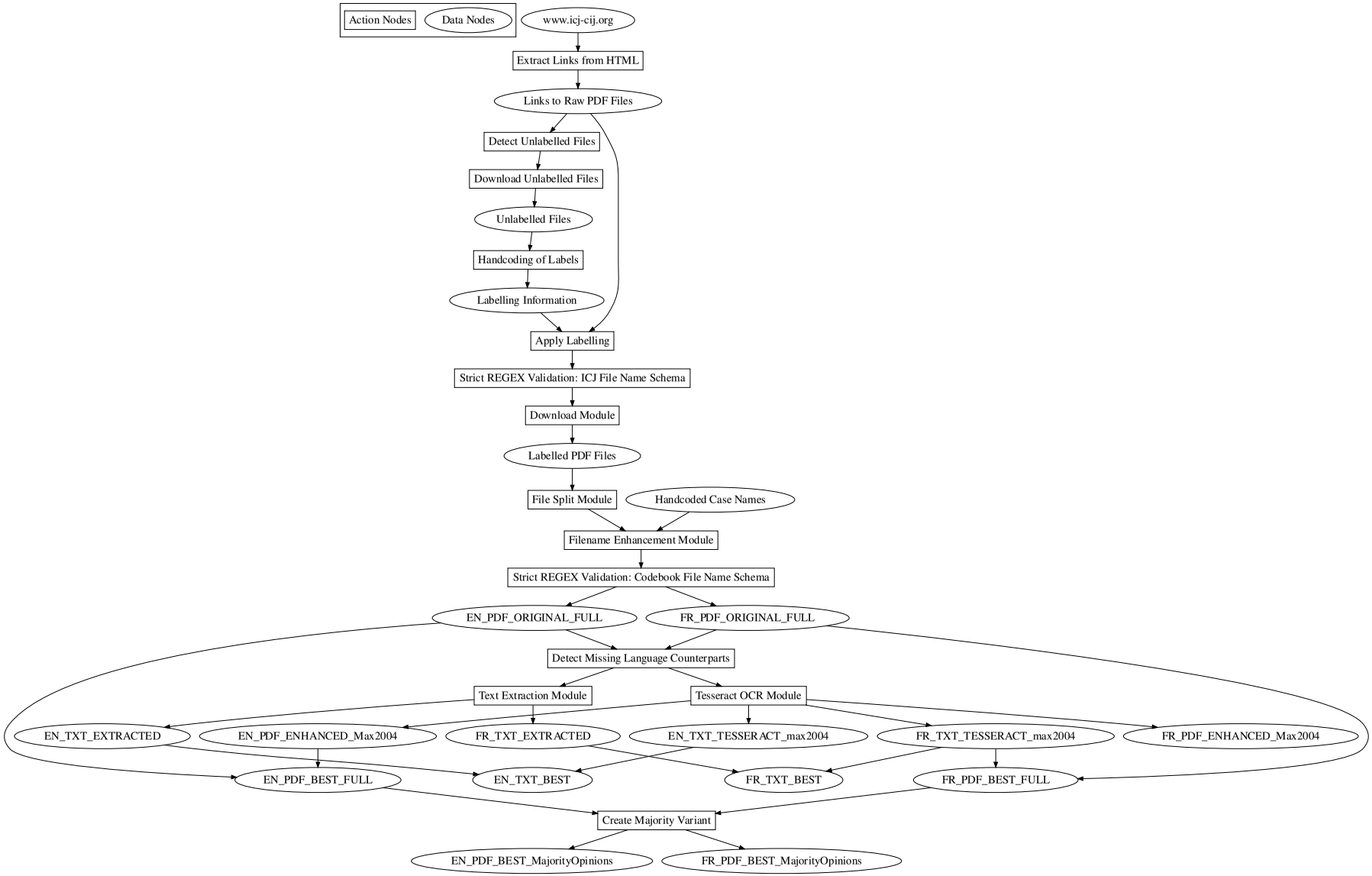


Figure 1: Workflow Part 1: Download, Labelling, Conversion and Sorting of Documents

```
  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")
```

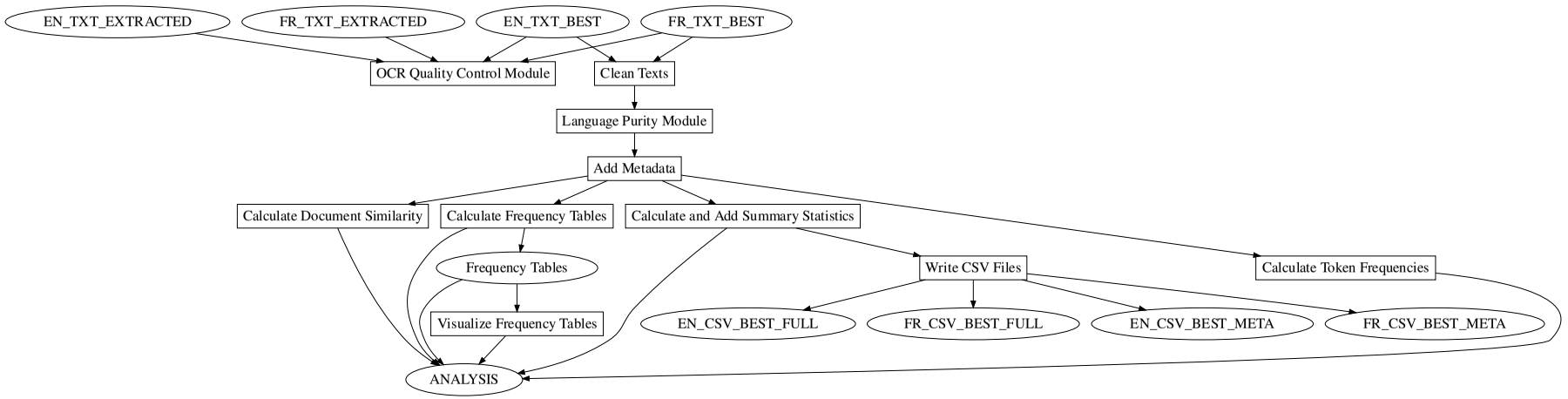


Figure 2: Workflow Part 2: Ingestion, Pre-Processing, Analysis and Creation of CSV Files

7 Prepare Download

7.1 Define Download Scope

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

7.2 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))
    caseno.full <- sort(caseno.full)
    }
```

7.3 Show Function: f.linkextract

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

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

7.4 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)
## }
```

7.5 Prepare Empty Link List

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

7.6 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
```

7.7 Clean Links

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

links.unique <- unique(links)

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

7.8 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)
```

7.9 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")
```

8 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.

8.1 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"
```

8.2 Write to Disk

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

8.3 Download Unlabelled Files

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

8.3.1 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)
```

8.3.2 Number of Unlabelled Files to Download

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

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

8.3.3 Timestamp (Unlabelled Download Begin)

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

```
## [1] "2021-11-23 04:05:33 CET"
```

8.3.4 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))
    }
```

8.3.5 Timestamp (Unlabelled Download End)

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

```
## [1] "2021-11-23 04:06:00 CET"
```

8.3.6 Duration (Download)

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

```
## Time difference of 27.60043 secs
```

8.4 Download Result

8.4.1 Number of Files to Download

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

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

8.4.2 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
```

8.4.3 Number of Missing Files

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

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

8.4.4 Names of Missing Files

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

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

8.5 Store Unlabelled Files

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

8.6 Manual Coding

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

8.7 Read in Corrected Labels

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

8.8 Apply Correct Labels to Link List

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

8.9 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!

8.9.1 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)
```

8.9.2 Results of Validation

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

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

8.9.3 Stop Script on Failure

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

8.10 Detect Duplicate Filenames

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

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

8.11 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)
```

8.12 Difference in Number of Files

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

```
## [1] 9
```

8.13 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/125/125-20040709-ORD
    -01-00-FR.pdf"
##  [3] "https://www.icj-cij.org/public/files/case-related/156/156-20150422-ORD
    -01-01-FR.pdf"
##  [4] "https://www.icj-cij.org/public/files/case-related/161/161-20211012-JUD
    -01-01-FR.pdf"
##  [5] "https://www.icj-cij.org/public/files/case-related/161/161-20211012-JUD
    -01-03-FR.pdf"
##  [6] "https://www.icj-cij.org/public/files/case-related/161/161-20211012-JUD
    -01-04-FR.pdf"
##  [7] "https://www.icj-cij.org/public/files/case-related/161/161-20211012-JUD
    -01-05-FR.pdf"
##  [8] "https://www.icj-cij.org/public/files/case-related/172/172-20210204-JUD
    -01-01-FR.pdf"
##  [9] "https://www.icj-cij.org/public/files/case-related/172/172-20210204-JUD
    -01-02-FR.pdf"
## [10] "https://www.icj-cij.org/public/files/case-related/172/172-20210204-JUD
    -01-03-FR.pdf"
```

```
## [11] "https://www.icj-cij.org/public/files/case-related/172/172-20210204-JUD
    -01-04-FR.pdf"
## [12] "https://www.icj-cij.org/public/files/case-related/172/172-20210204-JUD
    -01-05-FR.pdf"
```

8.14 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/161/161-20211012-JUD
    -01-02-EN.pdf"
## [2] "https://www.icj-cij.org/public/files/case-related/161/161-20211012-JUD
    -01-06-EN.pdf"
## [3] "https://www.icj-cij.org/public/files/case-related/172/172-20210204-JUD
    -01-06-EN.pdf"
```

9 Download Module

9.1 Prepare Download Table

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

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

9.2 Timestamp (Download Begin)

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

```
## [1] "2021-11-23 04:06:02 CET"
```

9.3 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))

}
```

9.4 Timestamp (Download End)

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

```
## [1] "2021-11-23 05:38:50 CET"
```

9.5 Duration (Download)

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

```
## Time difference of 1.546703 hours
```

9.6 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))

}
```

9.7 Download Result

9.7.1 Number of Files to Download

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

```
## [1] 4326
```

9.7.2 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] 4326
```

9.7.3 Number of Missing Files

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

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

9.7.4 Names of Missing Files

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

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

9.8 Timestamp (Retry Download Begin)

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

```
## [1] "2021-11-23 05:38:50 CET"
```

9.9 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))
    }
}
```

9.10 Timestamp (Retry Download End)

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

```
## [1] "2021-11-23 05:38:50 CET"
```

9.11 Duration (Retry Download)

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

```
## Time difference of 0.006034613 secs
```

9.12 Retry Result

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

9.12.1 Successful during Retry

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

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

9.12.2 Missing after Retry

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

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

9.13 Final Download Result

9.13.1 Number of Files to Download

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

```
## [1] 4326
```

9.13.2 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] 4326
```

9.13.3 Number of Missing Files

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

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

9.13.4 Names of Missing Files

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

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

10 File Split Module

10.1 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)
```

10.2 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.

10.2.1 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)
}
```

10.2.2 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)
}
```

10.2.3 Delete Bilingual Files

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

11 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")
```

11.1 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)
```

11.2 Manual Coding

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

11.3 Read Hand Coded Data

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

11.4 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)
```

11.5 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)
```

11.6 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!

11.6.1 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)
```

11.6.2 Results of Validation

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

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

11.6.3 Stop Script on Failure

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

11.7 Execute Rename

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

12 Detect Missing Counterparts for each Language Variant

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

12.1 Difference between French and English File Lists

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

```
## [1] 9
```

12.2 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_125_FrontierDispute_BEN_NER_2004-07-09_ORD_01_NA_00_FR.pdf"
##  [3] "ICJ_156_CertainDocumentsSeizure_TLS_AUS_2015-04-22_ORD_01_NA_01_FR.pdf"
##  [4] "ICJ_161_MaritimeDelimitation-IndianOcean_SOM_KEN_2021-10-12_JUD_01_NA
    _01_FR.pdf"
##  [5] "ICJ_161_MaritimeDelimitation-IndianOcean_SOM_KEN_2021-10-12_JUD_01_NA
    _03_FR.pdf"
##  [6] "ICJ_161_MaritimeDelimitation-IndianOcean_SOM_KEN_2021-10-12_JUD_01_NA
    _04_FR.pdf"
##  [7] "ICJ_161_MaritimeDelimitation-IndianOcean_SOM_KEN_2021-10-12_JUD_01_NA
    _05_FR.pdf"
##  [8] "ICJ_172_CERD_QAT_ARE_2021-02-04_JUD_01_PO_01_FR.pdf"
##  [9] "ICJ_172_CERD_QAT_ARE_2021-02-04_JUD_01_PO_02_FR.pdf"
## [10] "ICJ_172_CERD_QAT_ARE_2021-02-04_JUD_01_PO_03_FR.pdf"
## [11] "ICJ_172_CERD_QAT_ARE_2021-02-04_JUD_01_PO_04_FR.pdf"
## [12] "ICJ_172_CERD_QAT_ARE_2021-02-04_JUD_01_PO_05_FR.pdf"
```

12.3 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_161_MaritimeDelimitation-IndianOcean_SOM_KEN_2021-10-12_JUD_01_NA_02_
    EN.pdf"
## [2] "ICJ_161_MaritimeDelimitation-IndianOcean_SOM_KEN_2021-10-12_JUD_01_NA_06_
    EN.pdf"
## [3] "ICJ_172_CERD_QAT_ARE_2021-02-04_JUD_01_PO_06_EN.pdf"
```

13 Text Extraction Module

13.1 Define Set of Files to Process

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

13.2 Number of Files to Process

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

```
## [1] 4329
```

13.3 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)
}
```

}

13.4 Count Pages

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

```
## [1] "Parallel processing using 16 threads."
## [1] "Total number of pages: 62901"
```

```
## [1] 62901
```

13.5 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)
```

}

13.6 Extract Text

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

```
## [1] "Parallel processing using 16 threads. Begin at 2021-11-23 05:39:51"
## [1] "Processed 4329 files. Runtime was 11.74 secs. Ended at 2021-11-23
    05:40:02."
```

13.7 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")
```

14 Tesseract OCR Module

14.1 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")
```

14.2 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")
```

14.3 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)
```

}

14.4 English

14.4.1 Number of English Documents to Process

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

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

14.4.2 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: 20508"
```

```
## [1] 20508
```

14.4.3 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 2021-11-23 05:40:03"
## [1] "Processed 1484 files. Runtime was 3.62 hours. Ended at 2021-11-23
    09:17:24."
```

14.5 French

14.5.1 Number of French Documents to Process

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

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

14.5.2 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
```

14.5.3 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 2021-11-23 09:17:24"
## [1] "Processed 1482 files. Runtime was 4.69 hours. Ended at 2021-11-23
    13:58:39."
```

14.6 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)
```

14.7 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")
```

14.8 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")
```

15 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")
```

16 Read in TXT Files

16.1 Define Variable Names

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

16.2 BEST Variants

16.2.1 English

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

16.2.2 French

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

16.3 EXTRACTED Variants

16.3.1 English

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

16.3.2 French

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

16.4 Convert to Data Tables

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

17 Clean Texts

17.1 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.

17.1.1 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)
## }
```

17.1.2 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)]
```

17.2 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.

17.2.1 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)
## }
```

17.2.2 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)]
```

18 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.

18.1 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)
```

18.2 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)
```

18.3 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)
##     }
```

18.4 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)
```

18.5 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)
```

18.6 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	115816	57686	58130	50.19 %
French	135811	78031	57780	42.54 %

19 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.

19.1 Limit Detection to English and French

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

19.2 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)
```

19.3 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"
```

19.4 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
```

19.5 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.

20 Add and Delete Variables

20.1 Delete Textcat Classifications

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

20.2 Add Variable "year"

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

20.3 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
```

20.4 Add Variable "fullname"

20.4.1 Read Hand Coded Data

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

20.4.2 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)]
```

20.5 Add Variable "applicant_region"

20.5.1 Read Hand Coded Data

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

20.5.2 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
```

20.5.3 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
```

20.6 Add Variable “respondent_region”

20.6.1 Read Hand Coded Data

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

20.6.2 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
```

20.6.3 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
```

20.7 Add Variable "applicant_subregion"

20.7.1 Read Hand Coded Data

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

20.7.2 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
```

20.7.3 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
```

20.8 Add Variable "respondent_subregion"

20.8.1 Read Hand Coded Data

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

20.8.2 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
```

20.8.3 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
```

20.9 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])
```

20.10 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])
```

20.11 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]))
```

20.12 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]))
```

21 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.

21.1 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)
}
```

}

21.2 English Corpus

21.2.1 Variables to Ignore

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

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

21.2.2 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"
```

21.2.3 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	2169	100	100	100
Total	2169	100	100	100

Frequency Table for Variable: caseno

178 unique value(s) detected.

caseno	N	exactpercent	roundedpercent	cumulpercent
1	19	0.8759797	0.88	0.88
3	7	0.3227294	0.32	1.20
4	7	0.3227294	0.32	1.52
5	9	0.4149378	0.41	1.94
6	1	0.0461042	0.05	1.98
7	9	0.4149378	0.41	2.40
8	10	0.4610420	0.46	2.86
9	4	0.1844168	0.18	3.04
10	7	0.3227294	0.32	3.37
11	6	0.2766252	0.28	3.64
12	4	0.1844168	0.18	3.83
13	1	0.0461042	0.05	3.87
14	2	0.0922084	0.09	3.96
15	16	0.7376671	0.74	4.70
16	12	0.5532503	0.55	5.26
17	7	0.3227294	0.32	5.58
18	11	0.5071462	0.51	6.09
19	7	0.3227294	0.32	6.41
20	3	0.1383126	0.14	6.55
21	6	0.2766252	0.28	6.82
22	1	0.0461042	0.05	6.87
23	1	0.0461042	0.05	6.92
24	5	0.2305210	0.23	7.15
25	1	0.0461042	0.05	7.19
26	1	0.0461042	0.05	7.24
27	1	0.0461042	0.05	7.28
28	1	0.0461042	0.05	7.33
29	11	0.5071462	0.51	7.84
30	9	0.4149378	0.41	8.25
31	6	0.2766252	0.28	8.53

(continued)

caseno	N	exactpercent	roundedpercent	cumulpercent
32	25	1.1526049	1.15	9.68
33	11	0.5071462	0.51	10.19
34	16	0.7376671	0.74	10.93
35	8	0.3688336	0.37	11.30
36	8	0.3688336	0.37	11.66
37	5	0.2305210	0.23	11.89
38	8	0.3688336	0.37	12.26
39	6	0.2766252	0.28	12.54
40	1	0.0461042	0.05	12.59
41	5	0.2305210	0.23	12.82
42	4	0.1844168	0.18	13.00
43	4	0.1844168	0.18	13.19
44	1	0.0461042	0.05	13.23
45	16	0.7376671	0.74	13.97
46	30	1.3831259	1.38	15.35
47	30	1.3831259	1.38	16.74
48	18	0.8298755	0.83	17.57
49	11	0.5071462	0.51	18.07
50	31	1.4292301	1.43	19.50
51	15	0.6915629	0.69	20.19
52	15	0.6915629	0.69	20.89
53	16	0.7376671	0.74	21.62
54	14	0.6454587	0.65	22.27
55	24	1.1065007	1.11	23.37
56	25	1.1526049	1.15	24.53
57	11	0.5071462	0.51	25.03
58	31	1.4292301	1.43	26.46
59	29	1.3370217	1.34	27.80
60	5	0.2305210	0.23	28.03

(continued)

caseno	N	exactpercent	roundedpercent	cumulpercent
61	15	0.6915629	0.69	28.72
62	20	0.9220839	0.92	29.64
63	14	0.6454587	0.65	30.29
64	7	0.3227294	0.32	30.61
65	11	0.5071462	0.51	31.12
66	11	0.5071462	0.51	31.63
67	12	0.5532503	0.55	32.18
68	22	1.0142923	1.01	33.20
69	8	0.3688336	0.37	33.56
70	32	1.4753343	1.48	35.04
71	5	0.2305210	0.23	35.27
72	10	0.4610420	0.46	35.73
73	3	0.1383126	0.14	35.87
74	11	0.5071462	0.51	36.38
75	20	0.9220839	0.92	37.30
76	6	0.2766252	0.28	37.57
77	6	0.2766252	0.28	37.85
78	13	0.5993545	0.60	38.45
79	9	0.4149378	0.41	38.87
80	11	0.5071462	0.51	39.37
81	5	0.2305210	0.23	39.60
82	15	0.6915629	0.69	40.30
83	8	0.3688336	0.37	40.66
84	10	0.4610420	0.46	41.12
85	1	0.0461042	0.05	41.17
86	7	0.3227294	0.32	41.49
87	29	1.3370217	1.34	42.83
88	28	1.2909175	1.29	44.12
89	26	1.1987091	1.20	45.32

(continued)

caseno	N	exactpercent	roundedpercent	cumulpercent
90	32	1.4753343	1.48	46.80
91	43	1.9824804	1.98	48.78
92	16	0.7376671	0.74	49.52
93	9	0.4149378	0.41	49.93
94	34	1.5675426	1.57	51.50
95	16	0.7376671	0.74	52.24
96	13	0.5993545	0.60	52.84
97	8	0.3688336	0.37	53.20
98	12	0.5532503	0.55	53.76
99	7	0.3227294	0.32	54.08
100	6	0.2766252	0.28	54.36
101	4	0.1844168	0.18	54.54
102	14	0.6454587	0.65	55.19
103	24	1.1065007	1.11	56.29
104	11	0.5071462	0.51	56.80
105	21	0.9681881	0.97	57.77
106	21	0.9681881	0.97	58.74
107	19	0.8759797	0.88	59.61
108	19	0.8759797	0.88	60.49
109	20	0.9220839	0.92	61.41
110	21	0.9681881	0.97	62.38
111	21	0.9681881	0.97	63.35
112	9	0.4149378	0.41	63.76
113	21	0.9681881	0.97	64.73
114	7	0.3227294	0.32	65.05
115	3	0.1383126	0.14	65.19
116	28	1.2909175	1.29	66.48
117	3	0.1383126	0.14	66.62
118	31	1.4292301	1.43	68.05

(continued)

caseno	N	exactpercent	roundedpercent	cumulpercent
119	7	0.3227294	0.32	68.37
120	8	0.3688336	0.37	68.74
121	22	1.0142923	1.01	69.76
122	6	0.2766252	0.28	70.03
123	8	0.3688336	0.37	70.40
124	35	1.6136468	1.61	72.01
125	7	0.3227294	0.32	72.34
126	16	0.7376671	0.74	73.08
127	4	0.1844168	0.18	73.26
128	11	0.5071462	0.51	73.77
129	11	0.5071462	0.51	74.27
130	9	0.4149378	0.41	74.69
131	11	0.5071462	0.51	75.20
132	4	0.1844168	0.18	75.38
133	6	0.2766252	0.28	75.66
134	1	0.0461042	0.05	75.70
135	20	0.9220839	0.92	76.63
136	11	0.5071462	0.51	77.13
137	13	0.5993545	0.60	77.73
138	4	0.1844168	0.18	77.92
139	8	0.3688336	0.37	78.28
140	16	0.7376671	0.74	79.02
141	11	0.5071462	0.51	79.53
142	8	0.3688336	0.37	79.90
143	15	0.6915629	0.69	80.59
144	17	0.7837713	0.78	81.37
145	3	0.1383126	0.14	81.51
146	5	0.2305210	0.23	81.74
147	1	0.0461042	0.05	81.79

(continued)

caseno	N	exactpercent	roundedpercent	cumulpercent
148	17	0.7837713	0.78	82.57
149	8	0.3688336	0.37	82.94
150	40	1.8441678	1.84	84.79
151	14	0.6454587	0.65	85.43
152	16	0.7376671	0.74	86.17
153	14	0.6454587	0.65	86.81
154	13	0.5993545	0.60	87.41
155	15	0.6915629	0.69	88.11
156	11	0.5071462	0.51	88.61
157	13	0.5993545	0.60	89.21
158	17	0.7837713	0.78	90.00
159	17	0.7837713	0.78	90.78
160	17	0.7837713	0.78	91.56
161	15	0.6915629	0.69	92.25
162	4	0.1844168	0.18	92.44
163	26	1.1987091	1.20	93.64
164	12	0.5532503	0.55	94.19
165	10	0.4610420	0.46	94.65
166	23	1.0603965	1.06	95.71
167	1	0.0461042	0.05	95.76
168	11	0.5071462	0.51	96.27
169	15	0.6915629	0.69	96.96
170	1	0.0461042	0.05	97.00
171	9	0.4149378	0.41	97.42
172	22	1.0142923	1.01	98.43
173	6	0.2766252	0.28	98.71
174	6	0.2766252	0.28	98.99
175	11	0.5071462	0.51	99.49
176	1	0.0461042	0.05	99.54

(continued)

caseno	N	exactpercent	roundedpercent	cumulpercent
177	2	0.0922084	0.09	99.63
178	7	0.3227294	0.32	99.95
179	1	0.0461042	0.05	100.00
Total	2169	100.0000000	100.00	100.00

Frequency Table for Variable: shortname

147 unique value(s) detected.

shortname	N	exactpercent	roundedpercent	cumulpercent
1955AmityTreaty	11	0.5071462	0.51	0.51
ATILO-UNESCO	9	0.4149378	0.41	0.92
AccessPacificOcean	14	0.6454587	0.65	1.57
AdmissionUN	7	0.3227294	0.32	1.89
AegeanSeaContinentalShelf	20	0.9220839	0.92	2.81
AerialHerbicideSpraying	4	0.1844168	0.18	3.00
AerialIncident1952	1	0.0461042	0.05	3.04
AerialIncident1953	1	0.0461042	0.05	3.09
AerialIncident1988	9	0.4149378	0.41	3.50
AerialIncident1999	7	0.3227294	0.32	3.83
AerialIncidentNov1954	1	0.0461042	0.05	3.87
AerialIncidentSept1954	1	0.0461042	0.05	3.92
AerialIndicent1955	21	0.9681881	0.97	4.89
Ambatielos	16	0.7376671	0.74	5.62
AngloIranianOil	12	0.5532503	0.55	6.18
Antarctica	2	0.0922084	0.09	6.27
ApplicationGenocideConvention	81	3.7344398	3.73	10.00
ApplicationGenocideConvention-Revision	6	0.2766252	0.28	10.28
ArbitralAward1899	9	0.4149378	0.41	10.70

(continued)

shortname	N	exactpercent	roundedpercent	cumulpercent
ArbitralAward1989	15	0.6915629	0.69	11.39
ArbitralAwardKingOfSpain	6	0.2766252	0.28	11.66
ArbitrationUNHQAgreement	6	0.2766252	0.28	11.94
ArmedActivities	34	1.5675426	1.57	13.51
ArmedActivitiesApp2002	16	0.7376671	0.74	14.25
ArrestWarrant	22	1.0142923	1.01	15.26
Asylum	9	0.4149378	0.41	15.68
Asylum-Interpretation	1	0.0461042	0.05	15.72
Avena	11	0.5071462	0.51	16.23
Avena-Interpretation	8	0.3688336	0.37	16.60
BarcelonaTraction1958	5	0.2305210	0.23	16.83
BarcelonaTraction1962	31	1.4292301	1.43	18.26
CERD	22	1.0142923	1.01	19.27
CertainActivitiesBorderArea	40	1.8441678	1.84	21.12
CertainCriminalProceedings	11	0.5071462	0.51	21.62
CertainDocumentsSeizure	11	0.5071462	0.51	22.13
CertainExpensesUN	11	0.5071462	0.51	22.64
CertainPhosphateLands	11	0.5071462	0.51	23.14
CertainProperty	8	0.3688336	0.37	23.51
ChagosArchipelago	15	0.6915629	0.69	24.20
CompensationUNAT	6	0.2766252	0.28	24.48
CompetenceAdmissionGA	4	0.1844168	0.18	24.67
ConstitutionMaritimeSafetyCommittee	4	0.1844168	0.18	24.85
ConstructionWallOPT	11	0.5071462	0.51	25.36
ContinentalShelf	36	1.6597510	1.66	27.02
ContinentalShelf-InterpretationRevision	5	0.2305210	0.23	27.25
ConventionPrivilegesImmunitiesUN	5	0.2305210	0.23	27.48
ConventionTerrorismFinancingCERD	23	1.0603965	1.06	28.54

(continued)

shortname	N	exactpercent	roundedpercent	cumulpercent
CorfuChannel	19	0.8759797	0.88	29.41
DelimitationContinentalShelf	13	0.5993545	0.60	30.01
Diallo	24	1.1065007	1.11	31.12
DiplomaticEnvoyUN	1	0.0461042	0.05	31.17
DiplomaticRelations	1	0.0461042	0.05	31.21
ELSI	6	0.2766252	0.28	31.49
EastTimor	10	0.4610420	0.46	31.95
ElectriciteBeyrouth	3	0.1383126	0.14	32.09
Fisheries	9	0.4149378	0.41	32.50
FisheriesJurisdiction	62	2.8584601	2.86	35.36
FrenchNationalsEgypt	1	0.0461042	0.05	35.41
FrontierDispute	23	1.0603965	1.06	36.47
GabcikovoNagymaros	16	0.7376671	0.74	37.21
GuardianshipInfantsConvention	11	0.5071462	0.51	37.71
GuatemalaTerritorialInsularMaritimeClaim	2	0.0922084	0.09	37.81
GulfOfMaine	12	0.5532503	0.55	38.36
HayaDeLaTorre	2	0.0922084	0.09	38.45
ICAOCouncil	14	0.6454587	0.65	39.10
ICAOCouncil-CICA	6	0.2766252	0.28	39.37
ICAOCouncil-IASTA	6	0.2766252	0.28	39.65
ICERD	16	0.7376671	0.74	40.39
ImmunitiesCriminalProceedings	26	1.1987091	1.20	41.59
ImmunitySRCommHR	6	0.2766252	0.28	41.86
IndependenceDeclarationKosovo	11	0.5071462	0.51	42.37
Interhandel	16	0.7376671	0.74	43.11
InterimAccord1995	8	0.3688336	0.37	43.48
IranianAssets	12	0.5532503	0.55	44.03
IslaPortillos	10	0.4610420	0.46	44.49
Jadhav	11	0.5071462	0.51	45.00

(continued)

shortname	N	exactpercent	roundedpercent	cumulpercent
Judgment2867ATILO-IFAD	5	0.2305210	0.23	45.23
JudgmentsCivilCommercialMatters	3	0.1383126	0.14	45.37
JurisdictionalImmunities	15	0.6915629	0.69	46.06
KasikiliSedudu	12	0.5532503	0.55	46.61
LaGrand	11	0.5071462	0.51	47.12
LandIslandMaritimeFrontier	20	0.9220839	0.92	48.04
LandIslandMaritimeFrontier-Revision	4	0.1844168	0.18	48.22
LandMaritimeBoundary	34	1.5675426	1.57	49.79
LandMaritimeBoundary-Interpretation	4	0.1844168	0.18	49.98
LandMaritimeDelimitationSovereigntyIslands	1	0.0461042	0.05	50.02
LegalityNuclearWeaponsArmedConflict	9	0.4149378	0.41	50.44
LegalityThreatUseNuclearWeapons	16	0.7376671	0.74	51.18
Lockerbie	54	2.4896266	2.49	53.67
MaritimeDelimitation	30	1.3831259	1.38	55.05
MaritimeDelimitation-BlackSea	4	0.1844168	0.18	55.23
MaritimeDelimitation-CaribbeanPacific	13	0.5993545	0.60	55.83
MaritimeDelimitation-GreenlandJanMayen	13	0.5993545	0.60	56.43
MaritimeDelimitation-IndianOcean	15	0.6915629	0.69	57.12
MaritimeDispute	13	0.5993545	0.60	57.72
MilitaryParamilitaryActivitiesNicaragua	32	1.4753343	1.48	59.20
MinquiersEcrehos	7	0.3227294	0.32	59.52
MonetaryGold	7	0.3227294	0.32	59.84
MutualAssistanceCriminalMatters	11	0.5071462	0.51	60.35
Namibia	16	0.7376671	0.74	61.09
NavigationalRights	6	0.2766252	0.28	61.36

(continued)

shortname	N	exactpercent	roundedpercent	cumulpercent
NorthSeaContinentalShelf	30	1.3831259	1.38	62.75
NorthernCameroons	18	0.8298755	0.83	63.58
NorwegianLoans	11	0.5071462	0.51	64.08
Nottebohm	11	0.5071462	0.51	64.59
NuclearDisarmament	51	2.3513140	2.35	66.94
NuclearTests	60	2.7662517	2.77	69.71
NuclearTests-ExaminationSituation	8	0.3688336	0.37	70.08
ObligationProsecuteExtradite	17	0.7837713	0.78	70.86
OilPlatforms	32	1.4753343	1.48	72.34
PassageGreatBelt	7	0.3227294	0.32	72.66
PassageIndianTerritory	25	1.1526049	1.15	73.81
PeaceTreaties	10	0.4610420	0.46	74.27
PedraBranca	9	0.4149378	0.41	74.69
PedraBranca-Interpretation	1	0.0461042	0.05	74.73
PedraBranca-Revision	1	0.0461042	0.05	74.78
PetitionersComitteeSouthWestAfrica	6	0.2766252	0.28	75.06
PortBeyrouthSRO	4	0.1844168	0.18	75.24
PulpMills	20	0.9220839	0.92	76.16
RelocationEmbassyUSJerusalem	1	0.0461042	0.05	76.21
ReparationUN	7	0.3227294	0.32	76.53
ReservationsGenocideConvention	4	0.1844168	0.18	76.72
ReviewJudgment158UNAT	11	0.5071462	0.51	77.22
ReviewJudgment273UNAT	11	0.5071462	0.51	77.73
ReviewJudgment333UNAT	10	0.4610420	0.46	78.19
SanJuanRiver	16	0.7376671	0.74	78.93
SilalaWaters	4	0.1844168	0.18	79.11
SouthWestAfrica	60	2.7662517	2.77	81.88
SovereignRightsCaribbeanSea	15	0.6915629	0.69	82.57

(continued)

shortname	N	exactpercent	roundedpercent	cumulpercent
SovereigntyFrontierLand	8	0.3688336	0.37	82.94
SovereigntyPulau	14	0.6454587	0.65	83.59
StatusSouthWestAfrica	7	0.3227294	0.32	83.91
TemplePreahVihear	16	0.7376671	0.74	84.65
TemplePreahVihear-Interpretation	14	0.6454587	0.65	85.29
TerritorialDispute	43	1.9824804	1.98	87.28
TerritorialDispute-CaribbeanSea	8	0.3688336	0.37	87.64
TransborderArmedActions	14	0.6454587	0.65	88.29
TreatmentAirCrew	2	0.0922084	0.09	88.38
TrialPakistaniPOW	5	0.2305210	0.23	88.61
USDiplomaticStaffTehran	7	0.3227294	0.32	88.93
USNationalsMorocco	6	0.2766252	0.28	89.21
UseOfForce	179	8.2526510	8.25	97.46
ViennaConventionConsularRelations	7	0.3227294	0.32	97.79
VotingProcedureSouthWestAfrica	5	0.2305210	0.23	98.02
WHO-EgyptAgreement	11	0.5071462	0.51	98.52
WesternSahara	15	0.6915629	0.69	99.22
WhalingAntarctic	17	0.7837713	0.78	100.00
Total	2169	100.0000000	100.00	100.00

Frequency Table for Variable: applicant

79 unique value(s) detected.

applicant	N	exactpercent	roundedpercent	cumulpercent
ARG	20	0.9220839	0.92	0.92
AUS	48	2.2130014	2.21	3.14
BEL	64	2.9506685	2.95	6.09

(continued)

applicant	N	exactpercent	roundedpercent	cumulpercent
BEN	7	0.3227294	0.32	6.41
BFA	16	0.7376671	0.74	7.15
BHR-EGY-ARE	6	0.2766252	0.28	7.42
BHR-EGY-SAU-ARE	6	0.2766252	0.28	7.70
BIH	49	2.2591056	2.26	9.96
BOL	14	0.6454587	0.65	10.60
BWA	12	0.5532503	0.55	11.16
CAN	12	0.5532503	0.55	11.71
CARAT	32	1.4753343	1.48	13.19
CHE	16	0.7376671	0.74	13.92
CHL	4	0.1844168	0.18	14.11
CMR	56	2.5818349	2.58	16.69
COD	83	3.8266482	3.83	20.52
COL	12	0.5532503	0.55	21.07
CRI	69	3.1811895	3.18	24.25
DEU	81	3.7344398	3.73	27.99
DJI	11	0.5071462	0.51	28.49
DMA	1	0.0461042	0.05	28.54
DNK	13	0.5993545	0.60	29.14
ECOSOC	11	0.5071462	0.51	29.64
ECU	4	0.1844168	0.18	29.83
ESP	13	0.5993545	0.60	30.43
ETH	30	1.3831259	1.38	31.81
FIN	7	0.3227294	0.32	32.13
FRA	32	1.4753343	1.48	33.61
GAB	1	0.0461042	0.05	33.66
GBR	77	3.5500231	3.55	37.21
GEO	16	0.7376671	0.74	37.94
GIN	24	1.1065007	1.11	39.05

(continued)

applicant	N	exactpercent	roundedpercent	cumulpercent
GMB	7	0.3227294	0.32	39.37
GNB	16	0.7376671	0.74	40.11
GNQ	26	1.1987091	1.20	41.31
GRC	36	1.6597510	1.66	42.97
GTM	2	0.0922084	0.09	43.06
GUY	9	0.4149378	0.41	43.48
HND	7	0.3227294	0.32	43.80
HRV	31	1.4292301	1.43	45.23
HUN	16	0.7376671	0.74	45.97
IDN	14	0.6454587	0.65	46.61
IFAD	5	0.2305210	0.23	46.84
IMO	4	0.1844168	0.18	47.03
IND	25	1.1526049	1.15	48.18
IRN	64	2.9506685	2.95	51.13
ISR	8	0.3688336	0.37	51.50
ITA	7	0.3227294	0.32	51.82
KHM	30	1.3831259	1.38	53.20
LBR	30	1.3831259	1.38	54.59
LBY	84	3.8727524	3.87	58.46
LIE	19	0.8759797	0.88	59.34
MEX	19	0.8759797	0.88	60.21
MHL	51	2.3513140	2.35	62.56
MKD	8	0.3688336	0.37	62.93
MYS	11	0.5071462	0.51	63.44
NIC	133	6.1318580	6.13	69.57
NLD	11	0.5071462	0.51	70.08
NRU	11	0.5071462	0.51	70.59
NZL	37	1.7058552	1.71	72.29
PAK	12	0.5532503	0.55	72.84

(continued)

applicant	N	exactpercent	roundedpercent	cumulpercent
PER	13	0.5993545	0.60	73.44
PRT	35	1.6136468	1.61	75.06
PRY	7	0.3227294	0.32	75.38
PSE	1	0.0461042	0.05	75.43
QAT	51	2.3513140	2.35	77.78
ROU	4	0.1844168	0.18	77.96
SCG	163	7.5149839	7.51	85.48
SLV	24	1.1065007	1.11	86.58
SOM	15	0.6915629	0.69	87.28
TLS	11	0.5071462	0.51	87.78
TUN	19	0.8759797	0.88	88.66
UKR	23	1.0603965	1.06	89.72
UNESCO	9	0.4149378	0.41	90.13
UNGA	141	6.5006916	6.50	96.63
UNSC	16	0.7376671	0.74	97.37
USA	21	0.9681881	0.97	98.34
WHO	20	0.9220839	0.92	99.26
YUG	16	0.7376671	0.74	100.00
Total	2169	100.0000000	100.00	100.00

Frequency Table for Variable: respondent

72 unique value(s) detected.

respondent	N	exactpercent	roundedpercent	cumulpercent
NA	238	10.9727985	10.97	10.97
ALB	19	0.8759797	0.88	11.85
ARE	22	1.0142923	1.01	12.86
ARG	1	0.0461042	0.05	12.91
AUS	32	1.4753343	1.48	14.38

(continued)

respondent	N	exactpercent	roundedpercent	cumulpercent
BDI	3	0.1383126	0.14	14.52
BEL	43	1.9824804	1.98	16.51
BGR	21	0.9681881	0.97	17.47
BHR	29	1.3370217	1.34	18.81
BLZ	2	0.0922084	0.09	18.90
BOL	4	0.1844168	0.18	19.09
BRA	1	0.0461042	0.05	19.13
CAN	34	1.5675426	1.57	20.70
CHE	4	0.1844168	0.18	20.89
CHL	28	1.2909175	1.29	22.18
COD	24	1.1065007	1.11	23.28
COL	67	3.0889811	3.09	26.37
CRI	19	0.8759797	0.88	27.25
CSK	1	0.0461042	0.05	27.29
DEU	27	1.2448133	1.24	28.54
DNK	22	1.0142923	1.01	29.55
EGY	1	0.0461042	0.05	29.60
ESP	52	2.3974182	2.40	32.00
FRA	135	6.2240664	6.22	38.22
FRA-GBR-USA	7	0.3227294	0.32	38.54
GBR	107	4.9331489	4.93	43.48
GNQ	1	0.0461042	0.05	43.52
GRC	8	0.3688336	0.37	43.89
GTM	11	0.5071462	0.51	44.40
HND	43	1.9824804	1.98	46.38
HUN	1	0.0461042	0.05	46.43
IND	54	2.4896266	2.49	48.92
IRN	19	0.8759797	0.88	49.79
ISL	49	2.2591056	2.26	52.05

(continued)

respondent	N	exactpercent	roundedpercent	cumulpercent
ITA	41	1.8902720	1.89	53.94
JPN	17	0.7837713	0.78	54.73
KEN	15	0.6915629	0.69	55.42
LBN	7	0.3227294	0.32	55.74
LBY	19	0.8759797	0.88	56.62
MLI	8	0.3688336	0.37	56.98
MLT	22	1.0142923	1.01	58.00
MMR	7	0.3227294	0.32	58.32
MYS	14	0.6454587	0.65	58.97
NAM	12	0.5532503	0.55	59.52
NER	15	0.6915629	0.69	60.21
NGA	38	1.7519594	1.75	61.96
NIC	75	3.4578147	3.46	65.42
NLD	44	2.0285846	2.03	67.45
NOR	33	1.5214385	1.52	68.97
PAK	42	1.9363762	1.94	70.91
PER	12	0.5532503	0.55	71.46
PRT	21	0.9681881	0.97	72.43
QAT	12	0.5532503	0.55	72.98
RUS	39	1.7980636	1.80	74.78
RWA	19	0.8759797	0.88	75.66
SCG	43	1.9824804	1.98	77.64
SEN	33	1.5214385	1.52	79.16
SGP	11	0.5071462	0.51	79.67
SRB	31	1.4292301	1.43	81.10
SUN	4	0.1844168	0.18	81.28
SVK	16	0.7376671	0.74	82.02
SWE	11	0.5071462	0.51	82.53
TCD	8	0.3688336	0.37	82.90

(continued)

respondent	N	exactpercent	roundedpercent	cumulpercent
THA	30	1.3831259	1.38	84.28
TUR	20	0.9220839	0.92	85.20
UGA	28	1.2909175	1.29	86.49
UKR	4	0.1844168	0.18	86.68
URY	20	0.9220839	0.92	87.60
USA	194	8.9442139	8.94	96.54
VEN	9	0.4149378	0.41	96.96
YUG	6	0.2766252	0.28	97.23
ZAF	60	2.7662517	2.77	100.00
Total	2169	100.0000000	100.00	100.00

Frequency Table for Variable: doctype

3 unique value(s) detected.

doctype	N	exactpercent	roundedpercent	cumulpercent
ADV	194	8.944214	8.94	8.94
JUD	1040	47.948363	47.95	56.89
ORD	935	43.107423	43.11	100.00
Total	2169	100.000000	100.00	100.00

Frequency Table for Variable: collision

3 unique value(s) detected.

collision	N	exactpercent	roundedpercent	cumulpercent
1	2152	99.2162287	99.22	99.22
2	16	0.7376671	0.74	99.95
3	1	0.0461042	0.05	100.00

(continued)

collision	N	exactpercent	roundedpercent	cumulpercent
Total	2169	100.0000000	100.00	100.00

Frequency Table for Variable: stage

5 unique value(s) detected.

stage	N	exactpercent	roundedpercent	cumulpercent
NA	1134	52.282158	52.28	52.28
CO	33	1.521438	1.52	53.80
IN	36	1.659751	1.66	55.46
ME	543	25.034578	25.03	80.50
PO	423	19.502075	19.50	100.00
Total	2169	100.000000	100.00	100.00

Frequency Table for Variable: opinion

15 unique value(s) detected.

opinion	N	exactpercent	roundedpercent	cumulpercent
0	765	35.2697095	35.27	35.27
1	254	11.7104657	11.71	46.98
2	227	10.4656524	10.47	57.45
3	200	9.2208391	9.22	66.67
4	170	7.8377132	7.84	74.50
5	147	6.7773167	6.78	81.28
6	123	5.6708160	5.67	86.95
7	95	4.3798986	4.38	91.33
8	71	3.2733979	3.27	94.61
9	57	2.6279391	2.63	97.23

(continued)

opinion	N	exactpercent	roundedpercent	cumulpercent
10	30	1.3831259	1.38	98.62
11	14	0.6454587	0.65	99.26
12	8	0.3688336	0.37	99.63
13	4	0.1844168	0.18	99.82
14	4	0.1844168	0.18	100.00
Total	2169	100.0000000	100.00	100.00

Frequency Table for Variable: language

1 unique value(s) detected.

language	N	exactpercent	roundedpercent	cumulpercent
EN	2169	100	100	100
Total	2169	100	100	100

Frequency Table for Variable: year

75 unique value(s) detected.

year	N	exactpercent	roundedpercent	cumulpercent
1947	3	0.1383126	0.14	0.14
1948	12	0.5532503	0.55	0.69
1949	24	1.1065007	1.11	1.80
1950	31	1.4292301	1.43	3.23
1951	21	0.9681881	0.97	4.20
1952	26	1.1987091	1.20	5.39
1953	11	0.5071462	0.51	5.90
1954	19	0.8759797	0.88	6.78
1955	11	0.5071462	0.51	7.28

(continued)

year	N	exactpercent	roundedpercent	cumulpercent
1956	22	1.0142923	1.01	8.30
1957	22	1.0142923	1.01	9.31
1958	28	1.2909175	1.29	10.60
1959	32	1.4753343	1.48	12.08
1960	26	1.1987091	1.20	13.28
1961	17	0.7837713	0.78	14.06
1962	42	1.9363762	1.94	16.00
1963	17	0.7837713	0.78	16.78
1964	15	0.6915629	0.69	17.47
1965	5	0.2305210	0.23	17.70
1966	24	1.1065007	1.11	18.81
1967	4	0.1844168	0.18	18.99
1968	5	0.2305210	0.23	19.23
1969	24	1.1065007	1.11	20.33
1970	14	0.6454587	0.65	20.98
1971	16	0.7376671	0.74	21.72
1972	23	1.0603965	1.06	22.78
1973	62	2.8584601	2.86	25.63
1974	52	2.3974182	2.40	28.03
1975	15	0.6915629	0.69	28.72
1976	11	0.5071462	0.51	29.23
1977	1	0.0461042	0.05	29.28
1978	8	0.3688336	0.37	29.64
1979	3	0.1383126	0.14	29.78
1980	16	0.7376671	0.74	30.52
1981	8	0.3688336	0.37	30.89
1982	24	1.1065007	1.11	32.00
1983	2	0.0922084	0.09	32.09
1984	35	1.6136468	1.61	33.70

(continued)

year	N	exactpercent	roundedpercent	cumulpercent
1985	18	0.8298755	0.83	34.53
1986	17	0.7837713	0.78	35.32
1987	17	0.7837713	0.78	36.10
1988	14	0.6454587	0.65	36.75
1989	21	0.9681881	0.97	37.71
1990	15	0.6915629	0.69	38.40
1991	23	1.0603965	1.06	39.47
1992	43	1.9824804	1.98	41.45
1993	29	1.3370217	1.34	42.78
1994	14	0.6454587	0.65	43.43
1995	29	1.3370217	1.34	44.77
1996	55	2.5357308	2.54	47.30
1997	19	0.8759797	0.88	48.18
1998	61	2.8123559	2.81	50.99
1999	135	6.2240664	6.22	57.22
2000	37	1.7058552	1.71	58.92
2001	44	2.0285846	2.03	60.95
2002	49	2.2591056	2.26	63.21
2003	35	1.6136468	1.61	64.82
2004	78	3.5961272	3.60	68.42
2005	21	0.9681881	0.97	69.39
2006	18	0.8298755	0.83	70.22
2007	39	1.7980636	1.80	72.01
2008	42	1.9363762	1.94	73.95
2009	20	0.9220839	0.92	74.87
2010	42	1.9363762	1.94	76.81
2011	57	2.6279391	2.63	79.44
2012	36	1.6597510	1.66	81.10
2013	33	1.5214385	1.52	82.62

(continued)

year	N	exactpercent	roundedpercent	cumulpercent
2014	40	1.8441678	1.84	84.46
2015	52	2.3974182	2.40	86.86
2016	75	3.4578147	3.46	90.32
2017	35	1.6136468	1.61	91.93
2018	64	2.9506685	2.95	94.88
2019	54	2.4896266	2.49	97.37
2020	35	1.6136468	1.61	98.99
2021	22	1.0142923	1.01	100.00
Total	2169	100.0000000	100.00	100.00

Frequency Table for Variable: minority

2 unique value(s) detected.

minority	N	exactpercent	roundedpercent	cumulpercent
0	765	35.26971	35.27	35.27
1	1404	64.73029	64.73	100.00
Total	2169	100.00000	100.00	100.00

Frequency Table for Variable: fullname

178 unique value(s) detected.

fullname	N	exactpercent	roundedpercent	cumulpercent
Accordance with international law of the unilateral declaration of independence in respect of Kosovo	11	0.5071462	0.51	0.51
Admissibility of Hearings of Petitioners by the Committee on South West Africa	6	0.2766252	0.28	0.78

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Aegean Sea Continental Shelf (Greece v. Turkey)	20	0.9220839	0.92	1.71
Aerial Herbicide Spraying (Ecuador v. Colombia)	4	0.1844168	0.18	1.89
Aerial Incident of 10 August 1999 (Pakistan v. India)	7	0.3227294	0.32	2.21
Aerial Incident of 10 March 1953 (United States of America v. Czechoslovakia)	1	0.0461042	0.05	2.26
Aerial Incident of 27 July 1955 (Israel v. Bulgaria)	8	0.3688336	0.37	2.63
Aerial Incident of 27 July 1955 (United Kingdom v. Bulgaria)	5	0.2305210	0.23	2.86
Aerial Incident of 27 July 1955 (United States of America v. Bulgaria)	8	0.3688336	0.37	3.23
Aerial Incident of 3 July 1988 (Islamic Republic of Iran v. United States of America)	9	0.4149378	0.41	3.64
Aerial Incident of 4 September 1954 (United States of America v. Union of Soviet Socialist Republics)	1	0.0461042	0.05	3.69
Aerial Incident of 7 November 1954 (United States of America v. Union of Soviet Socialist Republics)	1	0.0461042	0.05	3.73
Aerial Incident of 7 October 1952 (United States of America v. Union of Soviet Socialist Republics)	1	0.0461042	0.05	3.78
Ahmadou Sadio Diallo (Republic of Guinea v. Democratic Republic of the Congo)	24	1.1065007	1.11	4.89
Alleged Violations of Sovereign Rights and Maritime Spaces in the Caribbean Sea (Nicaragua v. Colombia)	15	0.6915629	0.69	5.58

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Alleged violations of the 1955 Treaty of Amity, Economic Relations, and Consular Rights (Islamic Republic of Iran v. United States of America)	11	0.5071462	0.51	6.09
Ambatielos (Greece v. United Kingdom)	16	0.7376671	0.74	6.82
Anglo-Iranian Oil Co. (United Kingdom v. Iran)	12	0.5532503	0.55	7.38
Antarctica (United Kingdom v. Argentina)	1	0.0461042	0.05	7.42
Antarctica (United Kingdom v. Chile)	1	0.0461042	0.05	7.47
Appeal Relating to the Jurisdiction of the ICAO Council (India v. Pakistan)	14	0.6454587	0.65	8.11
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	0.2766252	0.28	8.39
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	0.2766252	0.28	8.67
Applicability of Article VI, Section 22, of the Convention on the Privileges and Immunities of the United Nations	5	0.2305210	0.23	8.90
Applicability of the Obligation to Arbitrate under Section 21 of the United Nations Headquarters Agreement of 26 June 1947	6	0.2766252	0.28	9.17

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Application for Review of Judgment No. 158 of the United Nations Administrative Tribunal	11	0.5071462	0.51	9.68
Application for Review of Judgment No. 273 of the United Nations Administrative Tribunal	11	0.5071462	0.51	10.19
Application for Review of Judgment No. 333 of the United Nations Administrative Tribunal	10	0.4610420	0.46	10.65
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	0.2305210	0.23	10.88
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	0.2766252	0.28	11.16
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	0.1844168	0.18	11.34

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
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	0.0461042	0.05	11.39
Application of the Convention of 1902 Governing the Guardianship of Infants (Netherlands v. Sweden)	11	0.5071462	0.51	11.89
Application of the Convention on the Prevention and Punishment of the Crime of Genocide (Bosnia and Herzegovina v. Serbia and Montenegro)	43	1.9824804	1.98	13.88
Application of the Convention on the Prevention and Punishment of the Crime of Genocide (Croatia v. Serbia)	31	1.4292301	1.43	15.31
Application of the Convention on the Prevention and Punishment of the Crime of Genocide (The Gambia v. Myanmar)	7	0.3227294	0.32	15.63
Application of the Interim Accord of 13 September 1995 (the former Yugoslav Republic of Macedonia v. Greece)	8	0.3688336	0.37	16.00
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)	23	1.0603965	1.06	17.06

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Application of the International Convention on the Elimination of All Forms of Racial Discrimination (Georgia v. Russian Federation)	16	0.7376671	0.74	17.80
Application of the International Convention on the Elimination of All Forms of Racial Discrimination (Qatar v. United Arab Emirates)	22	1.0142923	1.01	18.81
Arbitral Award Made by the King of Spain on 23 December 1906 (Honduras v. Nicaragua)	6	0.2766252	0.28	19.09
Arbitral Award of 3 October 1899 (Guyana v. Venezuela)	9	0.4149378	0.41	19.50
Arbitral Award of 31 July 1989 (Guinea-Bissau v. Senegal)	15	0.6915629	0.69	20.19
Armed Activities on the Territory of the Congo (Democratic Republic of the Congo v. Burundi)	3	0.1383126	0.14	20.33
Armed Activities on the Territory of the Congo (Democratic Republic of the Congo v. Rwanda)	3	0.1383126	0.14	20.47
Armed Activities on the Territory of the Congo (Democratic Republic of the Congo v. Uganda)	28	1.2909175	1.29	21.76
Armed Activities on the Territory of the Congo (New Application: 2002) (Democratic Republic of the Congo v. Rwanda)	16	0.7376671	0.74	22.50
Arrest Warrant of 11 April 2000 (Democratic Republic of the Congo v. Belgium)	22	1.0142923	1.01	23.51
Asylum (Colombia v. Peru)	9	0.4149378	0.41	23.93

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Avena and Other Mexican Nationals (Mexico v. United States of America)	11	0.5071462	0.51	24.44
Barcelona Traction, Light and Power Company, Limited (Belgium v. Spain)	5	0.2305210	0.23	24.67
Barcelona Traction, Light and Power Company, Limited (Belgium v. Spain) (New Application: 1962)	31	1.4292301	1.43	26.09
Border and Transborder Armed Actions (Nicaragua v. Costa Rica)	3	0.1383126	0.14	26.23
Border and Transborder Armed Actions (Nicaragua v. Honduras)	11	0.5071462	0.51	26.74
Certain Activities Carried Out by Nicaragua in the Border Area (Costa Rica v. Nicaragua)	40	1.8441678	1.84	28.58
Certain Criminal Proceedings in France (Republic of the Congo v. France)	11	0.5071462	0.51	29.09
Certain Expenses of the United Nations (Article 17, paragraph 2, of the Charter)	11	0.5071462	0.51	29.60
Certain Iranian Assets (Islamic Republic of Iran v. United States of America)	12	0.5532503	0.55	30.15
Certain Norwegian Loans (France v. Norway)	11	0.5071462	0.51	30.66
Certain Phosphate Lands in Nauru (Nauru v. Australia)	11	0.5071462	0.51	31.17
Certain Property (Liechtenstein v. Germany)	8	0.3688336	0.37	31.54
Certain Questions concerning Diplomatic Relations (Honduras v. Brazil)	1	0.0461042	0.05	31.58

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Certain Questions of Mutual Assistance in Criminal Matters (Djibouti v. France)	11	0.5071462	0.51	32.09
Compagnie du Port, des Quais et des Entrepôts de Beyrouth and Société Radio-Orient (France v. Lebanon)	4	0.1844168	0.18	32.27
Competence of the General Assembly for the Admission of a State to the United Nations	4	0.1844168	0.18	32.46
Conditions of Admission of a State to Membership in the United Nations (Article 4 of the Charter)	7	0.3227294	0.32	32.78
Constitution of the Maritime Safety Committee of the Inter-Governmental Maritime Consultative Organization	4	0.1844168	0.18	32.96
Construction of a Road in Costa Rica along the San Juan River (Nicaragua v. Costa Rica)	16	0.7376671	0.74	33.70
Continental Shelf (Libyan Arab Jamahiriya/Malta)	22	1.0142923	1.01	34.72
Continental Shelf (Tunisia/Libyan Arab Jamahiriya)	14	0.6454587	0.65	35.36
Corfu Channel (United Kingdom of Great Britain and Northern Ireland v. Albania)	19	0.8759797	0.88	36.24
Delimitation of the Maritime Boundary in the Gulf of Maine Area (Canada/United States of America)	12	0.5532503	0.55	36.79
Difference Relating to Immunity from Legal Process of a Special Rapporteur of the Commission on Human Rights	6	0.2766252	0.28	37.07

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Dispute over the Status and Use of the Waters of the Silala (Chile v. Bolivia)	4	0.1844168	0.18	37.25
Dispute regarding Navigational and Related Rights (Costa Rica v. Nicaragua)	6	0.2766252	0.28	37.53
East Timor (Portugal v. Australia)	10	0.4610420	0.46	37.99
Effect of Awards of Compensation Made by the United Nations Administrative Tribunal	6	0.2766252	0.28	38.27
Electricité de Beyrouth Company (France v. Lebanon)	3	0.1383126	0.14	38.40
Elettronica Sicula S.p.A. (ELSI) (United States of America v. Italy)	6	0.2766252	0.28	38.68
Fisheries (United Kingdom v. Norway)	9	0.4149378	0.41	39.10
Fisheries Jurisdiction (Federal Republic of Germany v. Iceland)	25	1.1526049	1.15	40.25
Fisheries Jurisdiction (Spain v. Canada)	13	0.5993545	0.60	40.85
Fisheries Jurisdiction (United Kingdom v. Iceland)	24	1.1065007	1.11	41.95
Frontier Dispute (Benin/Niger)	7	0.3227294	0.32	42.28
Frontier Dispute (Burkina Faso/Niger)	8	0.3688336	0.37	42.65
Frontier Dispute (Burkina Faso/Republic of Mali)	8	0.3688336	0.37	43.02
Gabčíkovo-Nagymaros Project (Hungary/Slovakia)	16	0.7376671	0.74	43.75
Guatemala's Territorial, Insular and Maritime Claim (Guatemala/Belize)	2	0.0922084	0.09	43.85

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Haya de la Torre (Colombia v. Peru)	2	0.0922084	0.09	43.94
Immunities and Criminal Proceedings (Equatorial Guinea v. France)	26	1.1987091	1.20	45.14
Interhandel (Switzerland v. United States of America)	16	0.7376671	0.74	45.87
International Status of South West Africa	7	0.3227294	0.32	46.20
Interpretation of Peace Treaties with Bulgaria, Hungary and Romania	10	0.4610420	0.46	46.66
Interpretation of the Agreement of 25 March 1951 between the WHO and Egypt	11	0.5071462	0.51	47.16
Jadhav (India v. Pakistan)	11	0.5071462	0.51	47.67
Judgment No.2867 of the Administrative Tribunal of the International Labour Organization upon a Complaint Filed against the International Fund for Agricultural Development	5	0.2305210	0.23	47.90
Judgments of the Administrative Tribunal of the ILO upon Complaints Made against UNESCO	9	0.4149378	0.41	48.32
Jurisdiction and Enforcement of Judgments in Civil and Commercial Matters (Belgium v. Switzerland)	3	0.1383126	0.14	48.46
Jurisdictional Immunities of the State (Germany v. Italy: Greece intervening)	15	0.6915629	0.69	49.15
Kasikili/Sedudu Island (Botswana/Namibia)	12	0.5532503	0.55	49.70
LaGrand (Germany v. United States of America)	11	0.5071462	0.51	50.21

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Land Boundary in the Northern Part of Isla Portillos (Costa Rica v. Nicaragua)	10	0.4610420	0.46	50.67
Land and Maritime Boundary between Cameroon and Nigeria (Cameroon v. Nigeria: Equatorial Guinea intervening)	34	1.5675426	1.57	52.24
Land and Maritime Delimitation and Sovereignty over Islands (Gabon/Equatorial Guinea)	1	0.0461042	0.05	52.28
Land, Island and Maritime Frontier Dispute (El Salvador/Honduras: Nicaragua intervening)	20	0.9220839	0.92	53.20
Legal Consequences for States of the Continued Presence of South Africa in Namibia (South West Africa) notwithstanding Security Council Resolution 276 (1970)	16	0.7376671	0.74	53.94
Legal Consequences of the Construction of a Wall in the Occupied Palestinian Territory	11	0.5071462	0.51	54.45
Legal Consequences of the Separation of the Chagos Archipelago from Mauritius in 1965	15	0.6915629	0.69	55.14
Legality of Use of Force (Serbia and Montenegro v. Belgium)	21	0.9681881	0.97	56.11
Legality of Use of Force (Serbia and Montenegro v. Canada)	21	0.9681881	0.97	57.08
Legality of Use of Force (Serbia and Montenegro v. France)	19	0.8759797	0.88	57.95

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Legality of Use of Force (Serbia and Montenegro v. Germany)	19	0.8759797	0.88	58.83
Legality of Use of Force (Serbia and Montenegro v. Italy)	20	0.9220839	0.92	59.75
Legality of Use of Force (Serbia and Montenegro v. Netherlands)	21	0.9681881	0.97	60.72
Legality of Use of Force (Serbia and Montenegro v. Portugal)	21	0.9681881	0.97	61.69
Legality of Use of Force (Serbia and Montenegro v. United Kingdom)	21	0.9681881	0.97	62.66
Legality of Use of Force (Yugoslavia v. Spain)	9	0.4149378	0.41	63.07
Legality of Use of Force (Yugoslavia v. United States of America)	7	0.3227294	0.32	63.39
Legality of the Threat or Use of Nuclear Weapons	16	0.7376671	0.74	64.13
Legality of the Use by a State of Nuclear Weapons in Armed Conflict	9	0.4149378	0.41	64.55
Maritime Delimitation and Territorial Questions between Qatar and Bahrain (Qatar v. Bahrain)	29	1.3370217	1.34	65.88
Maritime Delimitation between Guinea-Bissau and Senegal (Guinea-Bissau v. Senegal)	1	0.0461042	0.05	65.93
Maritime Delimitation in the Area between Greenland and Jan Mayen (Denmark v. Norway)	13	0.5993545	0.60	66.53

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Maritime Delimitation in the Black Sea (Romania v. Ukraine)	4	0.1844168	0.18	66.71
Maritime Delimitation in the Caribbean Sea and the Pacific Ocean (Costa Rica v. Nicaragua)	13	0.5993545	0.60	67.31
Maritime Delimitation in the Indian Ocean (Somalia v. Kenya)	15	0.6915629	0.69	68.00
Maritime Dispute (Peru v. Chile)	13	0.5993545	0.60	68.60
Military and Paramilitary Activities in and against Nicaragua (Nicaragua v. United States of America)	32	1.4753343	1.48	70.08
Minquiers and Ecrehos (France/United Kingdom)	7	0.3227294	0.32	70.40
Monetary Gold Removed from Rome in 1943 (Italy v. France, United Kingdom of Great Britain and Northern Ireland and United States of America)	7	0.3227294	0.32	70.72
North Sea Continental Shelf (Federal Republic of Germany/Denmark)	15	0.6915629	0.69	71.42
North Sea Continental Shelf (Federal Republic of Germany/Netherlands)	15	0.6915629	0.69	72.11
Northern Cameroons (Cameroon v. United Kingdom)	18	0.8298755	0.83	72.94
Nottebohm (Liechtenstein v. Guatemala)	11	0.5071462	0.51	73.44
Nuclear Tests (Australia v. France)	31	1.4292301	1.43	74.87
Nuclear Tests (New Zealand v. France)	29	1.3370217	1.34	76.21

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Obligation to Negotiate Access to the Pacific Ocean (Bolivia v. Chile)	14	0.6454587	0.65	76.86
Obligations concerning Negotiations relating to Cessation of the Nuclear Arms Race and to Nuclear Disarmament (Marshall Islands v. India)	17	0.7837713	0.78	77.64
Obligations concerning Negotiations relating to Cessation of the Nuclear Arms Race and to Nuclear Disarmament (Marshall Islands v. Pakistan)	17	0.7837713	0.78	78.42
Obligations concerning Negotiations relating to Cessation of the Nuclear Arms Race and to Nuclear Disarmament (Marshall Islands v. United Kingdom)	17	0.7837713	0.78	79.21
Oil Platforms (Islamic Republic of Iran v. United States of America)	32	1.4753343	1.48	80.68
Passage through the Great Belt (Finland v. Denmark)	7	0.3227294	0.32	81.01
Protection of French Nationals and Protected Persons in Egypt (France v. Egypt)	1	0.0461042	0.05	81.05
Pulp Mills on the River Uruguay (Argentina v. Uruguay)	20	0.9220839	0.92	81.97
Question of the Delimitation of the Continental Shelf between Nicaragua and Colombia beyond 200 nautical miles from the Nicaraguan Coast (Nicaragua v. Colombia)	13	0.5993545	0.60	82.57

(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 Kingdom)	28	1.2909175	1.29	83.86
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	1.1987091	1.20	85.06
Questions relating to the Obligation to Prosecute or Extradite (Belgium v. Senegal)	17	0.7837713	0.78	85.85
Questions relating to the Seizure and Detention of Certain Documents and Data (Timor-Leste v. Australia)	11	0.5071462	0.51	86.35
Relocation of the United States Embassy to Jerusalem (Palestine v. United States of America)	1	0.0461042	0.05	86.40
Reparation for Injuries Suffered in the Service of the United Nations	7	0.3227294	0.32	86.72
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	0.1844168	0.18	86.91

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
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	0.6454587	0.65	87.55
Request for Interpretation of the Judgment of 20 November 1950 in the Asylum Case (Colombia v. Peru)	1	0.0461042	0.05	87.60
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	0.0461042	0.05	87.64
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	0.3688336	0.37	88.01
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	0.3688336	0.37	88.38
Reservations to the Convention on the Prevention and Punishment of the Crime of Genocide	4	0.1844168	0.18	88.57
Right of Passage over Indian Territory (Portugal v. India)	25	1.1526049	1.15	89.72

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Rights of Nationals of the United States of America in Morocco (France v. United States of America)	6	0.2766252	0.28	90.00
South West Africa (Ethiopia v. South Africa)	30	1.3831259	1.38	91.38
South West Africa (Liberia v. South Africa)	30	1.3831259	1.38	92.76
Sovereignty over Certain Frontier Land (Belgium/Netherlands)	8	0.3688336	0.37	93.13
Sovereignty over Pedra Branca/Pulau Batu Puteh, Middle Rocks and South Ledge (Malaysia/Singapore)	9	0.4149378	0.41	93.55
Sovereignty over Pulau Ligitan and Pulau Sipadan (Indonesia/Malaysia)	14	0.6454587	0.65	94.19
Status vis-à-vis the Host State of a Diplomatic Envoy to the United Nations (Commonwealth of Dominica v. Switzerland)	1	0.0461042	0.05	94.24
Temple of Preah Vihear (Cambodia v. Thailand)	16	0.7376671	0.74	94.97
Territorial Dispute (Libyan Arab Jamahiriya/Chad)	8	0.3688336	0.37	95.34
Territorial and Maritime Dispute (Nicaragua v. Colombia)	35	1.6136468	1.61	96.96
Territorial and Maritime Dispute between Nicaragua and Honduras in the Caribbean Sea (Nicaragua v. Honduras)	8	0.3688336	0.37	97.33
Treatment in Hungary of Aircraft and Crew of United States of America (United States of America v. Hungarian People's Republic)	1	0.0461042	0.05	97.37

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Treatment in Hungary of Aircraft and Crew of United States of America (United States of America v. Union of Soviet Socialist Republics)	1	0.0461042	0.05	97.42
Trial of Pakistani Prisoners of War (Pakistan v. India)	5	0.2305210	0.23	97.65
United States Diplomatic and Consular Staff in Tehran (United States of America v. Iran)	7	0.3227294	0.32	97.97
Vienna Convention on Consular Relations (Paraguay v. United States of America)	7	0.3227294	0.32	98.29
Voting Procedure on Questions relating to Reports and Petitions concerning the Territory of South West Africa	5	0.2305210	0.23	98.52
Western Sahara	15	0.6915629	0.69	99.22
Whaling in the Antarctic (Australia v. Japan: New Zealand intervening)	17	0.7837713	0.78	100.00
Total	2169	100.0000000	100.00	100.00

Frequency Table for Variable: applicant_region

8 unique value(s) detected.

applicant_region	N	exactpercent	roundedpercent	cumulpercent
Africa	437	20.1475334	20.15	20.15
Americas	371	17.1046565	17.10	37.25
Asia	243	11.2033195	11.20	48.46
Asia|Africa|Asia	6	0.2766252	0.28	48.73
Asia|Africa|Asia|Asia	6	0.2766252	0.28	49.01
Europe	721	33.2411249	33.24	82.25

(continued)

applicant_region	N	exactpercent	roundedpercent	cumulpercent
NA	238	10.9727985	10.97	93.22
Oceania	147	6.7773167	6.78	100.00
Total	2169	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	10.9727985	10.97	10.97
Africa	284	13.0935915	13.09	24.07
Americas	520	23.9741817	23.97	48.04
Asia	284	13.0935915	13.09	61.13
Europe	804	37.0677732	37.07	98.20
Europe|Europe|Americas	7	0.3227294	0.32	98.52
Oceania	32	1.4753343	1.48	100.00
Total	2169	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	3.9188566	3.92	3.92
Eastern Europe	43	1.9824804	1.98	5.90
Latin America and the Caribbean	338	15.5832181	15.58	21.48
Micronesia	62	2.8584601	2.86	24.34
NA	238	10.9727985	10.97	35.32
Northern Africa	103	4.7487321	4.75	40.06

(continued)

applicant_subregion	N	exactpercent	roundedpercent	cumulpercent
Northern America	33	1.5214385	1.52	41.59
Northern Europe	97	4.4721070	4.47	46.06
South-eastern Asia	66	3.0428769	3.04	49.10
Southern Asia	101	4.6565237	4.66	53.76
Southern Europe	358	16.5053020	16.51	70.26
Sub-Saharan Africa	334	15.3988013	15.40	85.66
Western Asia	76	3.5039189	3.50	89.17
Western Asia|Northern Africa|Western Asia	6	0.2766252	0.28	89.44
Western Asia|Northern Africa|Western Asia|Western Asia	6	0.2766252	0.28	89.72
Western Europe	223	10.2812356	10.28	100.00
Total	2169	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	10.9727985	10.97	10.97
Australia and New Zealand	32	1.4753343	1.48	12.45
Eastern Asia	17	0.7837713	0.78	13.23
Eastern Europe	86	3.9649608	3.96	17.20
Latin America and the Caribbean	292	13.4624251	13.46	30.66
Northern Africa	20	0.9220839	0.92	31.58
Northern America	228	10.5117566	10.51	42.09
Northern Europe	222	10.2351314	10.24	52.33
South-eastern Asia	62	2.8584601	2.86	55.19
Southern Asia	115	5.3019825	5.30	60.49

(continued)

respondent_subregion	N	exactpercent	roundedpercent	cumulpercent
Southern Europe	243	11.2033195	11.20	71.69
Sub-Saharan Africa	264	12.1715076	12.17	83.86
Western Asia	90	4.1493776	4.15	88.01
Western Europe	253	11.6643615	11.66	99.68
Western Europe|Northern Europe|Northern America	7	0.3227294	0.32	100.00
Total	2169	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	2169	100	100	100
Total	2169	100	100	100

Frequency Table for Variable: doi_version

1 unique value(s) detected.

doi_version	N	exactpercent	roundedpercent	cumulpercent
10.5281/zenodo.3826445	2169	100	100	100
Total	2169	100	100	100

Frequency Table for Variable: version

1 unique value(s) detected.

version	N	exactpercent	roundedpercent	cumulpercent
2021-11-23	2169	100	100	100
Total	2169	100	100	100

Frequency Table for Variable: license

1 unique value(s) detected.

license	N	exactpercent	roundedpercent	cumulpercent
Creative Commons Zero 1.0 Universal	2169	100	100	100
Total	2169	100	100	100

21.3 French Corpus

21.3.1 Variables to Ignore

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

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

21.3.2 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"
```

21.3.3 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	2160	100	100	100
Total	2160	100	100	100

Frequency Table for Variable: caseno

178 unique value(s) detected.

caseno	N	exactpercent	roundedpercent	cumulpercent
1	19	0.8796296	0.88	0.88
3	7	0.3240741	0.32	1.20
4	7	0.3240741	0.32	1.53
5	9	0.4166667	0.42	1.94
6	1	0.0462963	0.05	1.99
7	9	0.4166667	0.42	2.41
8	10	0.4629630	0.46	2.87
9	4	0.1851852	0.19	3.06
10	7	0.3240741	0.32	3.38
11	6	0.2777778	0.28	3.66
12	4	0.1851852	0.19	3.84
13	1	0.0462963	0.05	3.89
14	2	0.0925926	0.09	3.98
15	16	0.7407407	0.74	4.72
16	12	0.5555556	0.56	5.28
17	7	0.3240741	0.32	5.60
18	11	0.5092593	0.51	6.11
19	7	0.3240741	0.32	6.44
20	3	0.1388889	0.14	6.57

(continued)

caseno	N	exactpercent	roundedpercent	cumulpercent
21	6	0.2777778	0.28	6.85
22	1	0.0462963	0.05	6.90
23	1	0.0462963	0.05	6.94
24	5	0.2314815	0.23	7.18
25	1	0.0462963	0.05	7.22
26	1	0.0462963	0.05	7.27
27	1	0.0462963	0.05	7.31
28	1	0.0462963	0.05	7.36
29	11	0.5092593	0.51	7.87
30	9	0.4166667	0.42	8.29
31	6	0.2777778	0.28	8.56
32	25	1.1574074	1.16	9.72
33	11	0.5092593	0.51	10.23
34	16	0.7407407	0.74	10.97
35	8	0.3703704	0.37	11.34
36	8	0.3703704	0.37	11.71
37	5	0.2314815	0.23	11.94
38	8	0.3703704	0.37	12.31
39	6	0.2777778	0.28	12.59
40	1	0.0462963	0.05	12.64
41	5	0.2314815	0.23	12.87
42	4	0.1851852	0.19	13.06
43	4	0.1851852	0.19	13.24
44	1	0.0462963	0.05	13.29
45	16	0.7407407	0.74	14.03
46	30	1.3888889	1.39	15.42
47	30	1.3888889	1.39	16.81
48	18	0.8333333	0.83	17.64
49	11	0.5092593	0.51	18.15

(continued)

caseno	N	exactpercent	roundedpercent	cumulpercent
50	31	1.4351852	1.44	19.58
51	15	0.6944444	0.69	20.28
52	15	0.6944444	0.69	20.97
53	16	0.7407407	0.74	21.71
54	14	0.6481481	0.65	22.36
55	24	1.1111111	1.11	23.47
56	25	1.1574074	1.16	24.63
57	11	0.5092593	0.51	25.14
58	31	1.4351852	1.44	26.57
59	29	1.3425926	1.34	27.92
60	5	0.2314815	0.23	28.15
61	15	0.6944444	0.69	28.84
62	20	0.9259259	0.93	29.77
63	14	0.6481481	0.65	30.42
64	7	0.3240741	0.32	30.74
65	11	0.5092593	0.51	31.25
66	11	0.5092593	0.51	31.76
67	12	0.5555556	0.56	32.31
68	22	1.0185185	1.02	33.33
69	8	0.3703704	0.37	33.70
70	32	1.4814815	1.48	35.19
71	5	0.2314815	0.23	35.42
72	10	0.4629630	0.46	35.88
73	3	0.1388889	0.14	36.02
74	11	0.5092593	0.51	36.53
75	20	0.9259259	0.93	37.45
76	6	0.2777778	0.28	37.73
77	6	0.2777778	0.28	38.01
78	13	0.6018519	0.60	38.61

(continued)

caseno	N	exactpercent	roundedpercent	cumulpercent
79	9	0.4166667	0.42	39.03
80	11	0.5092593	0.51	39.54
81	5	0.2314815	0.23	39.77
82	15	0.6944444	0.69	40.46
83	8	0.3703704	0.37	40.83
84	10	0.4629630	0.46	41.30
85	1	0.0462963	0.05	41.34
86	7	0.3240741	0.32	41.67
87	29	1.3425926	1.34	43.01
88	28	1.2962963	1.30	44.31
89	25	1.1574074	1.16	45.46
90	32	1.4814815	1.48	46.94
91	43	1.9907407	1.99	48.94
92	16	0.7407407	0.74	49.68
93	9	0.4166667	0.42	50.09
94	34	1.5740741	1.57	51.67
95	16	0.7407407	0.74	52.41
96	13	0.6018519	0.60	53.01
97	8	0.3703704	0.37	53.38
98	12	0.5555556	0.56	53.94
99	7	0.3240741	0.32	54.26
100	6	0.2777778	0.28	54.54
101	4	0.1851852	0.19	54.72
102	14	0.6481481	0.65	55.37
103	24	1.1111111	1.11	56.48
104	11	0.5092593	0.51	56.99
105	21	0.9722222	0.97	57.96
106	21	0.9722222	0.97	58.94
107	19	0.8796296	0.88	59.81

(continued)

caseno	N	exactpercent	roundedpercent	cumulpercent
108	19	0.8796296	0.88	60.69
109	20	0.9259259	0.93	61.62
110	21	0.9722222	0.97	62.59
111	21	0.9722222	0.97	63.56
112	9	0.4166667	0.42	63.98
113	21	0.9722222	0.97	64.95
114	7	0.3240741	0.32	65.28
115	3	0.1388889	0.14	65.42
116	28	1.2962963	1.30	66.71
117	3	0.1388889	0.14	66.85
118	31	1.4351852	1.44	68.29
119	7	0.3240741	0.32	68.61
120	8	0.3703704	0.37	68.98
121	22	1.0185185	1.02	70.00
122	6	0.2777778	0.28	70.28
123	8	0.3703704	0.37	70.65
124	35	1.6203704	1.62	72.27
125	6	0.2777778	0.28	72.55
126	16	0.7407407	0.74	73.29
127	4	0.1851852	0.19	73.47
128	11	0.5092593	0.51	73.98
129	11	0.5092593	0.51	74.49
130	9	0.4166667	0.42	74.91
131	11	0.5092593	0.51	75.42
132	4	0.1851852	0.19	75.60
133	6	0.2777778	0.28	75.88
134	1	0.0462963	0.05	75.93
135	20	0.9259259	0.93	76.85
136	11	0.5092593	0.51	77.36

(continued)

caseno	N	exactpercent	roundedpercent	cumulpercent
137	13	0.6018519	0.60	77.96
138	4	0.1851852	0.19	78.15
139	8	0.3703704	0.37	78.52
140	16	0.7407407	0.74	79.26
141	11	0.5092593	0.51	79.77
142	8	0.3703704	0.37	80.14
143	15	0.6944444	0.69	80.83
144	17	0.7870370	0.79	81.62
145	3	0.1388889	0.14	81.76
146	5	0.2314815	0.23	81.99
147	1	0.0462963	0.05	82.04
148	17	0.7870370	0.79	82.82
149	8	0.3703704	0.37	83.19
150	40	1.8518519	1.85	85.05
151	14	0.6481481	0.65	85.69
152	16	0.7407407	0.74	86.44
153	14	0.6481481	0.65	87.08
154	13	0.6018519	0.60	87.69
155	15	0.6944444	0.69	88.38
156	10	0.4629630	0.46	88.84
157	13	0.6018519	0.60	89.44
158	17	0.7870370	0.79	90.23
159	17	0.7870370	0.79	91.02
160	17	0.7870370	0.79	91.81
161	13	0.6018519	0.60	92.41
162	4	0.1851852	0.19	92.59
163	26	1.2037037	1.20	93.80
164	12	0.5555556	0.56	94.35
165	10	0.4629630	0.46	94.81

(continued)

caseno	N	exactpercent	roundedpercent	cumulpercent
166	23	1.0648148	1.06	95.88
167	1	0.0462963	0.05	95.93
168	11	0.5092593	0.51	96.44
169	15	0.6944444	0.69	97.13
170	1	0.0462963	0.05	97.18
171	9	0.4166667	0.42	97.59
172	18	0.8333333	0.83	98.43
173	6	0.2777778	0.28	98.70
174	6	0.2777778	0.28	98.98
175	11	0.5092593	0.51	99.49
176	1	0.0462963	0.05	99.54
177	2	0.0925926	0.09	99.63
178	7	0.3240741	0.32	99.95
179	1	0.0462963	0.05	100.00
Total	2160	100.0000000	100.00	100.00

Frequency Table for Variable: shortname

147 unique value(s) detected.

shortname	N	exactpercent	roundedpercent	cumulpercent
1955AmityTreaty	11	0.5092593	0.51	0.51
ATILO-UNESCO	9	0.4166667	0.42	0.93
AccessPacificOcean	14	0.6481481	0.65	1.57
AdmissionUN	7	0.3240741	0.32	1.90
AegeanSeaContinentalShelf	20	0.9259259	0.93	2.82
AerialHerbicideSpraying	4	0.1851852	0.19	3.01
AerialIncident1952	1	0.0462963	0.05	3.06
AerialIncident1953	1	0.0462963	0.05	3.10

(continued)

shortname	N	exactpercent	roundedpercent	cumulpercent
AerialIncident1988	9	0.4166667	0.42	3.52
AerialIncident1999	7	0.3240741	0.32	3.84
AerialIncidentNov1954	1	0.0462963	0.05	3.89
AerialIncidentSept1954	1	0.0462963	0.05	3.94
AerialIndicent1955	21	0.9722222	0.97	4.91
Ambatielos	16	0.7407407	0.74	5.65
AngloIranianOil	12	0.5555556	0.56	6.20
Antarctica	2	0.0925926	0.09	6.30
ApplicationGenocideConvention	81	3.7500000	3.75	10.05
ApplicationGenocideConvention-Revision	6	0.2777778	0.28	10.32
ArbitralAward1899	9	0.4166667	0.42	10.74
ArbitralAward1989	15	0.6944444	0.69	11.44
ArbitralAwardKingOfSpain	6	0.2777778	0.28	11.71
ArbitrationUNHQAgreement	6	0.2777778	0.28	11.99
ArmedActivities	34	1.5740741	1.57	13.56
ArmedActivitiesApp2002	16	0.7407407	0.74	14.31
ArrestWarrant	22	1.0185185	1.02	15.32
Asylum	9	0.4166667	0.42	15.74
Asylum-Interpretation	1	0.0462963	0.05	15.79
Avena	11	0.5092593	0.51	16.30
Avena-Interpretation	8	0.3703704	0.37	16.67
BarcelonaTraction1958	5	0.2314815	0.23	16.90
BarcelonaTraction1962	31	1.4351852	1.44	18.33
CERD	18	0.8333333	0.83	19.17
CertainActivitiesBorderArea	40	1.8518519	1.85	21.02
CertainCriminalProceedings	11	0.5092593	0.51	21.53
CertainDocumentsSeizure	10	0.4629630	0.46	21.99
CertainExpensesUN	11	0.5092593	0.51	22.50

(continued)

shortname	N	exactpercent	roundedpercent	cumulpercent
CertainPhosphateLands	11	0.5092593	0.51	23.01
CertainProperty	8	0.3703704	0.37	23.38
ChagosArchipelago	15	0.6944444	0.69	24.07
CompensationUNAT	6	0.2777778	0.28	24.35
CompetenceAdmissionGA	4	0.1851852	0.19	24.54
ConstitutionMaritimeSafetyCommittee	4	0.1851852	0.19	24.72
ConstructionWallOPT	11	0.5092593	0.51	25.23
ContinentalShelf	36	1.6666667	1.67	26.90
ContinentalShelf-InterpretationRevision	5	0.2314815	0.23	27.13
ConventionPrivilegesImmunitiesUN	5	0.2314815	0.23	27.36
ConventionTerrorismFinancingCERD	23	1.0648148	1.06	28.43
CorfuChannel	19	0.8796296	0.88	29.31
DelimitationContinentalShelf	13	0.6018519	0.60	29.91
Diallo	24	1.1111111	1.11	31.02
DiplomaticEnvoyUN	1	0.0462963	0.05	31.06
DiplomaticRelations	1	0.0462963	0.05	31.11
ELSI	6	0.2777778	0.28	31.39
EastTimor	10	0.4629630	0.46	31.85
ElectriciteBeyrouth	3	0.1388889	0.14	31.99
Fisheries	9	0.4166667	0.42	32.41
FisheriesJurisdiction	62	2.8703704	2.87	35.28
FrenchNationalsEgypt	1	0.0462963	0.05	35.32
FrontierDispute	22	1.0185185	1.02	36.34
GabcikovoNagymaros	16	0.7407407	0.74	37.08
GuardianshipInfantsConvention	11	0.5092593	0.51	37.59
GuatemalaTerritorialInsularMaritimeClaim	2	0.0925926	0.09	37.69
GulfOfMaine	12	0.5555556	0.56	38.24
HayaDeLaTorre	2	0.0925926	0.09	38.33

(continued)

shortname	N	exactpercent	roundedpercent	cumulpercent
ICAOCouncil	14	0.6481481	0.65	38.98
ICAOCouncil-CICA	6	0.2777778	0.28	39.26
ICAOCouncil-IASTA	6	0.2777778	0.28	39.54
ICERD	16	0.7407407	0.74	40.28
ImmunitiesCriminalProceedings	26	1.2037037	1.20	41.48
ImmunitySRCommHR	6	0.2777778	0.28	41.76
IndependenceDeclarationKosovo	11	0.5092593	0.51	42.27
Interhandel	16	0.7407407	0.74	43.01
InterimAccord1995	8	0.3703704	0.37	43.38
IranianAssets	12	0.5555556	0.56	43.94
IslaPortillos	10	0.4629630	0.46	44.40
Jadhav	11	0.5092593	0.51	44.91
Judgment2867ATILO-IFAD	5	0.2314815	0.23	45.14
JudgmentsCivilCommercialMatters	3	0.1388889	0.14	45.28
JurisdictionalImmunities	15	0.6944444	0.69	45.97
KasikiliSedudu	12	0.5555556	0.56	46.53
LaGrand	11	0.5092593	0.51	47.04
LandIslandMaritimeFrontier	20	0.9259259	0.93	47.96
LandIslandMaritimeFrontier-Revision	4	0.1851852	0.19	48.15
LandMaritimeBoundary	34	1.5740741	1.57	49.72
LandMaritimeBoundary-Interpretation	4	0.1851852	0.19	49.91
LandMaritimeDelimitationSovereigntyIslands	1	0.0462963	0.05	49.95
LegalityNuclearWeaponsArmedConflict	9	0.4166667	0.42	50.37
LegalityThreatUseNuclearWeapons	16	0.7407407	0.74	51.11
Lockerbie	53	2.4537037	2.45	53.56
MaritimeDelimitation	30	1.3888889	1.39	54.95
MaritimeDelimitation-BlackSea	4	0.1851852	0.19	55.14

(continued)

shortname	N	exactpercent	roundedpercent	cumulpercent
MaritimeDelimitation-CaribbeanPacific	13	0.6018519	0.60	55.74
MaritimeDelimitation-GreenlandJanMayen	13	0.6018519	0.60	56.34
MaritimeDelimitation-IndianOcean	13	0.6018519	0.60	56.94
MaritimeDispute	13	0.6018519	0.60	57.55
MilitaryParamilitaryActivitiesNicaragua	32	1.4814815	1.48	59.03
MinquiersEcrehos	7	0.3240741	0.32	59.35
MonetaryGold	7	0.3240741	0.32	59.68
MutualAssistanceCriminalMatters	11	0.5092593	0.51	60.19
Namibia	16	0.7407407	0.74	60.93
NavigationalRights	6	0.2777778	0.28	61.20
NorthSeaContinentalShelf	30	1.3888889	1.39	62.59
NorthernCameroons	18	0.8333333	0.83	63.43
NorwegianLoans	11	0.5092593	0.51	63.94
Nottebohm	11	0.5092593	0.51	64.44
NuclearDisarmament	51	2.3611111	2.36	66.81
NuclearTests	60	2.7777778	2.78	69.58
NuclearTests-ExaminationSituation	8	0.3703704	0.37	69.95
ObligationProsecuteExtradite	17	0.7870370	0.79	70.74
OilPlatforms	32	1.4814815	1.48	72.22
PassageGreatBelt	7	0.3240741	0.32	72.55
PassageIndianTerritory	25	1.1574074	1.16	73.70
PeaceTreaties	10	0.4629630	0.46	74.17
PedraBranca	9	0.4166667	0.42	74.58
PedraBranca-Interpretation	1	0.0462963	0.05	74.63
PedraBranca-Revision	1	0.0462963	0.05	74.68
PetitionersComitteeSouthWestAfrica	6	0.2777778	0.28	74.95

(continued)

shortname	N	exactpercent	roundedpercent	cumulpercent
PortBeyrouthSRO	4	0.1851852	0.19	75.14
PulpMills	20	0.9259259	0.93	76.06
RelocationEmbassyUSJerusalem	1	0.0462963	0.05	76.11
ReparationUN	7	0.3240741	0.32	76.44
ReservationsGenocideConvention	4	0.1851852	0.19	76.62
ReviewJudgment158UNAT	11	0.5092593	0.51	77.13
ReviewJudgment273UNAT	11	0.5092593	0.51	77.64
ReviewJudgment333UNAT	10	0.4629630	0.46	78.10
SanJuanRiver	16	0.7407407	0.74	78.84
SilalaWaters	4	0.1851852	0.19	79.03
SouthWestAfrica	60	2.7777778	2.78	81.81
SovereignRightsCaribbeanSea	15	0.6944444	0.69	82.50
SovereigntyFrontierLand	8	0.3703704	0.37	82.87
SovereigntyPulau	14	0.6481481	0.65	83.52
StatusSouthWestAfrica	7	0.3240741	0.32	83.84
TemplePreahVihear	16	0.7407407	0.74	84.58
TemplePreahVihear-Interpretation	14	0.6481481	0.65	85.23
TerritorialDispute	43	1.9907407	1.99	87.22
TerritorialDispute-CaribbeanSea	8	0.3703704	0.37	87.59
TransborderArmedActions	14	0.6481481	0.65	88.24
TreatmentAirCrew	2	0.0925926	0.09	88.33
TrialPakistaniPOW	5	0.2314815	0.23	88.56
USDiplomaticStaffTehran	7	0.3240741	0.32	88.89
USNationalsMorocco	6	0.2777778	0.28	89.17
UseOfForce	179	8.2870370	8.29	97.45
ViennaConventionConsularRelations	7	0.3240741	0.32	97.78
VotingProcedureSouthWestAfrica	5	0.2314815	0.23	98.01
WHO-EgyptAgreement	11	0.5092593	0.51	98.52

(continued)

shortname	N	exactpercent	roundedpercent	cumulpercent
WesternSahara	15	0.6944444	0.69	99.21
WhalingAntarctic	17	0.7870370	0.79	100.00
Total	2160	100.0000000	100.00	100.00

Frequency Table for Variable: applicant

79 unique value(s) detected.

applicant	N	exactpercent	roundedpercent	cumulpercent
ARG	20	0.9259259	0.93	0.93
AUS	48	2.2222222	2.22	3.15
BEL	64	2.9629630	2.96	6.11
BEN	6	0.2777778	0.28	6.39
BFA	16	0.7407407	0.74	7.13
BHR-EGY-ARE	6	0.2777778	0.28	7.41
BHR-EGY-SAU-ARE	6	0.2777778	0.28	7.69
BIH	49	2.2685185	2.27	9.95
BOL	14	0.6481481	0.65	10.60
BWA	12	0.5555556	0.56	11.16
CAN	12	0.5555556	0.56	11.71
CARAT	32	1.4814815	1.48	13.19
CHE	16	0.7407407	0.74	13.94
CHL	4	0.1851852	0.19	14.12
CMR	56	2.5925926	2.59	16.71
COD	83	3.8425926	3.84	20.56
COL	12	0.5555556	0.56	21.11
CRI	69	3.1944444	3.19	24.31
DEU	81	3.7500000	3.75	28.06
DJI	11	0.5092593	0.51	28.56

(continued)

applicant	N	exactpercent	roundedpercent	cumulpercent
DMA	1	0.0462963	0.05	28.61
DNK	13	0.6018519	0.60	29.21
ECOSOC	11	0.5092593	0.51	29.72
ECU	4	0.1851852	0.19	29.91
ESP	13	0.6018519	0.60	30.51
ETH	30	1.3888889	1.39	31.90
FIN	7	0.3240741	0.32	32.22
FRA	32	1.4814815	1.48	33.70
GAB	1	0.0462963	0.05	33.75
GBR	77	3.5648148	3.56	37.31
GEO	16	0.7407407	0.74	38.06
GIN	24	1.1111111	1.11	39.17
GMB	7	0.3240741	0.32	39.49
GNB	16	0.7407407	0.74	40.23
GNQ	26	1.2037037	1.20	41.44
GRC	36	1.6666667	1.67	43.10
GTM	2	0.0925926	0.09	43.19
GUY	9	0.4166667	0.42	43.61
HND	7	0.3240741	0.32	43.94
HRV	31	1.4351852	1.44	45.37
HUN	16	0.7407407	0.74	46.11
IDN	14	0.6481481	0.65	46.76
IFAD	5	0.2314815	0.23	46.99
IMO	4	0.1851852	0.19	47.18
IND	25	1.1574074	1.16	48.33
IRN	64	2.9629630	2.96	51.30
ISR	8	0.3703704	0.37	51.67
ITA	7	0.3240741	0.32	51.99
KHM	30	1.3888889	1.39	53.38

(continued)

applicant	N	exactpercent	roundedpercent	cumulpercent
LBR	30	1.3888889	1.39	54.77
LBY	83	3.8425926	3.84	58.61
LIE	19	0.8796296	0.88	59.49
MEX	19	0.8796296	0.88	60.37
MHL	51	2.3611111	2.36	62.73
MKD	8	0.3703704	0.37	63.10
MYS	11	0.5092593	0.51	63.61
NIC	133	6.1574074	6.16	69.77
NLD	11	0.5092593	0.51	70.28
NRU	11	0.5092593	0.51	70.79
NZL	37	1.7129630	1.71	72.50
PAK	12	0.5555556	0.56	73.06
PER	13	0.6018519	0.60	73.66
PRT	35	1.6203704	1.62	75.28
PRY	7	0.3240741	0.32	75.60
PSE	1	0.0462963	0.05	75.65
QAT	47	2.1759259	2.18	77.82
ROU	4	0.1851852	0.19	78.01
SCG	163	7.5462963	7.55	85.56
SLV	24	1.1111111	1.11	86.67
SOM	13	0.6018519	0.60	87.27
TLS	10	0.4629630	0.46	87.73
TUN	19	0.8796296	0.88	88.61
UKR	23	1.0648148	1.06	89.68
UNESCO	9	0.4166667	0.42	90.09
UNGA	141	6.5277778	6.53	96.62
UNSC	16	0.7407407	0.74	97.36
USA	21	0.9722222	0.97	98.33
WHO	20	0.9259259	0.93	99.26

(continued)

applicant	N	exactpercent	roundedpercent	cumulpercent
YUG	16	0.7407407	0.74	100.00
Total	2160	100.0000000	100.00	100.00

Frequency Table for Variable: respondent

72 unique value(s) detected.

respondent	N	exactpercent	roundedpercent	cumulpercent
NA	238	11.0185185	11.02	11.02
ALB	19	0.8796296	0.88	11.90
ARE	18	0.8333333	0.83	12.73
ARG	1	0.0462963	0.05	12.78
AUS	31	1.4351852	1.44	14.21
BDI	3	0.1388889	0.14	14.35
BEL	43	1.9907407	1.99	16.34
BGR	21	0.9722222	0.97	17.31
BHR	29	1.3425926	1.34	18.66
BLZ	2	0.0925926	0.09	18.75
BOL	4	0.1851852	0.19	18.94
BRA	1	0.0462963	0.05	18.98
CAN	34	1.5740741	1.57	20.56
CHE	4	0.1851852	0.19	20.74
CHL	28	1.2962963	1.30	22.04
COD	24	1.1111111	1.11	23.15
COL	67	3.1018519	3.10	26.25
CRI	19	0.8796296	0.88	27.13
CSK	1	0.0462963	0.05	27.18
DEU	27	1.2500000	1.25	28.43
DNK	22	1.0185185	1.02	29.44

(continued)

respondent	N	exactpercent	roundedpercent	cumulpercent
EGY	1	0.0462963	0.05	29.49
ESP	52	2.4074074	2.41	31.90
FRA	135	6.2500000	6.25	38.15
FRA-GBR-USA	7	0.3240741	0.32	38.47
GBR	107	4.9537037	4.95	43.43
GNQ	1	0.0462963	0.05	43.47
GRC	8	0.3703704	0.37	43.84
GTM	11	0.5092593	0.51	44.35
HND	43	1.9907407	1.99	46.34
HUN	1	0.0462963	0.05	46.39
IND	54	2.5000000	2.50	48.89
IRN	19	0.8796296	0.88	49.77
ISL	49	2.2685185	2.27	52.04
ITA	41	1.8981481	1.90	53.94
JPN	17	0.7870370	0.79	54.72
KEN	13	0.6018519	0.60	55.32
LBN	7	0.3240741	0.32	55.65
LBY	19	0.8796296	0.88	56.53
MLI	8	0.3703704	0.37	56.90
MLT	22	1.0185185	1.02	57.92
MMR	7	0.3240741	0.32	58.24
MYS	14	0.6481481	0.65	58.89
NAM	12	0.5555556	0.56	59.44
NER	14	0.6481481	0.65	60.09
NGA	38	1.7592593	1.76	61.85
NIC	75	3.4722222	3.47	65.32
NLD	44	2.0370370	2.04	67.36
NOR	33	1.5277778	1.53	68.89
PAK	42	1.9444444	1.94	70.83

(continued)

respondent	N	exactpercent	roundedpercent	cumulpercent
PER	12	0.5555556	0.56	71.39
PRT	21	0.9722222	0.97	72.36
QAT	12	0.5555556	0.56	72.92
RUS	39	1.8055556	1.81	74.72
RWA	19	0.8796296	0.88	75.60
SCG	43	1.9907407	1.99	77.59
SEN	33	1.5277778	1.53	79.12
SGP	11	0.5092593	0.51	79.63
SRB	31	1.4351852	1.44	81.06
SUN	4	0.1851852	0.19	81.25
SVK	16	0.7407407	0.74	81.99
SWE	11	0.5092593	0.51	82.50
TCD	8	0.3703704	0.37	82.87
THA	30	1.3888889	1.39	84.26
TUR	20	0.9259259	0.93	85.19
UGA	28	1.2962963	1.30	86.48
UKR	4	0.1851852	0.19	86.67
URY	20	0.9259259	0.93	87.59
USA	193	8.9351852	8.94	96.53
VEN	9	0.4166667	0.42	96.94
YUG	6	0.2777778	0.28	97.22
ZAF	60	2.7777778	2.78	100.00
Total	2160	100.0000000	100.00	100.00

Frequency Table for Variable: doctype

3 unique value(s) detected.

doctype	N	exactpercent	roundedpercent	cumulpercent
ADV	194	8.981481	8.98	8.98
JUD	1034	47.870370	47.87	56.85
ORD	932	43.148148	43.15	100.00
Total	2160	100.000000	100.00	100.00

Frequency Table for Variable: collision

3 unique value(s) detected.

collision	N	exactpercent	roundedpercent	cumulpercent
1	2143	99.2129630	99.21	99.21
2	16	0.7407407	0.74	99.95
3	1	0.0462963	0.05	100.00
Total	2160	100.0000000	100.00	100.00

Frequency Table for Variable: stage

5 unique value(s) detected.

stage	N	exactpercent	roundedpercent	cumulpercent
NA	1129	52.268518	52.27	52.27
CO	33	1.527778	1.53	53.80
IN	36	1.666667	1.67	55.46
ME	543	25.138889	25.14	80.60
PO	419	19.398148	19.40	100.00
Total	2160	100.000000	100.00	100.00

Frequency Table for Variable: opinion

15 unique value(s) detected.

opinion	N	exactpercent	roundedpercent	cumulpercent
0	763	35.3240741	35.32	35.32
1	251	11.6203704	11.62	46.94
2	227	10.5092593	10.51	57.45
3	198	9.1666667	9.17	66.62
4	168	7.7777778	7.78	74.40
5	145	6.7129630	6.71	81.11
6	125	5.7870370	5.79	86.90
7	95	4.3981481	4.40	91.30
8	71	3.2870370	3.29	94.58
9	57	2.6388889	2.64	97.22
10	30	1.3888889	1.39	98.61
11	14	0.6481481	0.65	99.26
12	8	0.3703704	0.37	99.63
13	4	0.1851852	0.19	99.81
14	4	0.1851852	0.19	100.00
Total	2160	100.0000000	100.00	100.00

Frequency Table for Variable: language

1 unique value(s) detected.

language	N	exactpercent	roundedpercent	cumulpercent
FR	2160	100	100	100
Total	2160	100	100	100

Frequency Table for Variable: year

75 unique value(s) detected.

year	N	exactpercent	roundedpercent	cumulpercent
1947	3	0.1388889	0.14	0.14
1948	12	0.5555556	0.56	0.69
1949	24	1.1111111	1.11	1.81
1950	31	1.4351852	1.44	3.24
1951	21	0.9722222	0.97	4.21
1952	26	1.2037037	1.20	5.42
1953	11	0.5092593	0.51	5.93
1954	19	0.8796296	0.88	6.81
1955	11	0.5092593	0.51	7.31
1956	22	1.0185185	1.02	8.33
1957	22	1.0185185	1.02	9.35
1958	28	1.2962963	1.30	10.65
1959	32	1.4814815	1.48	12.13
1960	26	1.2037037	1.20	13.33
1961	17	0.7870370	0.79	14.12
1962	42	1.9444444	1.94	16.06
1963	17	0.7870370	0.79	16.85
1964	15	0.6944444	0.69	17.55
1965	5	0.2314815	0.23	17.78
1966	24	1.1111111	1.11	18.89
1967	4	0.1851852	0.19	19.07
1968	5	0.2314815	0.23	19.31
1969	24	1.1111111	1.11	20.42
1970	14	0.6481481	0.65	21.06
1971	16	0.7407407	0.74	21.81
1972	23	1.0648148	1.06	22.87
1973	62	2.8703704	2.87	25.74
1974	52	2.4074074	2.41	28.15
1975	15	0.6944444	0.69	28.84
1976	11	0.5092593	0.51	29.35

(continued)

year	N	exactpercent	roundedpercent	cumulpercent
1977	1	0.0462963	0.05	29.40
1978	8	0.3703704	0.37	29.77
1979	3	0.1388889	0.14	29.91
1980	16	0.7407407	0.74	30.65
1981	8	0.3703704	0.37	31.02
1982	24	1.1111111	1.11	32.13
1983	2	0.0925926	0.09	32.22
1984	35	1.6203704	1.62	33.84
1985	18	0.8333333	0.83	34.68
1986	17	0.7870370	0.79	35.46
1987	17	0.7870370	0.79	36.25
1988	14	0.6481481	0.65	36.90
1989	21	0.9722222	0.97	37.87
1990	15	0.6944444	0.69	38.56
1991	23	1.0648148	1.06	39.63
1992	43	1.9907407	1.99	41.62
1993	29	1.3425926	1.34	42.96
1994	14	0.6481481	0.65	43.61
1995	29	1.3425926	1.34	44.95
1996	55	2.5462963	2.55	47.50
1997	19	0.8796296	0.88	48.38
1998	61	2.8240741	2.82	51.20
1999	134	6.2037037	6.20	57.41
2000	37	1.7129630	1.71	59.12
2001	44	2.0370370	2.04	61.16
2002	49	2.2685185	2.27	63.43
2003	35	1.6203704	1.62	65.05
2004	77	3.5648148	3.56	68.61
2005	21	0.9722222	0.97	69.58

(continued)

year	N	exactpercent	roundedpercent	cumulpercent
2006	18	0.8333333	0.83	70.42
2007	39	1.8055556	1.81	72.22
2008	42	1.9444444	1.94	74.17
2009	20	0.9259259	0.93	75.09
2010	42	1.9444444	1.94	77.04
2011	57	2.6388889	2.64	79.68
2012	36	1.6666667	1.67	81.34
2013	33	1.5277778	1.53	82.87
2014	40	1.8518519	1.85	84.72
2015	51	2.3611111	2.36	87.08
2016	75	3.4722222	3.47	90.56
2017	35	1.6203704	1.62	92.18
2018	64	2.9629630	2.96	95.14
2019	54	2.5000000	2.50	97.64
2020	35	1.6203704	1.62	99.26
2021	16	0.7407407	0.74	100.00
Total	2160	100.0000000	100.00	100.00

Frequency Table for Variable: minority

2 unique value(s) detected.

minority	N	exactpercent	roundedpercent	cumulpercent
0	763	35.32407	35.32	35.32
1	1397	64.67593	64.68	100.00
Total	2160	100.00000	100.00	100.00

Frequency Table for Variable: fullname

178 unique value(s) detected.

fullname	N	exactpercent	roundedpercent	cumulpercent
Accordance with international law of the unilateral declaration of independence in respect of Kosovo	11	0.5092593	0.51	0.51
Admissibility of Hearings of Petitioners by the Committee on South West Africa	6	0.2777778	0.28	0.79
Aegean Sea Continental Shelf (Greece v. Turkey)	20	0.9259259	0.93	1.71
Aerial Herbicide Spraying (Ecuador v. Colombia)	4	0.1851852	0.19	1.90
Aerial Incident of 10 August 1999 (Pakistan v. India)	7	0.3240741	0.32	2.22
Aerial Incident of 10 March 1953 (United States of America v. Czechoslovakia)	1	0.0462963	0.05	2.27
Aerial Incident of 27 July 1955 (Israel v. Bulgaria)	8	0.3703704	0.37	2.64
Aerial Incident of 27 July 1955 (United Kingdom v. Bulgaria)	5	0.2314815	0.23	2.87
Aerial Incident of 27 July 1955 (United States of America v. Bulgaria)	8	0.3703704	0.37	3.24
Aerial Incident of 3 July 1988 (Islamic Republic of Iran v. United States of America)	9	0.4166667	0.42	3.66
Aerial Incident of 4 September 1954 (United States of America v. Union of Soviet Socialist Republics)	1	0.0462963	0.05	3.70
Aerial Incident of 7 November 1954 (United States of America v. Union of Soviet Socialist Republics)	1	0.0462963	0.05	3.75
Aerial Incident of 7 October 1952 (United States of America v. Union of Soviet Socialist Republics)	1	0.0462963	0.05	3.80

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Ahmadou Sadio Diallo (Republic of Guinea v. Democratic Republic of the Congo)	24	1.1111111	1.11	4.91
Alleged Violations of Sovereign Rights and Maritime Spaces in the Caribbean Sea (Nicaragua v. Colombia)	15	0.6944444	0.69	5.60
Alleged violations of the 1955 Treaty of Amity, Economic Relations, and Consular Rights (Islamic Republic of Iran v. United States of America)	11	0.5092593	0.51	6.11
Ambatielos (Greece v. United Kingdom)	16	0.7407407	0.74	6.85
Anglo-Iranian Oil Co. (United Kingdom v. Iran)	12	0.5555556	0.56	7.41
Antarctica (United Kingdom v. Argentina)	1	0.0462963	0.05	7.45
Antarctica (United Kingdom v. Chile)	1	0.0462963	0.05	7.50
Appeal Relating to the Jurisdiction of the ICAO Council (India v. Pakistan)	14	0.6481481	0.65	8.15
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	0.2777778	0.28	8.43
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	0.2777778	0.28	8.70

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Applicability of Article VI, Section 22, of the Convention on the Privileges and Immunities of the United Nations	5	0.2314815	0.23	8.94
Applicability of the Obligation to Arbitrate under Section 21 of the United Nations Headquarters Agreement of 26 June 1947	6	0.2777778	0.28	9.21
Application for Review of Judgment No. 158 of the United Nations Administrative Tribunal	11	0.5092593	0.51	9.72
Application for Review of Judgment No. 273 of the United Nations Administrative Tribunal	11	0.5092593	0.51	10.23
Application for Review of Judgment No. 333 of the United Nations Administrative Tribunal	10	0.4629630	0.46	10.69
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	0.2314815	0.23	10.93
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	0.2777778	0.28	11.20

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
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	0.1851852	0.19	11.39
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	0.0462963	0.05	11.44
Application of the Convention of 1902 Governing the Guardianship of Infants (Netherlands v. Sweden)	11	0.5092593	0.51	11.94
Application of the Convention on the Prevention and Punishment of the Crime of Genocide (Bosnia and Herzegovina v. Serbia and Montenegro)	43	1.9907407	1.99	13.94
Application of the Convention on the Prevention and Punishment of the Crime of Genocide (Croatia v. Serbia)	31	1.4351852	1.44	15.37
Application of the Convention on the Prevention and Punishment of the Crime of Genocide (The Gambia v. Myanmar)	7	0.3240741	0.32	15.69
Application of the Interim Accord of 13 September 1995 (the former Yugoslav Republic of Macedonia v. Greece)	8	0.3703704	0.37	16.06

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
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)	23	1.0648148	1.06	17.13
Application of the International Convention on the Elimination of All Forms of Racial Discrimination (Georgia v. Russian Federation)	16	0.7407407	0.74	17.87
Application of the International Convention on the Elimination of All Forms of Racial Discrimination (Qatar v. United Arab Emirates)	18	0.8333333	0.83	18.70
Arbitral Award Made by the King of Spain on 23 December 1906 (Honduras v. Nicaragua)	6	0.2777778	0.28	18.98
Arbitral Award of 3 October 1899 (Guyana v. Venezuela)	9	0.4166667	0.42	19.40
Arbitral Award of 31 July 1989 (Guinea-Bissau v. Senegal)	15	0.6944444	0.69	20.09
Armed Activities on the Territory of the Congo (Democratic Republic of the Congo v. Burundi)	3	0.1388889	0.14	20.23
Armed Activities on the Territory of the Congo (Democratic Republic of the Congo v. Rwanda)	3	0.1388889	0.14	20.37
Armed Activities on the Territory of the Congo (Democratic Republic of the Congo v. Uganda)	28	1.2962963	1.30	21.67

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Armed Activities on the Territory of the Congo (New Application: 2002) (Democratic Republic of the Congo v. Rwanda)	16	0.7407407	0.74	22.41
Arrest Warrant of 11 April 2000 (Democratic Republic of the Congo v. Belgium)	22	1.0185185	1.02	23.43
Asylum (Colombia v. Peru)	9	0.4166667	0.42	23.84
Avena and Other Mexican Nationals (Mexico v. United States of America)	11	0.5092593	0.51	24.35
Barcelona Traction, Light and Power Company, Limited (Belgium v. Spain)	5	0.2314815	0.23	24.58
Barcelona Traction, Light and Power Company, Limited (Belgium v. Spain) (New Application: 1962)	31	1.4351852	1.44	26.02
Border and Transborder Armed Actions (Nicaragua v. Costa Rica)	3	0.1388889	0.14	26.16
Border and Transborder Armed Actions (Nicaragua v. Honduras)	11	0.5092593	0.51	26.67
Certain Activities Carried Out by Nicaragua in the Border Area (Costa Rica v. Nicaragua)	40	1.8518519	1.85	28.52
Certain Criminal Proceedings in France (Republic of the Congo v. France)	11	0.5092593	0.51	29.03
Certain Expenses of the United Nations (Article 17, paragraph 2, of the Charter)	11	0.5092593	0.51	29.54
Certain Iranian Assets (Islamic Republic of Iran v. United States of America)	12	0.5555556	0.56	30.09

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Certain Norwegian Loans (France v. Norway)	11	0.5092593	0.51	30.60
Certain Phosphate Lands in Nauru (Nauru v. Australia)	11	0.5092593	0.51	31.11
Certain Property (Liechtenstein v. Germany)	8	0.3703704	0.37	31.48
Certain Questions concerning Diplomatic Relations (Honduras v. Brazil)	1	0.0462963	0.05	31.53
Certain Questions of Mutual Assistance in Criminal Matters (Djibouti v. France)	11	0.5092593	0.51	32.04
Compagnie du Port, des Quais et des Entrepôts de Beyrouth and Société Radio-Orient (France v. Lebanon)	4	0.1851852	0.19	32.22
Competence of the General Assembly for the Admission of a State to the United Nations	4	0.1851852	0.19	32.41
Conditions of Admission of a State to Membership in the United Nations (Article 4 of the Charter)	7	0.3240741	0.32	32.73
Constitution of the Maritime Safety Committee of the Inter-Governmental Maritime Consultative Organization	4	0.1851852	0.19	32.92
Construction of a Road in Costa Rica along the San Juan River (Nicaragua v. Costa Rica)	16	0.7407407	0.74	33.66
Continental Shelf (Libyan Arab Jamahiriya/Malta)	22	1.0185185	1.02	34.68
Continental Shelf (Tunisia/Libyan Arab Jamahiriya)	14	0.6481481	0.65	35.32
Corfu Channel (United Kingdom of Great Britain and Northern Ireland v. Albania)	19	0.8796296	0.88	36.20

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Delimitation of the Maritime Boundary in the Gulf of Maine Area (Canada/United States of America)	12	0.5555556	0.56	36.76
Difference Relating to Immunity from Legal Process of a Special Rapporteur of the Commission on Human Rights	6	0.2777778	0.28	37.04
Dispute over the Status and Use of the Waters of the Silala (Chile v. Bolivia)	4	0.1851852	0.19	37.22
Dispute regarding Navigational and Related Rights (Costa Rica v. Nicaragua)	6	0.2777778	0.28	37.50
East Timor (Portugal v. Australia)	10	0.4629630	0.46	37.96
Effect of Awards of Compensation Made by the United Nations Administrative Tribunal	6	0.2777778	0.28	38.24
Electricité de Beyrouth Company (France v. Lebanon)	3	0.1388889	0.14	38.38
Elettronica Sicula S.p.A. (ELSI) (United States of America v. Italy)	6	0.2777778	0.28	38.66
Fisheries (United Kingdom v. Norway)	9	0.4166667	0.42	39.07
Fisheries Jurisdiction (Federal Republic of Germany v. Iceland)	25	1.1574074	1.16	40.23
Fisheries Jurisdiction (Spain v. Canada)	13	0.6018519	0.60	40.83
Fisheries Jurisdiction (United Kingdom v. Iceland)	24	1.1111111	1.11	41.94
Frontier Dispute (Benin/Niger)	6	0.2777778	0.28	42.22
Frontier Dispute (Burkina Faso/Niger)	8	0.3703704	0.37	42.59

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Frontier Dispute (Burkina Faso/Republic of Mali)	8	0.3703704	0.37	42.96
Gabčíkovo-Nagymaros Project (Hungary/Slovakia)	16	0.7407407	0.74	43.70
Guatemala's Territorial, Insular and Maritime Claim (Guatemala/Belize)	2	0.0925926	0.09	43.80
Haya de la Torre (Colombia v. Peru)	2	0.0925926	0.09	43.89
Immunities and Criminal Proceedings (Equatorial Guinea v. France)	26	1.2037037	1.20	45.09
Interhandel (Switzerland v. United States of America)	16	0.7407407	0.74	45.83
International Status of South West Africa	7	0.3240741	0.32	46.16
Interpretation of Peace Treaties with Bulgaria, Hungary and Romania	10	0.4629630	0.46	46.62
Interpretation of the Agreement of 25 March 1951 between the WHO and Egypt	11	0.5092593	0.51	47.13
Jadhav (India v. Pakistan)	11	0.5092593	0.51	47.64
Judgment No.2867 of the Administrative Tribunal of the International Labour Organization upon a Complaint Filed against the International Fund for Agricultural Development	5	0.2314815	0.23	47.87
Judgments of the Administrative Tribunal of the ILO upon Complaints Made against UNESCO	9	0.4166667	0.42	48.29
Jurisdiction and Enforcement of Judgments in Civil and Commercial Matters (Belgium v. Switzerland)	3	0.1388889	0.14	48.43

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Jurisdictional Immunities of the State (Germany v. Italy: Greece intervening)	15	0.6944444	0.69	49.12
Kasikili/Sedudu Island (Botswana/Namibia)	12	0.5555556	0.56	49.68
LaGrand (Germany v. United States of America)	11	0.5092593	0.51	50.19
Land Boundary in the Northern Part of Isla Portillos (Costa Rica v. Nicaragua)	10	0.4629630	0.46	50.65
Land and Maritime Boundary between Cameroon and Nigeria (Cameroon v. Nigeria: Equatorial Guinea intervening)	34	1.5740741	1.57	52.22
Land and Maritime Delimitation and Sovereignty over Islands (Gabon/Equatorial Guinea)	1	0.0462963	0.05	52.27
Land, Island and Maritime Frontier Dispute (El Salvador/Honduras: Nicaragua intervening)	20	0.9259259	0.93	53.19
Legal Consequences for States of the Continued Presence of South Africa in Namibia (South West Africa) notwithstanding Security Council Resolution 276 (1970)	16	0.7407407	0.74	53.94
Legal Consequences of the Construction of a Wall in the Occupied Palestinian Territory	11	0.5092593	0.51	54.44
Legal Consequences of the Separation of the Chagos Archipelago from Mauritius in 1965	15	0.6944444	0.69	55.14

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Legality of Use of Force (Serbia and Montenegro v. Belgium)	21	0.9722222	0.97	56.11
Legality of Use of Force (Serbia and Montenegro v. Canada)	21	0.9722222	0.97	57.08
Legality of Use of Force (Serbia and Montenegro v. France)	19	0.8796296	0.88	57.96
Legality of Use of Force (Serbia and Montenegro v. Germany)	19	0.8796296	0.88	58.84
Legality of Use of Force (Serbia and Montenegro v. Italy)	20	0.9259259	0.93	59.77
Legality of Use of Force (Serbia and Montenegro v. Netherlands)	21	0.9722222	0.97	60.74
Legality of Use of Force (Serbia and Montenegro v. Portugal)	21	0.9722222	0.97	61.71
Legality of Use of Force (Serbia and Montenegro v. United Kingdom)	21	0.9722222	0.97	62.69
Legality of Use of Force (Yugoslavia v. Spain)	9	0.4166667	0.42	63.10
Legality of Use of Force (Yugoslavia v. United States of America)	7	0.3240741	0.32	63.43
Legality of the Threat or Use of Nuclear Weapons	16	0.7407407	0.74	64.17
Legality of the Use by a State of Nuclear Weapons in Armed Conflict	9	0.4166667	0.42	64.58
Maritime Delimitation and Territorial Questions between Qatar and Bahrain (Qatar v. Bahrain)	29	1.3425926	1.34	65.93

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Maritime Delimitation between Guinea-Bissau and Senegal (Guinea-Bissau v. Senegal)	1	0.0462963	0.05	65.97
Maritime Delimitation in the Area between Greenland and Jan Mayen (Denmark v. Norway)	13	0.6018519	0.60	66.57
Maritime Delimitation in the Black Sea (Romania v. Ukraine)	4	0.1851852	0.19	66.76
Maritime Delimitation in the Caribbean Sea and the Pacific Ocean (Costa Rica v. Nicaragua)	13	0.6018519	0.60	67.36
Maritime Delimitation in the Indian Ocean (Somalia v. Kenya)	13	0.6018519	0.60	67.96
Maritime Dispute (Peru v. Chile)	13	0.6018519	0.60	68.56
Military and Paramilitary Activities in and against Nicaragua (Nicaragua v. United States of America)	32	1.4814815	1.48	70.05
Minquiers and Ecrehos (France/United Kingdom)	7	0.3240741	0.32	70.37
Monetary Gold Removed from Rome in 1943 (Italy v. France, United Kingdom of Great Britain and Northern Ireland and United States of America)	7	0.3240741	0.32	70.69
North Sea Continental Shelf (Federal Republic of Germany/Denmark)	15	0.6944444	0.69	71.39
North Sea Continental Shelf (Federal Republic of Germany/Netherlands)	15	0.6944444	0.69	72.08

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Northern Cameroons (Cameroon v. United Kingdom)	18	0.8333333	0.83	72.92
Nottebohm (Liechtenstein v. Guatemala)	11	0.5092593	0.51	73.43
Nuclear Tests (Australia v. France)	31	1.4351852	1.44	74.86
Nuclear Tests (New Zealand v. France)	29	1.3425926	1.34	76.20
Obligation to Negotiate Access to the Pacific Ocean (Bolivia v. Chile)	14	0.6481481	0.65	76.85
Obligations concerning Negotiations relating to Cessation of the Nuclear Arms Race and to Nuclear Disarmament (Marshall Islands v. India)	17	0.7870370	0.79	77.64
Obligations concerning Negotiations relating to Cessation of the Nuclear Arms Race and to Nuclear Disarmament (Marshall Islands v. Pakistan)	17	0.7870370	0.79	78.43
Obligations concerning Negotiations relating to Cessation of the Nuclear Arms Race and to Nuclear Disarmament (Marshall Islands v. United Kingdom)	17	0.7870370	0.79	79.21
Oil Platforms (Islamic Republic of Iran v. United States of America)	32	1.4814815	1.48	80.69
Passage through the Great Belt (Finland v. Denmark)	7	0.3240741	0.32	81.02
Protection of French Nationals and Protected Persons in Egypt (France v. Egypt)	1	0.0462963	0.05	81.06
Pulp Mills on the River Uruguay (Argentina v. Uruguay)	20	0.9259259	0.93	81.99

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Question of the Delimitation of the Continental Shelf between Nicaragua and Colombia beyond 200 nautical miles from the Nicaraguan Coast (Nicaragua v. Colombia)	13	0.6018519	0.60	82.59
Questions of Interpretation and Application of the 1971 Montreal Convention arising from the Aerial Incident at Lockerbie (Libyan Arab Jamahiriya v. United Kingdom)	28	1.2962963	1.30	83.89
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	1.1574074	1.16	85.05
Questions relating to the Obligation to Prosecute or Extradite (Belgium v. Senegal)	17	0.7870370	0.79	85.83
Questions relating to the Seizure and Detention of Certain Documents and Data (Timor-Leste v. Australia)	10	0.4629630	0.46	86.30
Relocation of the United States Embassy to Jerusalem (Palestine v. United States of America)	1	0.0462963	0.05	86.34
Reparation for Injuries Suffered in the Service of the United Nations	7	0.3240741	0.32	86.67

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
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	0.1851852	0.19	86.85
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	0.6481481	0.65	87.50
Request for Interpretation of the Judgment of 20 November 1950 in the Asylum Case (Colombia v. Peru)	1	0.0462963	0.05	87.55
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	0.0462963	0.05	87.59
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	0.3703704	0.37	87.96
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	0.3703704	0.37	88.33

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Reservations to the Convention on the Prevention and Punishment of the Crime of Genocide	4	0.1851852	0.19	88.52
Right of Passage over Indian Territory (Portugal v. India)	25	1.1574074	1.16	89.68
Rights of Nationals of the United States of America in Morocco (France v. United States of America)	6	0.2777778	0.28	89.95
South West Africa (Ethiopia v. South Africa)	30	1.3888889	1.39	91.34
South West Africa (Liberia v. South Africa)	30	1.3888889	1.39	92.73
Sovereignty over Certain Frontier Land (Belgium/Netherlands)	8	0.3703704	0.37	93.10
Sovereignty over Pedra Branca/Pulau Batu Puteh, Middle Rocks and South Ledge (Malaysia/Singapore)	9	0.4166667	0.42	93.52
Sovereignty over Pulau Ligitan and Pulau Sipadan (Indonesia/Malaysia)	14	0.6481481	0.65	94.17
Status vis-à-vis the Host State of a Diplomatic Envoy to the United Nations (Commonwealth of Dominica v. Switzerland)	1	0.0462963	0.05	94.21
Temple of Preah Vihear (Cambodia v. Thailand)	16	0.7407407	0.74	94.95
Territorial Dispute (Libyan Arab Jamahiriya/Chad)	8	0.3703704	0.37	95.32
Territorial and Maritime Dispute (Nicaragua v. Colombia)	35	1.6203704	1.62	96.94

(continued)

fullname	N	exactpercent	roundedpercent	cumulpercent
Territorial and Maritime Dispute between Nicaragua and Honduras in the Caribbean Sea (Nicaragua v. Honduras)	8	0.3703704	0.37	97.31
Treatment in Hungary of Aircraft and Crew of United States of America (United States of America v. Hungarian People's Republic)	1	0.0462963	0.05	97.36
Treatment in Hungary of Aircraft and Crew of United States of America (United States of America v. Union of Soviet Socialist Republics)	1	0.0462963	0.05	97.41
Trial of Pakistani Prisoners of War (Pakistan v. India)	5	0.2314815	0.23	97.64
United States Diplomatic and Consular Staff in Tehran (United States of America v. Iran)	7	0.3240741	0.32	97.96
Vienna Convention on Consular Relations (Paraguay v. United States of America)	7	0.3240741	0.32	98.29
Voting Procedure on Questions relating to Reports and Petitions concerning the Territory of South West Africa	5	0.2314815	0.23	98.52
Western Sahara	15	0.6944444	0.69	99.21
Whaling in the Antarctic (Australia v. Japan: New Zealand intervening)	17	0.7870370	0.79	100.00
Total	2160	100.0000000	100.00	100.00

Frequency Table for Variable: applicant_region

8 unique value(s) detected.

applicant_region	N	exactpercent	roundedpercent	cumulpercent
Africa	433	20.0462963	20.05	20.05
Americas	371	17.1759259	17.18	37.22
Asia	238	11.0185185	11.02	48.24
Asia|Africa|Asia	6	0.2777778	0.28	48.52
Asia|Africa|Asia|Asia	6	0.2777778	0.28	48.80
Europe	721	33.3796296	33.38	82.18
NA	238	11.0185185	11.02	93.19
Oceania	147	6.8055556	6.81	100.00
Total	2160	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	11.0185185	11.02	11.02
Africa	281	13.0092593	13.01	24.03
Americas	519	24.0277778	24.03	48.06
Asia	280	12.9629630	12.96	61.02
Europe	804	37.2222222	37.22	98.24
Europe|Europe|Americas	7	0.3240741	0.32	98.56
Oceania	31	1.4351852	1.44	100.00
Total	2160	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	3.9351852	3.94	3.94

(continued)

applicant_subregion	N	exactpercent	roundedpercent	cumulpercent
Eastern Europe	43	1.9907407	1.99	5.93
Latin America and the Caribbean	338	15.6481481	15.65	21.57
Micronesia	62	2.8703704	2.87	24.44
NA	229	10.6018519	10.60	35.05
Northern Africa	102	4.7222222	4.72	39.77
Northern America	33	1.5277778	1.53	41.30
Northern Europe	97	4.4907407	4.49	45.79
South-eastern Asia	65	3.0092593	3.01	48.80
Southern Asia	101	4.6759259	4.68	53.47
Southern Europe	358	16.5740741	16.57	70.05
Sub-Saharan Africa	331	15.3240741	15.32	85.37
UNESCO	9	0.4166667	0.42	85.79
Western Asia	72	3.3333333	3.33	89.12
Western Asia|Northern Africa|Western Asia	6	0.2777778	0.28	89.40
Western Asia|Northern Africa|Western Asia|Western Asia	6	0.2777778	0.28	89.68
Western Europe	223	10.3240741	10.32	100.00
Total	2160	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	11.0185185	11.02	11.02
Australia and New Zealand	31	1.4351852	1.44	12.45
Eastern Asia	17	0.7870370	0.79	13.24
Eastern Europe	86	3.9814815	3.98	17.22

(continued)

respondent_subregion	N	exactpercent	roundedpercent	cumulpercent
Latin America and the Caribbean	292	13.5185185	13.52	30.74
Northern Africa	20	0.9259259	0.93	31.67
Northern America	227	10.5092593	10.51	42.18
Northern Europe	222	10.2777778	10.28	52.45
South-eastern Asia	62	2.8703704	2.87	55.32
Southern Asia	115	5.3240741	5.32	60.65
Southern Europe	243	11.2500000	11.25	71.90
Sub-Saharan Africa	261	12.0833333	12.08	83.98
Western Asia	86	3.9814815	3.98	87.96
Western Europe	253	11.7129630	11.71	99.68
Western Europe|Northern Europe|Northern America	7	0.3240741	0.32	100.00
Total	2160	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	2160	100	100	100
Total	2160	100	100	100

Frequency Table for Variable: doi_version

1 unique value(s) detected.

doi_version	N	exactpercent	roundedpercent	cumulpercent
10.5281/zenodo.3826445	2160	100	100	100
Total	2160	100	100	100

(continued)

doi_version	N	exactpercent	roundedpercent	cumulpercent

Frequency Table for Variable: version

1 unique value(s) detected.

version	N	exactpercent	roundedpercent	cumulpercent
2021-11-23	2160	100	100	100
Total	2160	100	100	100

Frequency Table for Variable: license

1 unique value(s) detected.

license	N	exactpercent	roundedpercent	cumulpercent
Creative Commons Zero 1.0 Universal	2160	100	100	100
Total	2160	100	100	100

22 Visualize Frequency Tables

22.1 Load Tables

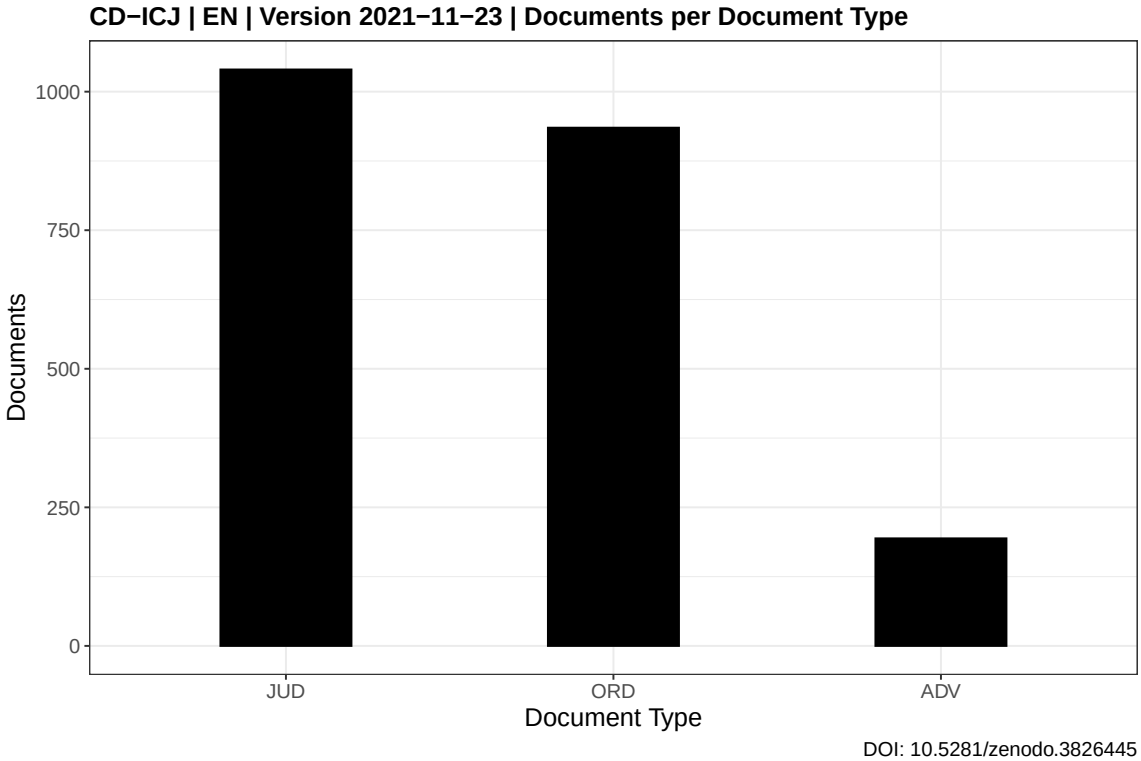
```
prefix.en <- paste0("ANALYSIS/",
                    datashort,
                    "_EN_01_FrequencyTable_var-")

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"))
```

22.2 Doctype

22.2.1 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)
    )
```

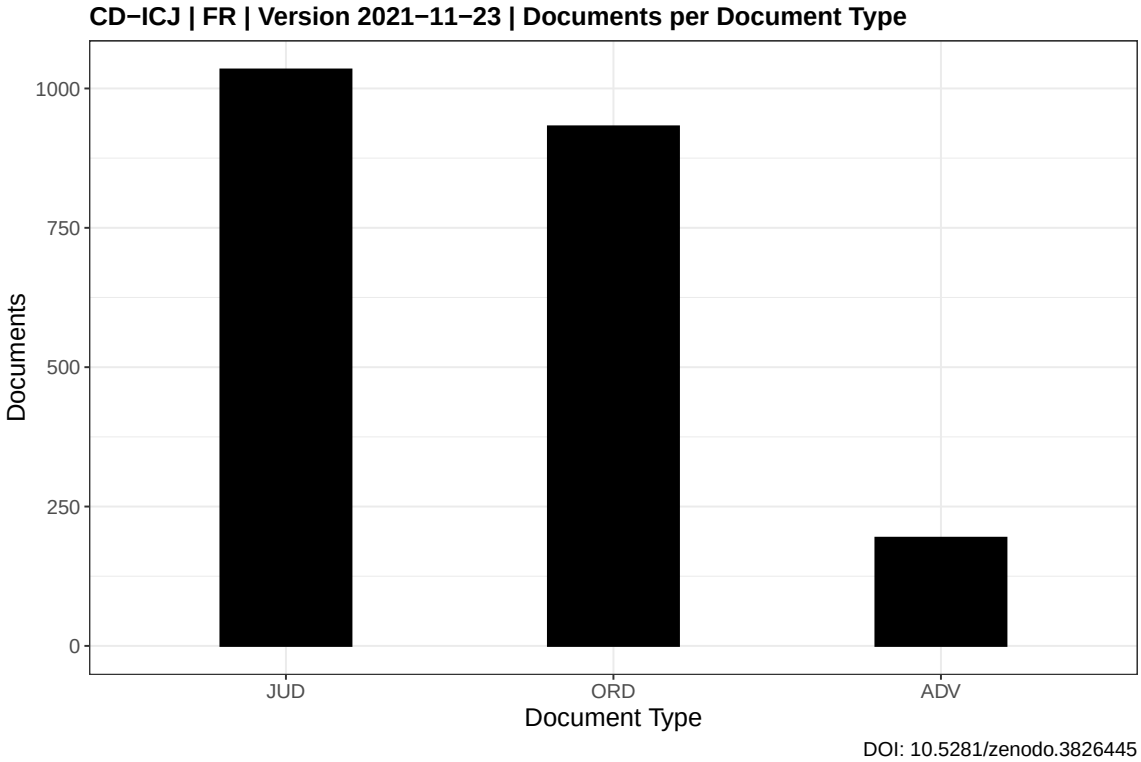


DOI: 10.5281/zenodo.3826445

22.2.2 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)
    )
```



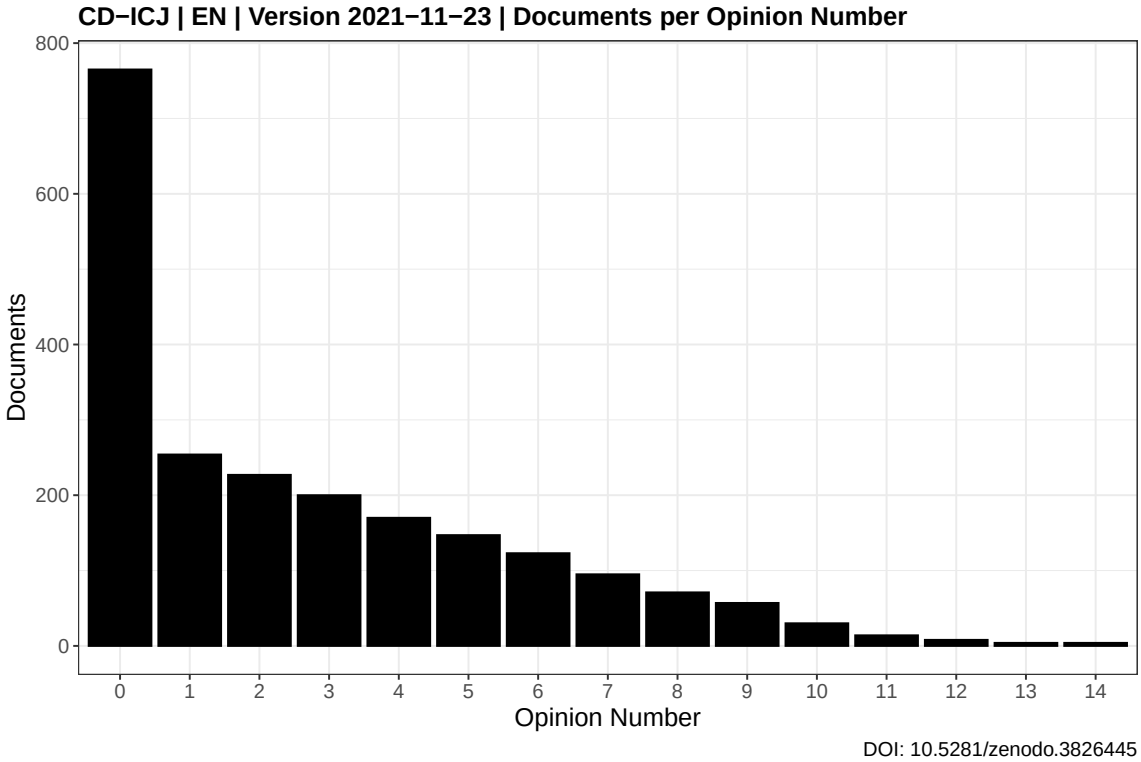
DOI: 10.5281/zenodo.3826445

22.3 Opinion

22.3.1 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)
    )
```

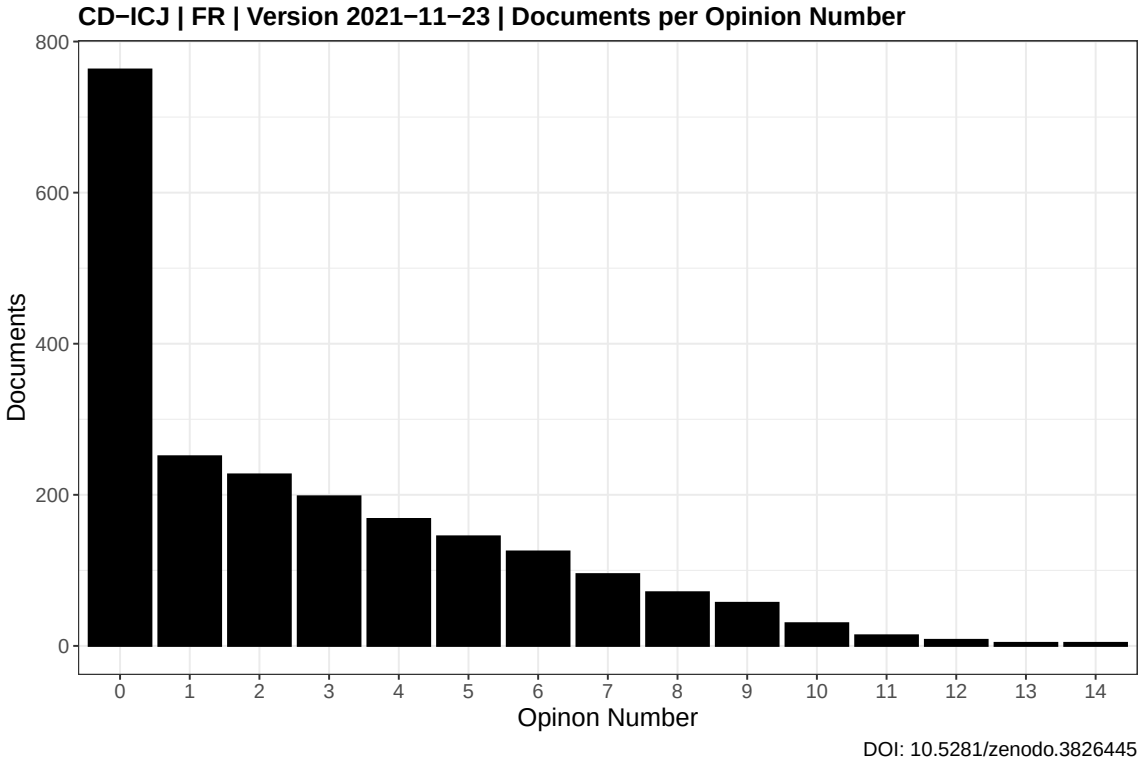


DOI: 10.5281/zenodo.3826445

22.3.2 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)
    )
```



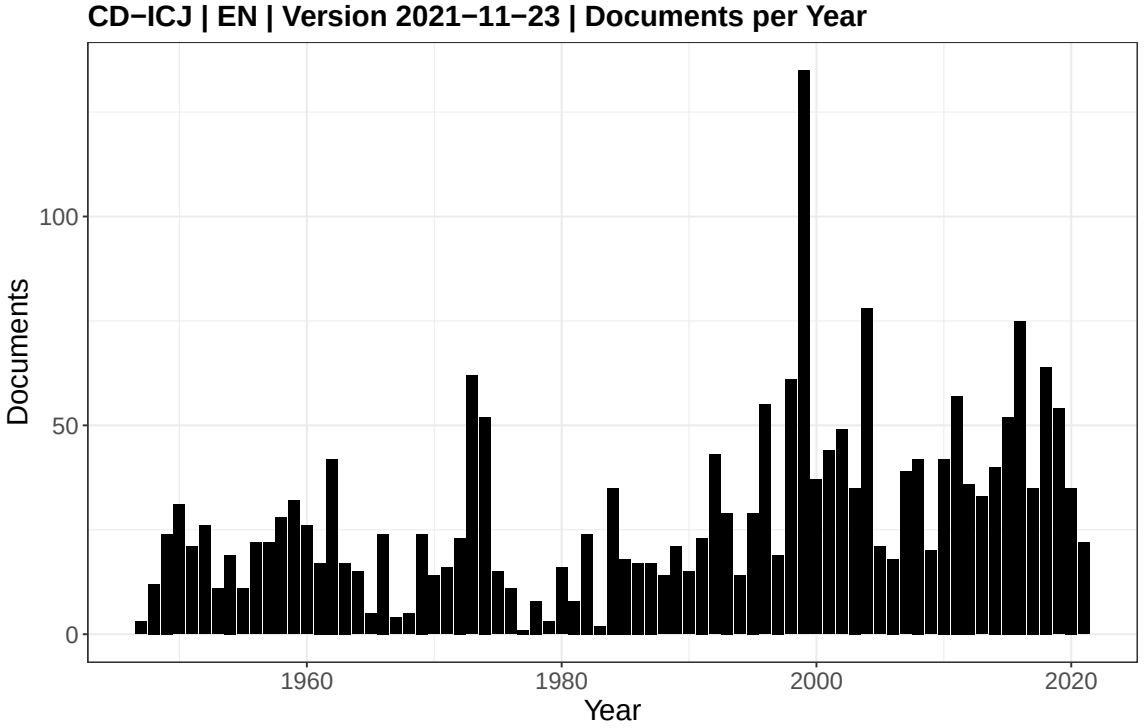
DOI: 10.5281/zenodo.3826445

22.4 Year

22.4.1 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)
    )
```

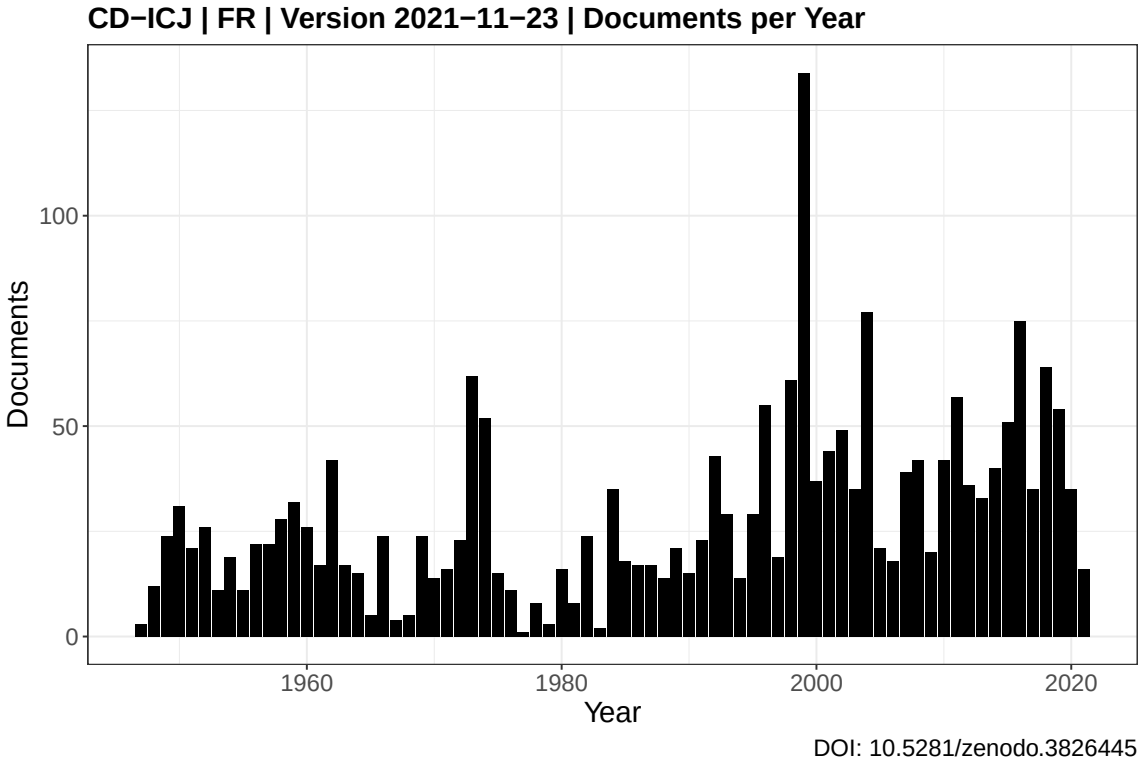


DOI: 10.5281/zenodo.3826445

22.4.2 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.3826445

23 Summary Statistics

23.1 Linguistic Metrics

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

23.1.1 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)
```

}

23.1.2 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 2021-11-23 14:05:40.
    Processing 2169 documents with a total length of 84637041 characters."
## [1] "Runtime was 11.64 secs. Ended at 2021-11-23 14:05:52."
```

```
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 2021-11-23 14:05:52.
    Processing 2160 documents with a total length of 89661729 characters."
## [1] "Runtime was 14.49 secs. Ended at 2021-11-23 14:06:06."
```

23.1.3 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)
```

23.1.4 Create Metadata-only Variants

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

23.1.5 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"))
```

23.1.6 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	84,637,041	376	4,436	16,373	39,021.2268	44,409	744,471
ntokens	15,108,060	71	754	2,895	6,965.4495	8,068	142,584
ntypes	89,901	53	290	720	1,050.9373	1,404	9,995
nsentences	512,598	1	20	94	236.3292	269	5,642

23.1.7 Write Summaries to Disk: English

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

23.1.8 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"))
```

23.1.9 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	89,661,729	398	4,566.50	17,220.0	41,510.0597	47,366.50	817,687
ntokens	15,463,747	69	780.75	2,950.5	7,159.1421	8,437.25	148,563
ntypes	114,131	55	319.00	839.0	1,242.0269	1,691.00	12,090
nsentences	506,597	1	23.00	94.5	234.5356	264.25	5,531

23.1.10 Write Summaries to Disk: French

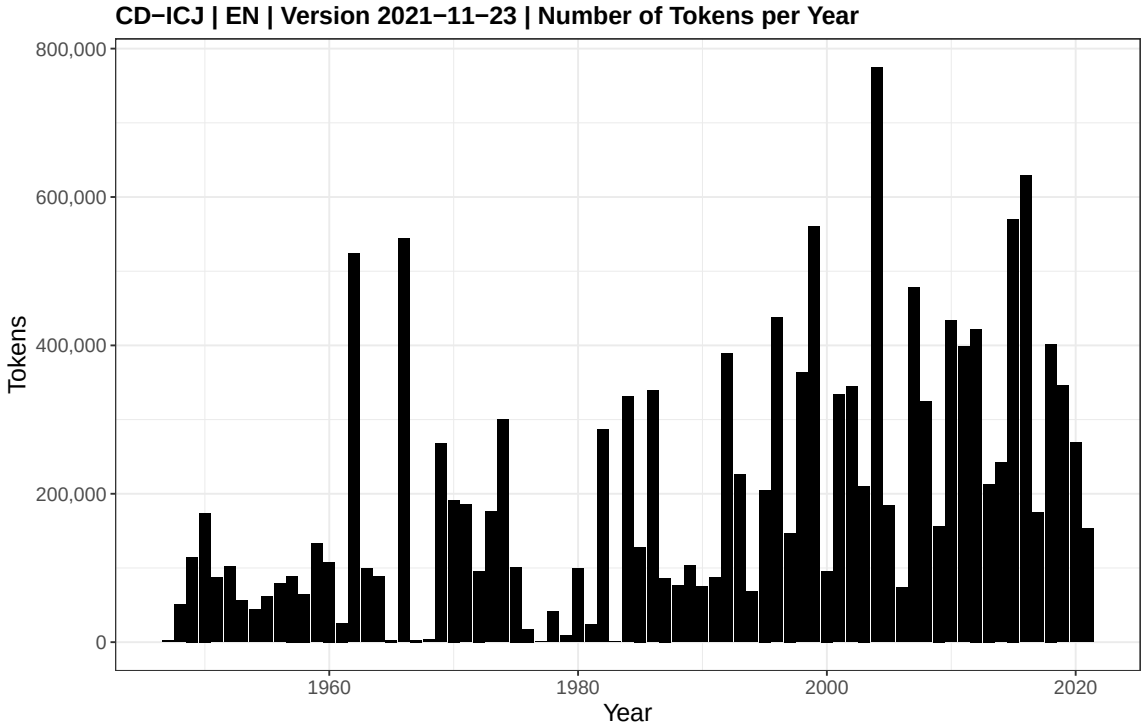
```
fwrite(dt.stats.ling,
       paste0(outputdir,
              datashort,
              "_FR_00_CorpusStatistics_Summaries_Linguistic.csv"),
       na = "NA")
```

23.2 Distributions

23.2.1 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")
    )
)
```

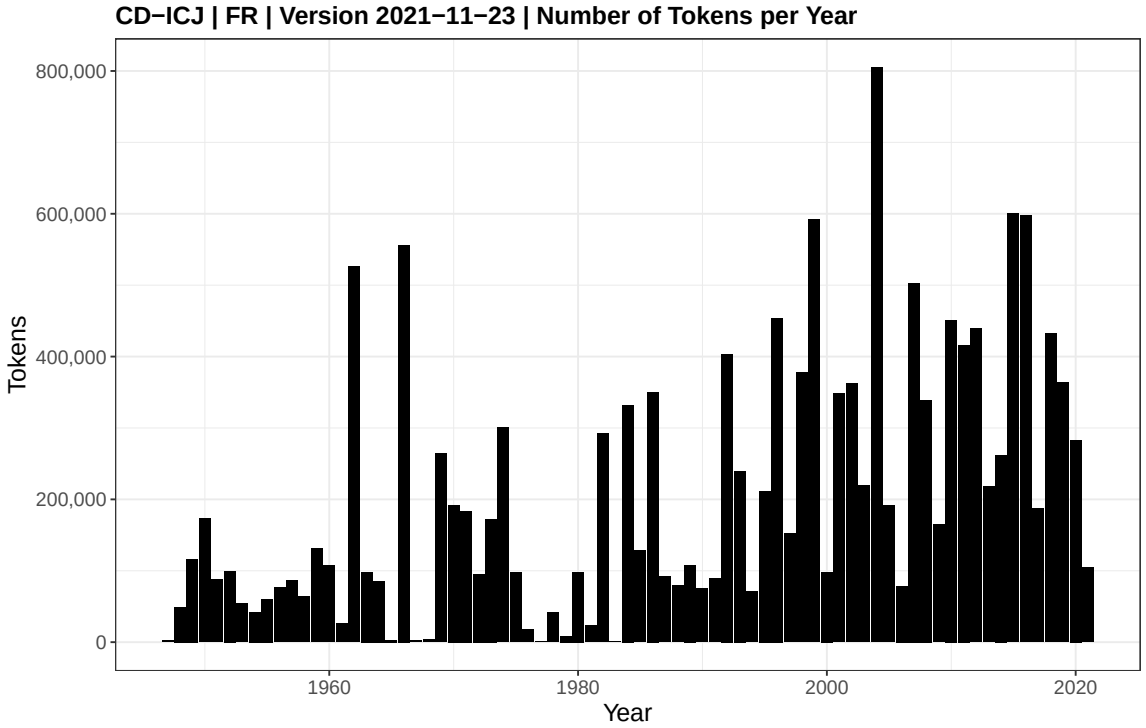


DOI: 10.5281/zenodo.3826445

23.2.2 Tokens per Year: French

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

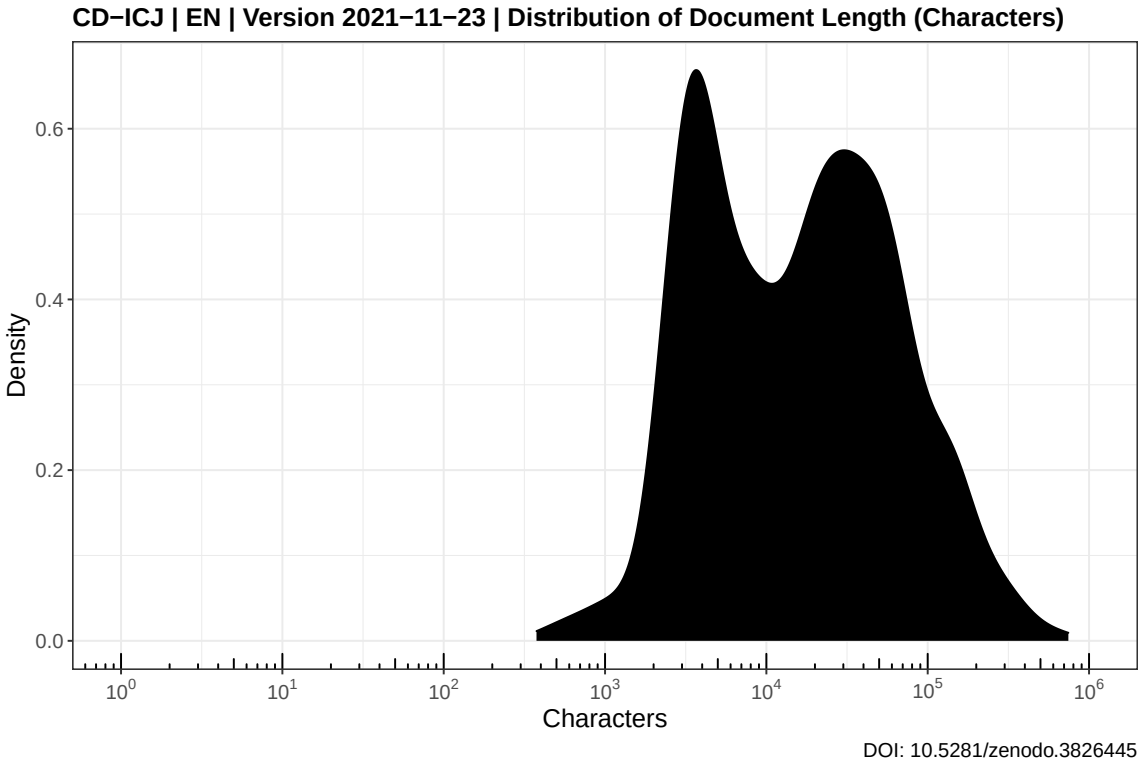
```
print(
    ggplot(data = tokens.year.fr,
           aes(x = year,
               y = V1))+
    geom_bar(stat = "identity",
             fill = "black")+
    scale_y_continuous(labels = comma)+
    theme_bw()+
    labs(
        title = paste(datashort,
                      "| FR | Version",
    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



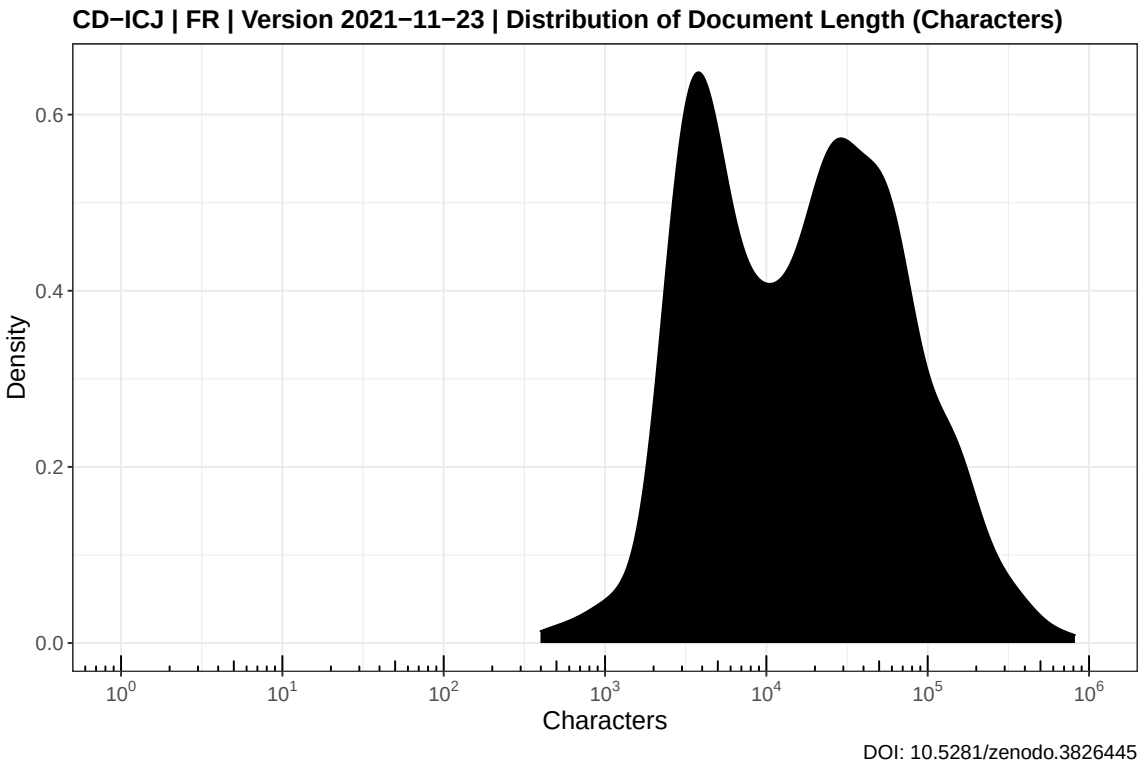
DOI: 10.5281/zenodo.3826445

23.2.3 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)
    )
```

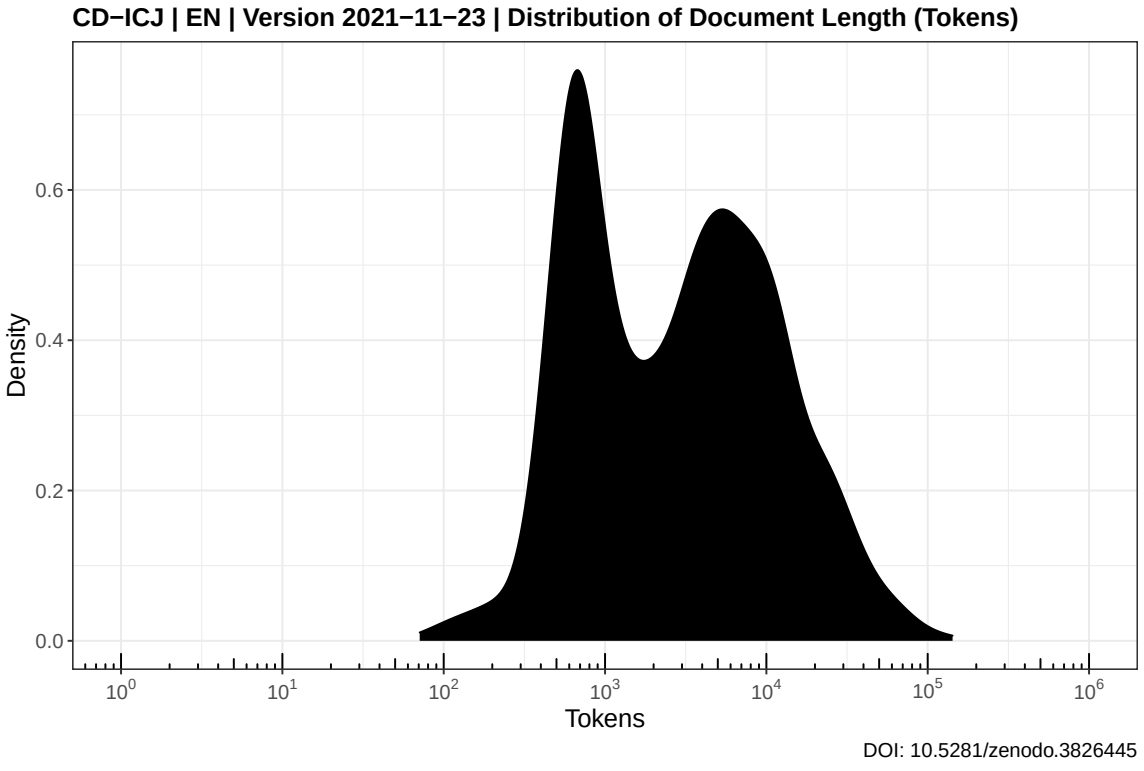


```
ggplot(data = meta.best.fr) +
    geom_density(aes(x = nchars),
                 fill = "black") +
    scale_x_log10(breaks = trans_breaks("log10", function(x) 10^x),
                  labels = trans_format("log10", math_format(10^.x)))+
    annotation_logticks(sides = "b")+
    coord_cartesian(xlim = 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)
    )
```

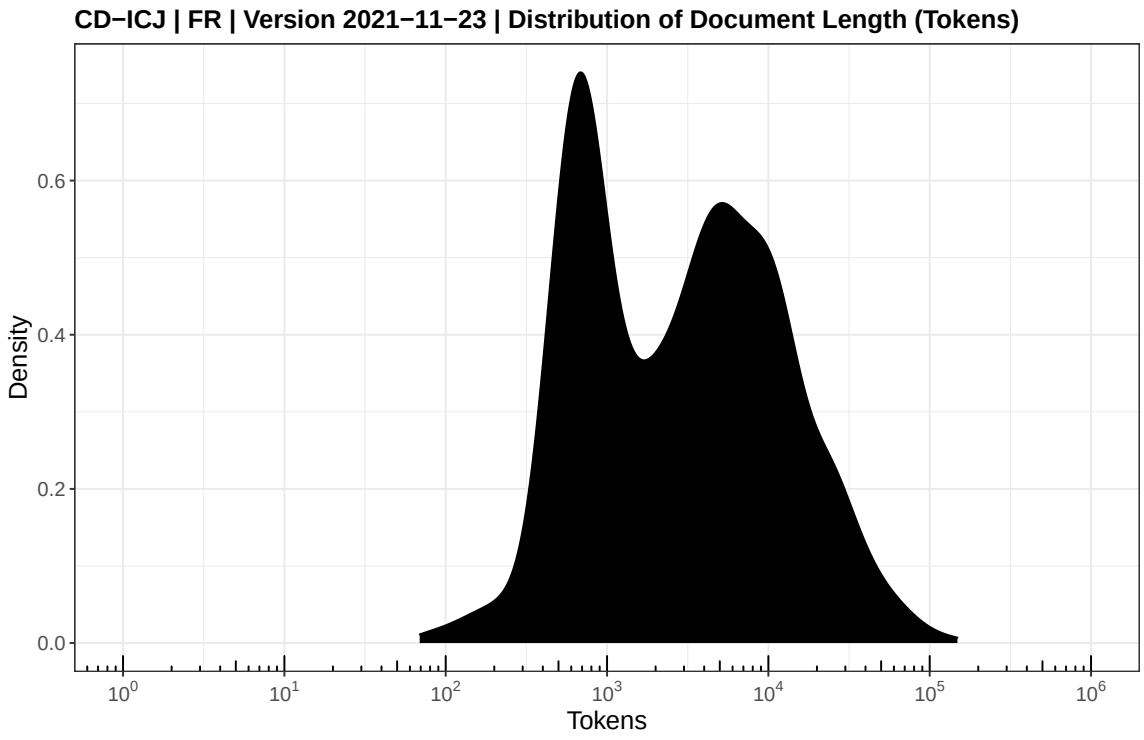


23.2.4 Density: Tokens

```
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)
    )
```

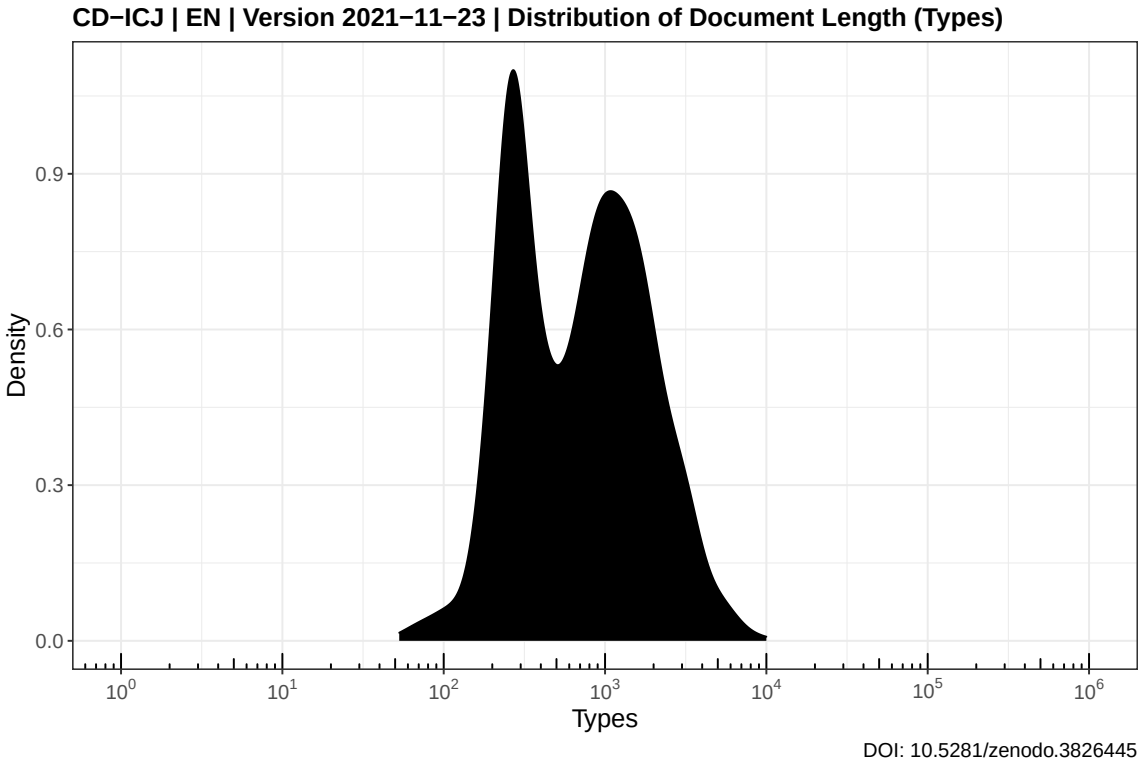


```
ggplot(data = meta.best.fr) +
    geom_density(aes(x = ntokens),
                 fill = "black") +
    scale_x_log10(breaks = trans_breaks("log10", function(x) 10^x),
                  labels = trans_format("log10", math_format(10^.x)))+
    annotation_logticks(sides = "b")+
    coord_cartesian(xlim = 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)
    )
```

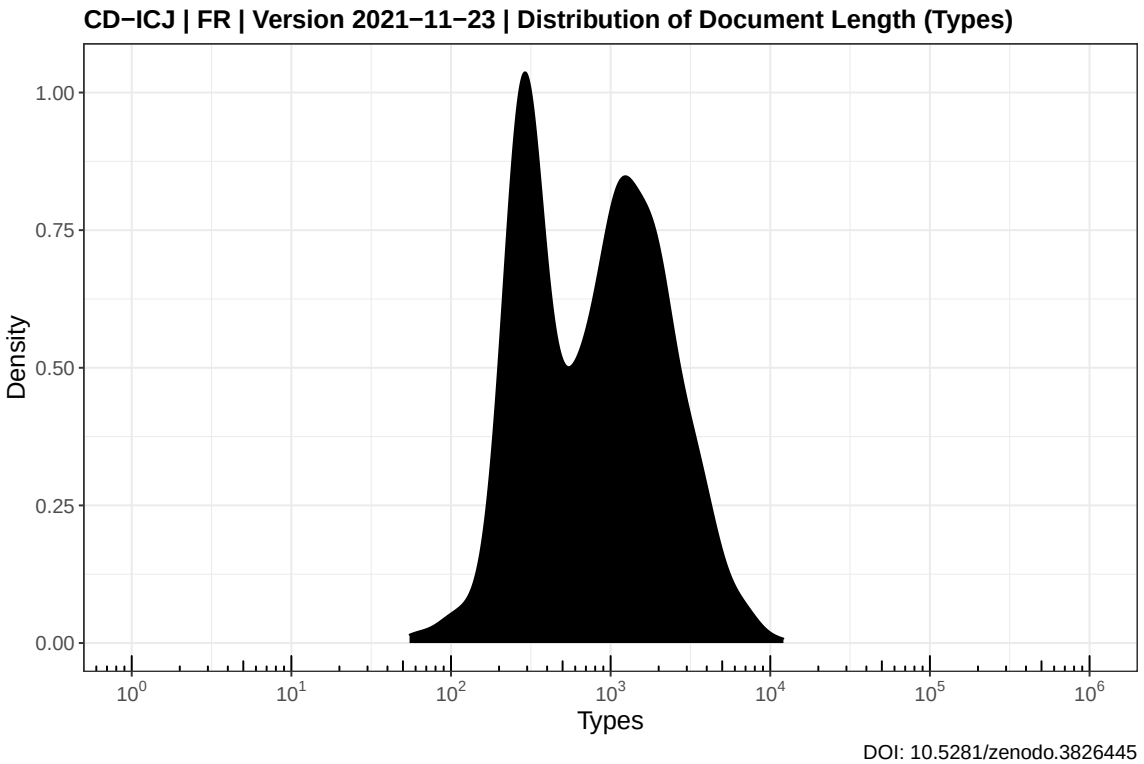


23.2.5 Density: Types

```
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)
    )
```

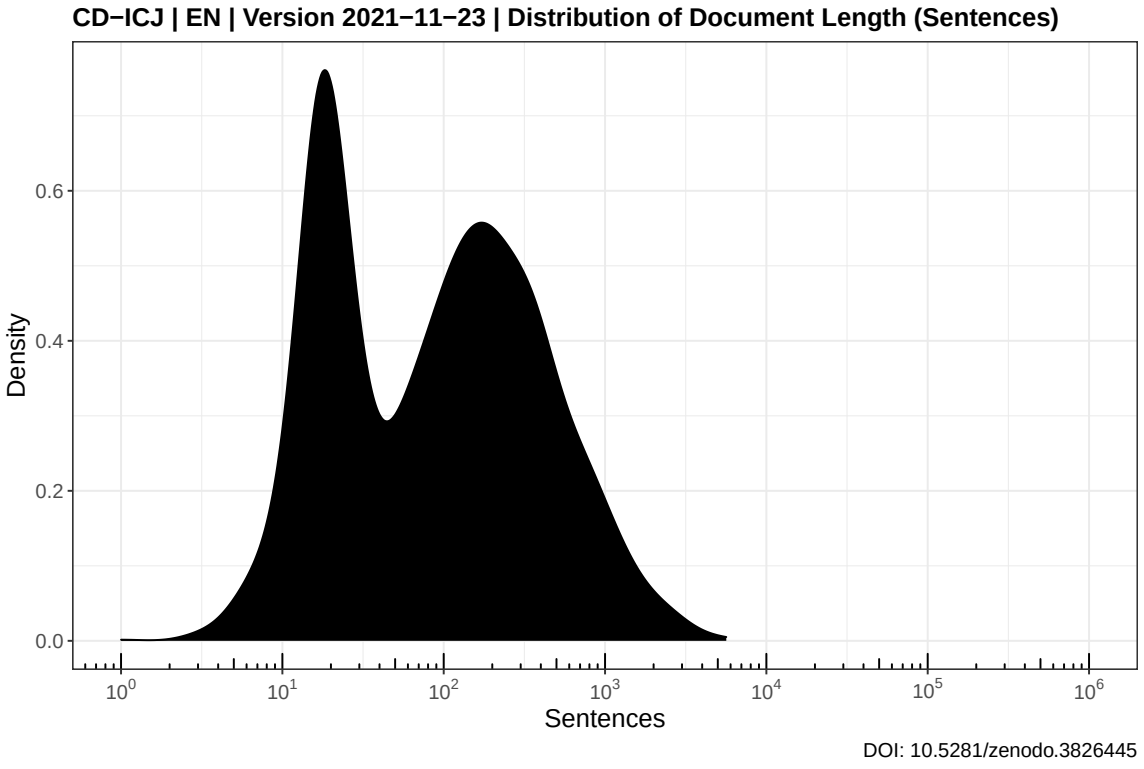


```
ggplot(data = meta.best.fr) +
    geom_density(aes(x = ntypes),
                 fill = "black") +
    scale_x_log10(breaks = trans_breaks("log10", function(x) 10^x),
                  labels = trans_format("log10", math_format(10^.x)))+
    annotation_logticks(sides = "b")+
    coord_cartesian(xlim = 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)
    )
```

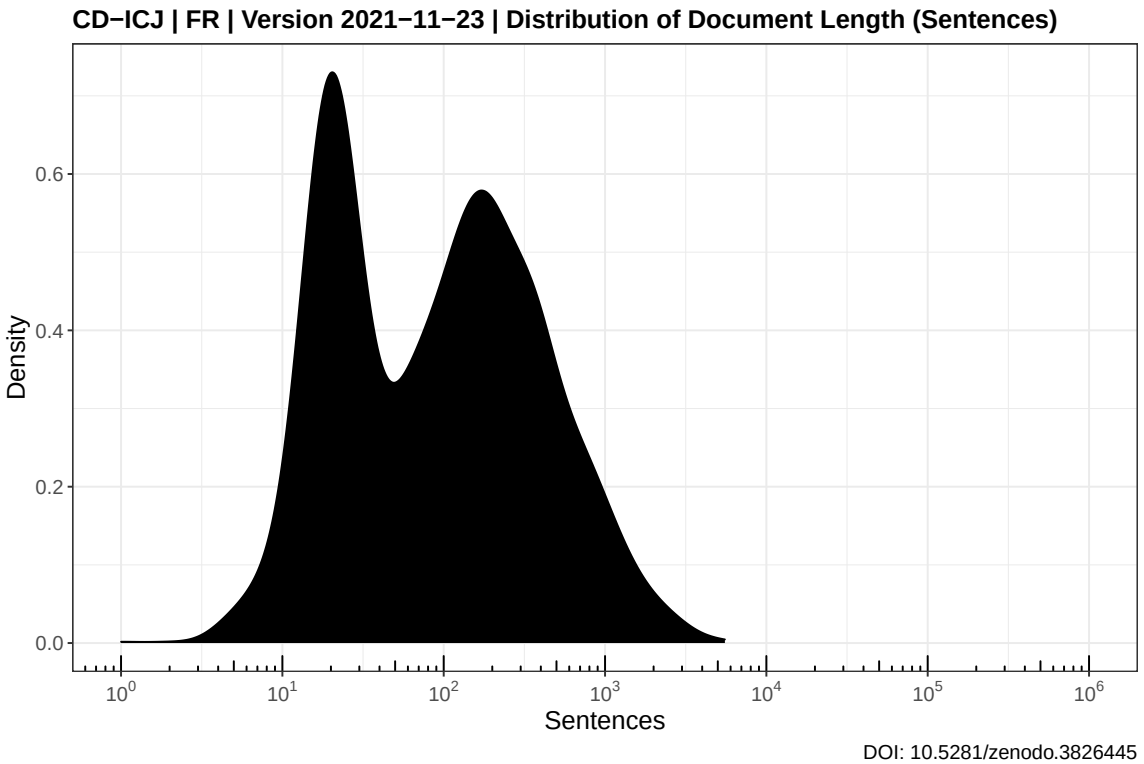


23.2.6 Density: Sentences

```
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)
    )
```



23.2.7 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)
```

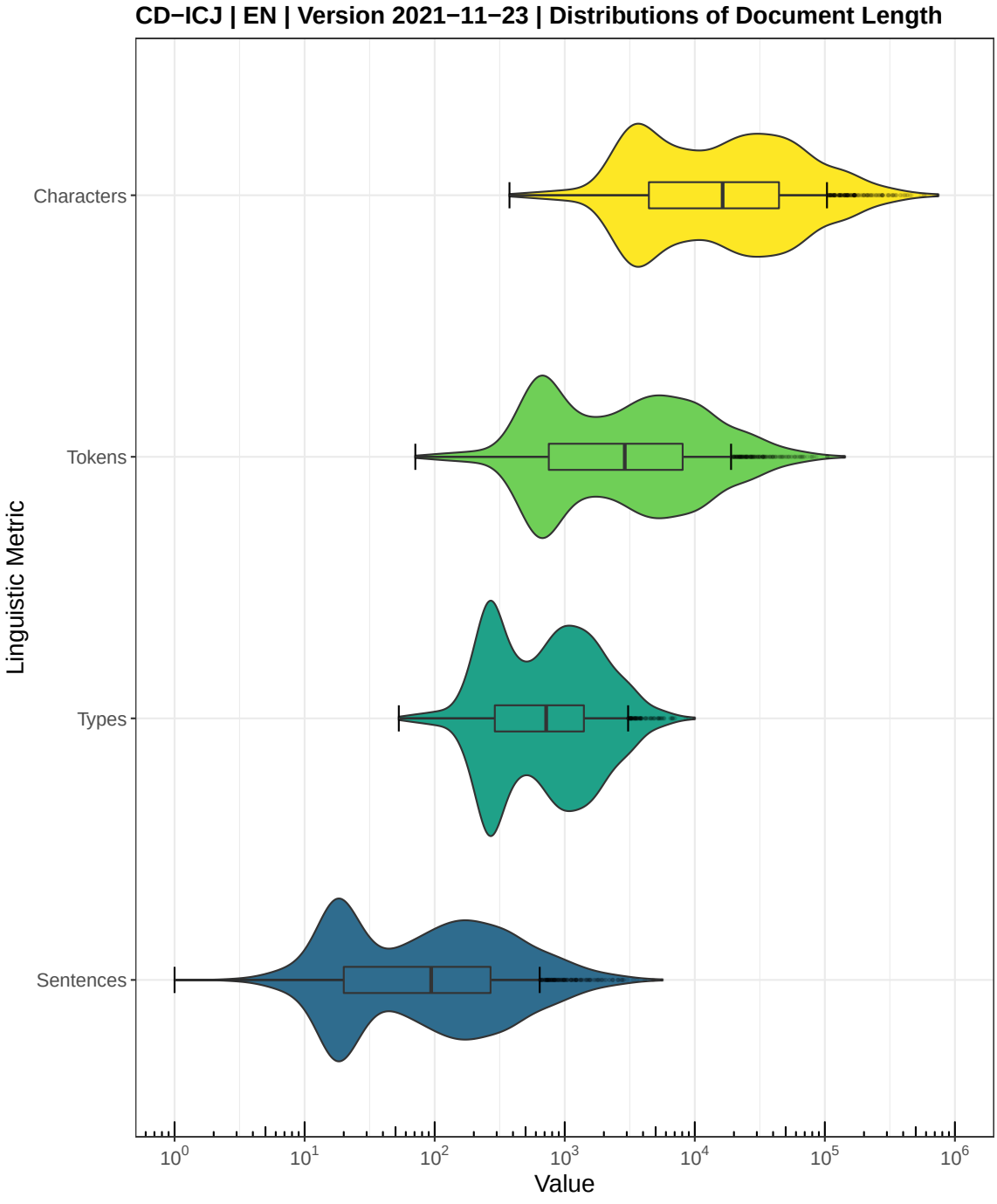
```
## function(x) {
##
##     body = log10(boxplot.stats(10^x)[["stats"]])
##
##     names(body) = c("ymin",
##                     "lower",
##                     "middle",
##                     "upper",
##                     "ymax")
##
##     return(body)
##
## }
```

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

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

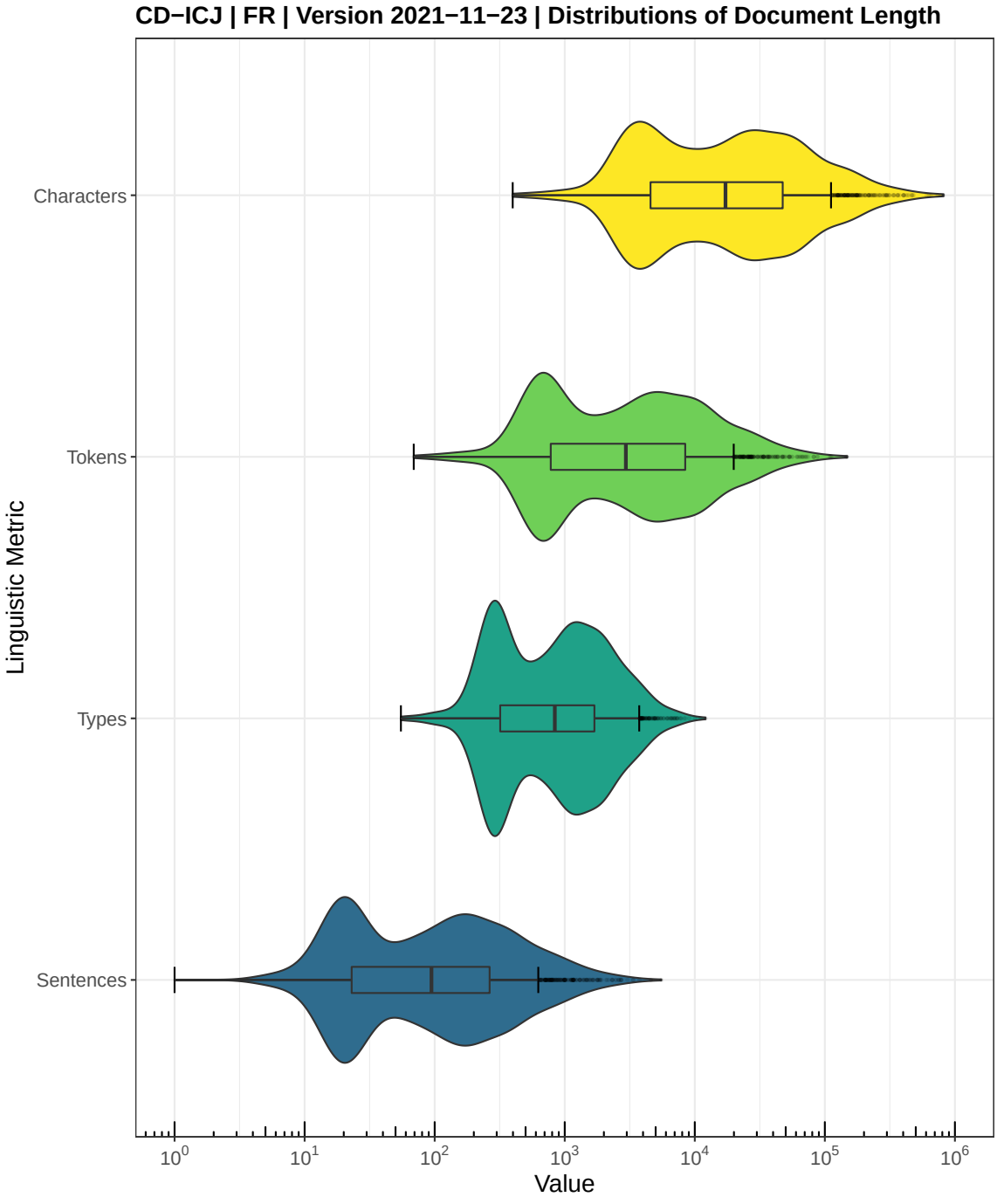
```
dt.allmetrics.en <- melt(summary.corpus.en,
                         measure.vars = rev(c("nchars",
                                              "ntokens",
                                              "ntypes",
                                              "nsentences")))
```

```
ggplot(dt.allmetrics.en, aes(x = value,
                             y = variable,
                             fill = 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() +
    scale_fill_viridis_d(begin = 0.35)+
    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)
    )
```


CD−ICJ | EN | Version 2021−11−23 | Distributions of Document Length
Characters
Tokens
Types
Sentences
Linguistic Metric
10^0
10^1
10^2
10^3
10^4
10^5
10^6
Value
DOI: 10.5281/zenodo.3826445

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

```
ggplot(dt.allmetrics.fr, aes(x = value,
                             y = variable,
                             fill = 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() +
    scale_fill_viridis_d(begin = 0.35)+
    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)
    )
```


CD−ICJ | FR | Version 2021−11−23 | Distributions of Document Length
Characters
Tokens
Types
Sentences
Linguistic Metric
10^0
10^1
10^2
10^3
10^4
10^5
10^6
Value
DOI: 10.5281/zenodo.3826445

23.3 Number of Majority Opinions

23.3.1 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	144
ORD	593
Total	765

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

23.3.2 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	144
ORD	591
Total	763

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

23.4 Number of Minority Opinions

23.4.1 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	896
ORD	342
Total	1404

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

23.4.2 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	890
ORD	341
Total	1397

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

23.5 Year Range

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1947    1973    1998    1992    2010    2021
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1947    1973    1998    1991    2009    2021
```

23.6 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" "1973-07-12" "1998-06-11" "1992-01-11" "2010-03-12" "2021-10-12"
```

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

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## "1947-07-31" "1973-07-12" "1998-06-11" "1991-12-05" "2009-08-13" "2021-10-12"
```

24 Test and Sort Variable Names

24.1 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.

24.1.1 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"))
```

24.1.2 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"))
```

24.2 Number of Variables: Full Data Set

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

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

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

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

24.3 Number of Variables: Metadata

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

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

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

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

24.4 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"
```

```r
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"
```

24.5 List All Variables: Metadata

```r
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"
```

```r
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"
```

25 Calculate Detailed Token Frequencies

25.1 Create Corpora

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

25.2 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)
```

25.3 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)
```

25.4 Most Frequent Tokens | TF Weighting | Tables

25.4.1 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	116549	1	2155	all
article	49390	2	2011	all
international	48416	3	1978	all
case	44807	4	1989	all
states	41884	5	1617	all
united	40859	6	1545	all
law	40795	7	1481	all
judgment	34875	8	1499	all
jurisdiction	31407	9	1528	all
state	31117	10	1503	all
parties	30019	11	1829	all
convention	28804	12	1175	all
p	26985	13	1807	all
paragraph	25295	14	1726	all
application	25123	15	1678	all
nations	24326	16	1285	all
may	24170	17	1797	all
dispute	24161	18	1512	all
legal	22850	19	1450	all
one	21184	20	1973	all
question	21006	21	1546	all
general	20633	22	1806	all
order	19805	23	1945	all
present	19536	24	1787	all
rights	19444	25	1244	all
para	19302	26	1124	all
opinion	18294	27	1511	all
treaty	18161	28	1041	all
also	17092	29	1474	all
government	16536	30	1453	all

(continued)

Feature	Frequency	Rank	Docfreq	Group
republic	16087	31	1275	all
statute	16050	32	1687	all
reports	15972	33	1740	all
two	15665	34	1631	all
op	15270	35	836	all
proceedings	15108	36	1629	all
whether	15084	37	1441	all
right	14721	38	1240	all
measures	14031	39	1023	all
agreement	13909	40	1253	all
must	13710	41	1362	all
view	13265	42	1488	all
nicaragua	13244	43	588	all
made	13218	44	1533	all
part	13047	45	1386	all
v	12876	46	1483	all
fact	12527	47	1364	all
upon	12264	48	1425	all
decision	12246	49	1756	all
court's	12205	50	1443	all
justice	12026	51	1620	all
first	11755	52	1419	all
within	11567	53	1584	all
rules	11540	54	1635	all
council	11470	55	723	all
respect	11424	56	1441	all
can	11394	57	1345	all
territory	11364	58	981	all
interpretation	11107	59	1155	all

(continued)

Feature	Frequency	Rank	Docfreq	Group
force	11101	60	1090	all
time	11033	61	1407	all
however	10812	62	1414	all
concerning	10744	63	1633	all
basis	10680	64	1340	all
assembly	10640	65	686	all
declaration	10556	66	1195	all
principle	10442	67	1097	all
regard	10408	68	1771	all
thus	10407	69	1404	all
request	10349	70	1344	all
obligation	10343	71	1072	all
line	10332	72	586	all
genocide	10312	73	451	all
use	10146	74	1022	all
delimitation	10082	75	431	all
mr	10042	76	670	all
maritime	10036	77	626	all
party	10019	78	1301	all
point	9921	79	1207	all
certain	9855	80	1360	all
claim	9714	81	1058	all
boundary	9708	82	507	all
pp	9568	83	976	all
nuclear	9471	84	385	all
even	9402	85	1296	all
obligations	9331	86	1019	all
preliminary	9281	87	1043	all
provisional	9192	88	766	all

(continued)

Feature	Frequency	Rank	Docfreq	Group
new	9168	89	1084	all
continental	9084	90	447	all
cases	9060	91	1145	all
shelf	8944	92	424	all
particular	8906	93	1304	all
yugoslavia	8872	94	422	all
resolution	8862	95	824	all
provisions	8781	96	1152	all
judge	8743	97	1577	all
effect	8686	98	1226	all
whereas	8683	99	1074	all
charter	8576	100	788	all

25.4.2 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	117425	1	2150	all
droit	55596	2	1624	all
l'article	42622	3	1964	all

(continued)

Feature	Frequency	Rank	Docfreq	Group
comme	37836	4	1837	all
p	35914	5	1835	all
fait	34280	6	2064	all
parties	31767	7	1843	all
être	31152	8	1720	all
d'une	30115	9	1744	all
si	29956	10	1611	all
entre	29176	11	1650	all
convention	29073	12	1179	all
plus	27696	13	1561	all
international	27675	14	1810	all
question	27656	15	1617	all
d'un	27329	16	1767	all
nations	27165	17	1275	all
qu'il	27003	18	1693	all
paragraphe	25726	19	1687	all
compétence	23070	20	1421	all
etats	22745	21	1354	all
différend	20686	22	1432	all
deux	20632	23	1691	all
statut	20003	24	1791	all
arrêt	19996	25	1233	all
peut	19897	26	1488	all
unies	19163	27	1220	all
droits	19111	28	1238	all
ainsi	19086	29	1797	all
tout	17896	30	1532	all
non	17735	31	1452	all
demande	17695	32	1495	all

(continued)

Feature	Frequency	Rank	Docfreq	Group
partie	17181	33	1487	all
gouvernement	16910	34	1447	all
dont	16789	35	1916	all
qu'elle	16683	36	1566	all
traité	16250	37	1029	all
république	16120	38	1283	all
l'affaire	15799	39	1570	all
op	15484	40	865	all
mesures	15424	41	1116	all
internationale	15363	42	1772	all
n'a	15317	43	1475	all
n'est	15258	44	1451	all
selon	15109	45	1392	all
cas	15027	46	1410	all
nicaragua	14859	47	583	all
recueil	14790	48	1717	all
point	14744	49	1371	all
juridique	14635	50	1299	all
laquelle	13579	51	1757	all
conseil	13576	52	1166	all
doit	13468	53	1548	all
procédure	13344	54	1688	all
requête	13315	55	1409	all
etats-unis	13136	56	874	all
bien	12734	57	1408	all
déclaration	12652	58	1300	all
justice	12487	59	1661	all
aussi	12083	60	1368	all
générale	12036	61	1101	all

(continued)

Feature	Frequency	Rank	Docfreq	Group
etat	12031	62	1156	all
contre	11926	63	1514	all
présente	11889	64	1542	all
l'arrêt	11775	65	1233	all
territoire	11763	66	989	all
décision	11646	67	1373	all
c'est	11639	68	1362	all
où	11634	69	1380	all
règlement	11319	70	1607	all
voir	11153	71	1159	all
l'etat	11022	72	1012	all
général	10975	73	1613	all
principe	10822	74	1131	all
faire	10706	75	1463	all
société	10643	76	527	all
donc	10543	77	1338	all
savoir	10510	78	1438	all
délimitation	10371	79	440	all
questions	10295	80	1346	all
autre	10275	81	1305	all
l'organisation	10168	82	880	all
devant	10093	83	1319	all
ni	10086	84	1291	all
affaire	10030	85	1630	all
dispositions	9947	86	1192	all
autres	9899	87	1726	all
l'assemblée	9893	88	672	all
génocide	9866	89	446	all
toute	9780	90	1352	all

(continued)

Feature	Frequency	Rank	Docfreq	Group
sens	9771	91	1274	all
avis	9675	92	1654	all
ligne	9509	93	499	all
tant	9296	94	1217	all
mandat	9226	95	348	all
termes	9213	96	1292	all
titre	9206	97	1194	all
fond	9167	98	1287	all
yougoslavie	9107	99	408	all
obligations	8971	100	1021	all

25.5 Most Frequent Tokens | TFIDF Weighting | Tables

25.5.1 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
convention	7668.249	1	1175	all
nicaragua	7507.788	2	588	all
nuclear	7110.816	3	385	all

(continued)

Feature	Weight	Rank	Docfreq	Group
delimitation	7075.369	4	431	all
league	7036.103	5	314	all
genocide	7033.640	6	451	all
mandate	6934.181	7	250	all
law	6759.915	8	1481	all
diss	6359.194	9	366	all
shelf	6340.345	10	424	all
op	6322.594	11	836	all
yugoslavia	6307.523	12	422	all
continental	6231.188	13	447	all
boundary	6128.190	14	507	all
costa	6107.420	15	238	all
united	6019.800	16	1545	all
line	5872.316	17	586	all
treaty	5789.887	18	1041	all
judgment	5595.970	19	1499	all
nations	5530.672	20	1285	all
para	5510.591	21	1124	all
africa	5484.429	22	431	all
council	5472.581	23	723	all
maritime	5416.281	24	626	all
states	5342.285	25	1617	all
assembly	5319.313	26	686	all
mr	5123.275	27	670	all
weapons	5083.309	28	240	all
islands	5005.073	29	282	all
state	4956.956	30	1503	all
south	4918.447	31	548	all
jurisdiction	4778.142	32	1528	all

(continued)

Feature	Weight	Rank	Docfreq	Group
rica	4722.963	33	229	all
rights	4694.543	34	1244	all
sep	4640.476	35	488	all
measures	4579.493	36	1023	all
qatar	4573.927	37	121	all
human	4513.798	38	563	all
federal	4471.704	39	554	all
river	4424.106	40	267	all
mandates	4205.621	41	109	all
el	4191.892	42	223	all
provisional	4155.067	43	766	all
legal	3996.272	44	1450	all
sea	3937.685	45	597	all
territory	3915.927	46	981	all
honduras	3892.451	47	251	all
bahrain	3889.015	48	94	all
dispute	3786.216	49	1512	all
equidistance	3775.724	50	103	all
salvador	3774.233	51	139	all
charter	3771.153	52	788	all
area	3763.270	53	645	all
chamber	3756.446	54	259	all
serbia	3755.332	55	278	all
resolution	3724.985	56	824	all
republic	3712.065	57	1275	all
indb	3672.541	58	277	all
right	3574.816	59	1240	all
territorial	3482.383	60	705	all
tribunal	3472.111	61	596	all

(continued)

Feature	Weight	Rank	Docfreq	Group
security	3362.508	62	785	all
respondent	3359.085	63	747	all
pp	3318.276	64	976	all
force	3317.346	65	1090	all
use	3315.801	66	1022	all
agreement	3314.633	67	1253	all
disarmament	3295.808	68	95	all
fry	3196.336	69	64	all
cerd	3192.948	70	87	all
obligation	3165.628	71	1072	all
cançado	3104.866	72	196	all
principle	3091.385	73	1097	all
trindade	3091.001	74	197	all
question	3088.934	75	1546	all
coast	3075.287	76	244	all
obligations	3061.365	77	1019	all
interpretation	3039.737	78	1155	all
claim	3028.572	79	1058	all
trusteeship	3014.962	80	107	all
ibid	3005.032	81	760	all
member	2999.044	82	708	all
kingdom	2979.525	83	831	all
west	2978.357	84	459	all
treaties	2961.809	85	773	all
committee	2958.880	86	556	all
preliminary	2951.128	87	1043	all
island	2949.826	88	254	all
government	2877.164	89	1453	all
opinion	2872.068	90	1511	all

(continued)

Feature	Weight	Rank	Docfreq	Group
bosnia	2870.686	91	257	all
also	2867.389	92	1474	all
montenegro	2852.224	93	205	all
commission	2845.263	94	679	all
protection	2833.795	95	800	all
applicant	2831.128	96	869	all
evidence	2830.098	97	824	all
mandatory	2823.776	98	239	all
application	2800.400	99	1678	all
drc	2777.896	100	54	all

25.5.2 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	8451.579	1	583	all
convention	7644.453	2	1179	all
mandat	7315.060	3	348	all
délimitation	7166.372	4	440	all

(continued)

Feature	Weight	Rank	Docfreq	Group
droit	6886.550	5	1624	all
génocide	6759.383	6	446	all
yougoslavie	6591.588	7	408	all
société	6520.361	8	527	all
plateau	6237.222	9	424	all
diss	6226.813	10	367	all
nations	6219.252	11	1275	all
op	6153.924	12	865	all
ligne	6051.083	13	499	all
continental	5993.433	14	433	all
frontière	5401.360	15	531	all
traité	5233.124	16	1029	all
nucléaires	5193.320	17	381	all
rica	5168.064	18	224	all
etats-unis	5161.690	19	874	all
costa	5100.002	20	230	all
l'assemblée	5016.587	21	672	all
arrêt	4868.840	22	1233	all
qatar	4851.384	23	116	all
unies	4754.224	24	1220	all
zone	4626.561	25	492	all
droits	4619.761	26	1238	all
etats	4613.484	27	1354	all
mer	4540.970	28	605	all
ind	4511.906	29	464	all
mesures	4423.442	30	1116	all
conservatoires	4410.784	31	638	all
maritime	4308.965	32	603	all
compétence	4195.503	33	1421	all

(continued)

Feature	Weight	Rank	Docfreq	Group
mandats	4193.447	34	124	all
bahreïn	4101.174	35	89	all
salvador	4053.645	36	142	all
honduras	4049.425	37	254	all
résolution	4043.070	38	702	all
territoire	3990.685	39	989	all
l'organisation	3965.226	40	880	all
armes	3922.120	41	270	all
plus	3906.544	42	1561	all
si	3815.143	43	1611	all
africain	3760.388	44	357	all
charte	3743.640	45	777	all
différend	3692.673	46	1432	all
tribunal	3676.197	47	686	all
indb	3665.320	48	276	all
république	3646.781	49	1283	all
conseil	3635.045	50	1166	all
l'etat	3629.250	51	1012	all
fédérale	3603.768	52	452	all
tutelle	3550.900	53	161	all
générale	3522.533	54	1101	all
îles	3484.252	55	286	all
question	3477.569	56	1617	all
commission	3474.884	57	763	all
entre	3412.711	58	1650	all
défendeur	3404.587	59	734	all
rfy	3383.727	60	79	all
mandataire	3314.116	61	110	all
etat	3266.367	62	1156	all

(continued)

Feature	Weight	Rank	Docfreq	Group
l'homme	3254.472	63	420	all
juridique	3232.061	64	1299	all
sécurité	3221.791	65	818	all
peut	3220.346	66	1488	all
traités	3110.158	67	813	all
colombie	3106.856	68	247	all
fleuve	3091.090	69	190	all
côte	3082.048	70	227	all
être	3081.721	71	1720	all
non	3059.059	72	1452	all
sud-ouest	3057.067	73	394	all
milles	3043.630	74	199	all
principe	3040.886	75	1131	all
membres	3026.601	76	890	all
voir	3015.440	77	1159	all
libye	2980.974	78	149	all
l'emploi	2980.310	79	566	all
pacte	2974.917	80	330	all
juridiction	2974.767	81	970	all
rdc	2965.413	82	54	all
cameroun	2952.707	83	356	all
gouvernement	2942.090	84	1447	all
cançado	2936.906	85	194	all
trindade	2925.392	86	194	all
obligations	2919.415	87	1021	all
ibid	2915.286	88	753	all
point	2910.706	89	1371	all
royaume-uni	2901.692	90	683	all
membre	2885.793	91	668	all

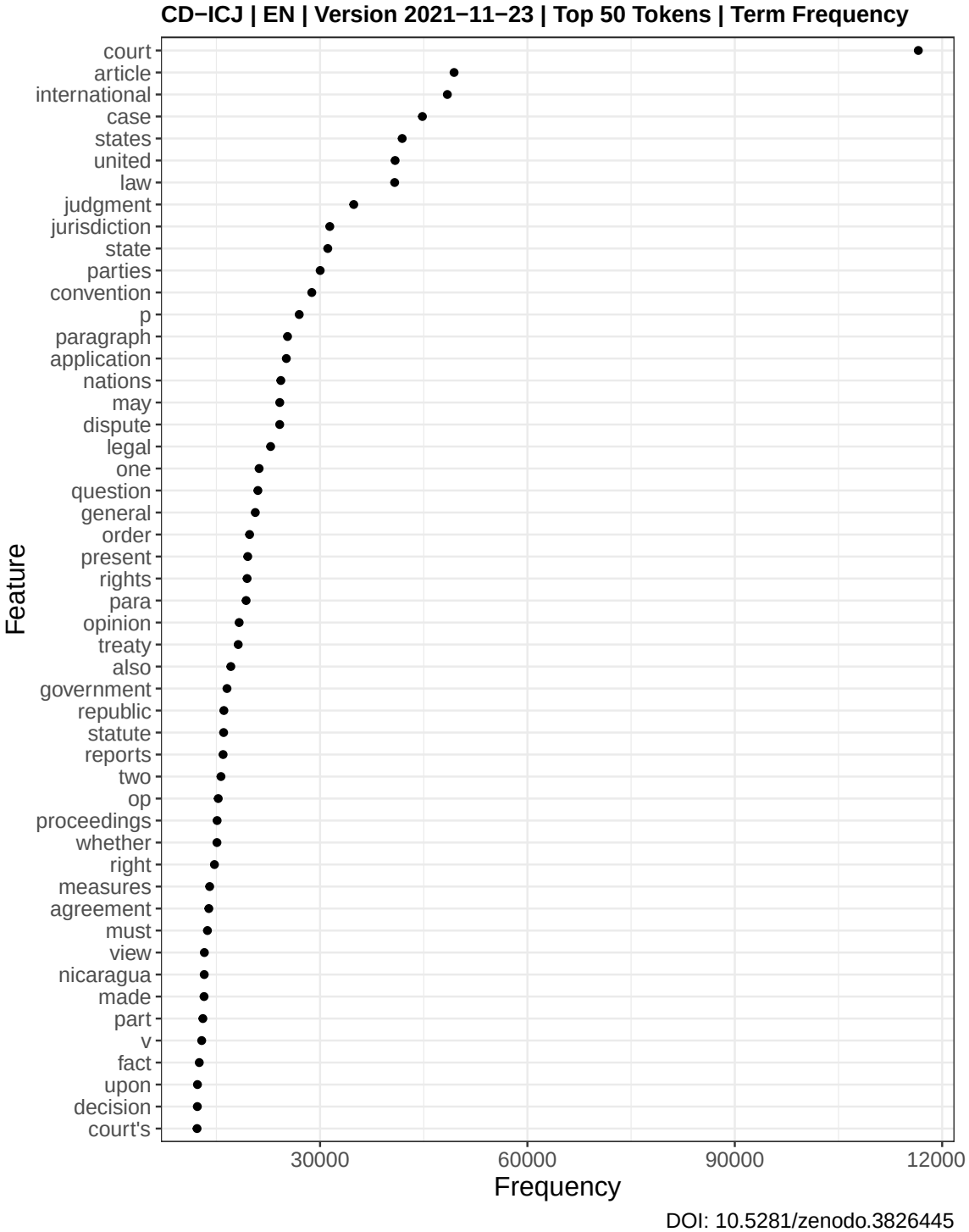
(continued)

Feature	Weight	Rank	Docfreq	Group
selon	2883.017	92	1392	all
force	2882.835	93	861	all
l'arrêt	2867.103	94	1233	all
congo	2865.781	95	268	all
qu'il	2856.831	96	1693	all
bosnie-herzégovine	2847.539	97	246	all
demande	2827.883	98	1495	all
demandeur	2813.379	99	814	all
consultatif	2809.056	100	632	all

25.6 Most Frequent Tokens | TF Weighting | Scatterplots

25.6.1 English

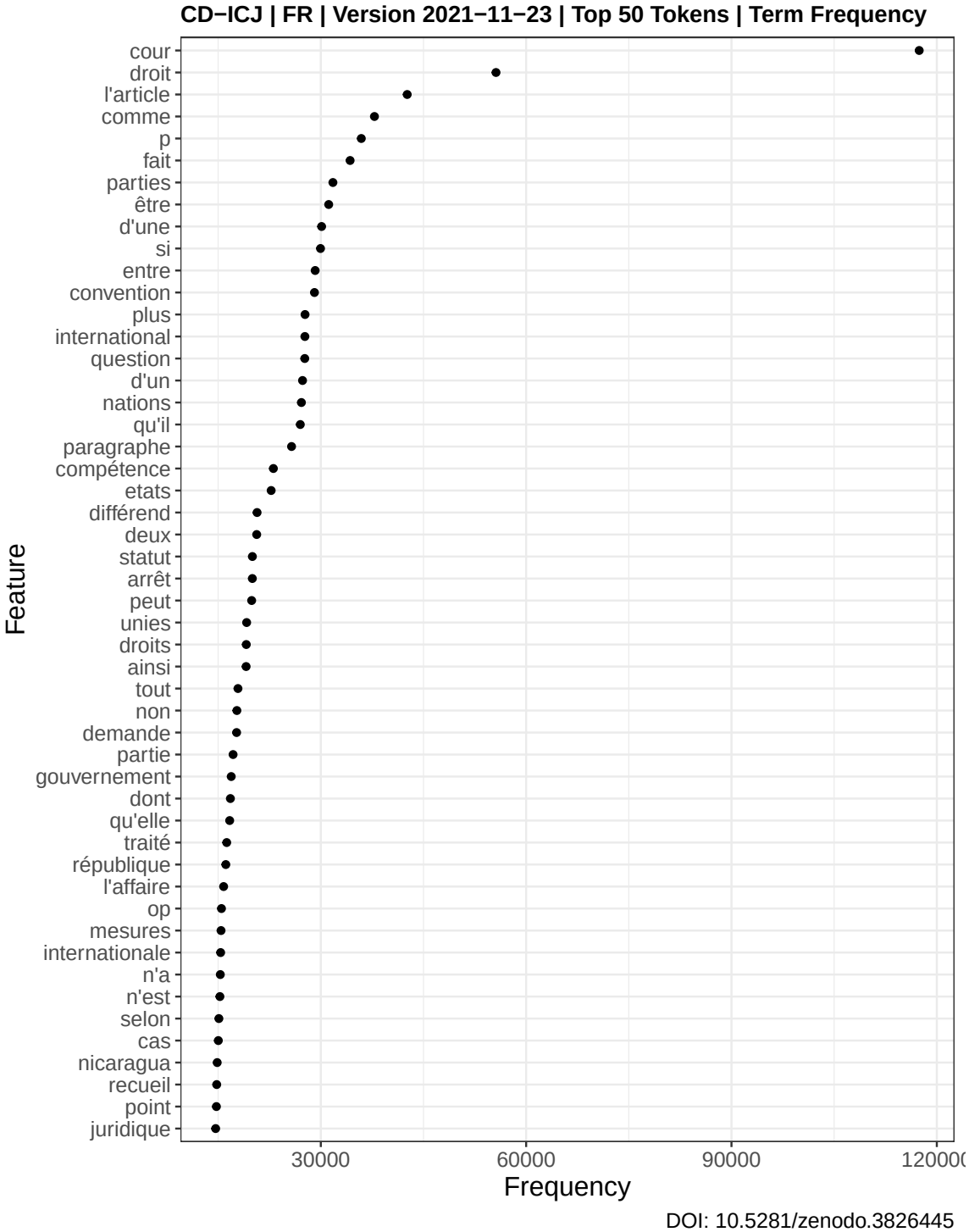
```
print(
    ggplot(data = tstat.en[1:50, ],
           aes(x = reorder(feature,
                           frequency),
               y = frequency))+
    geom_point()+
    coord_flip()+
    labs(
        title = paste(datashort,
                      "| EN | Version",
                      datestamp,
                      "| 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.3826445

25.6.2 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")
    )
)
```

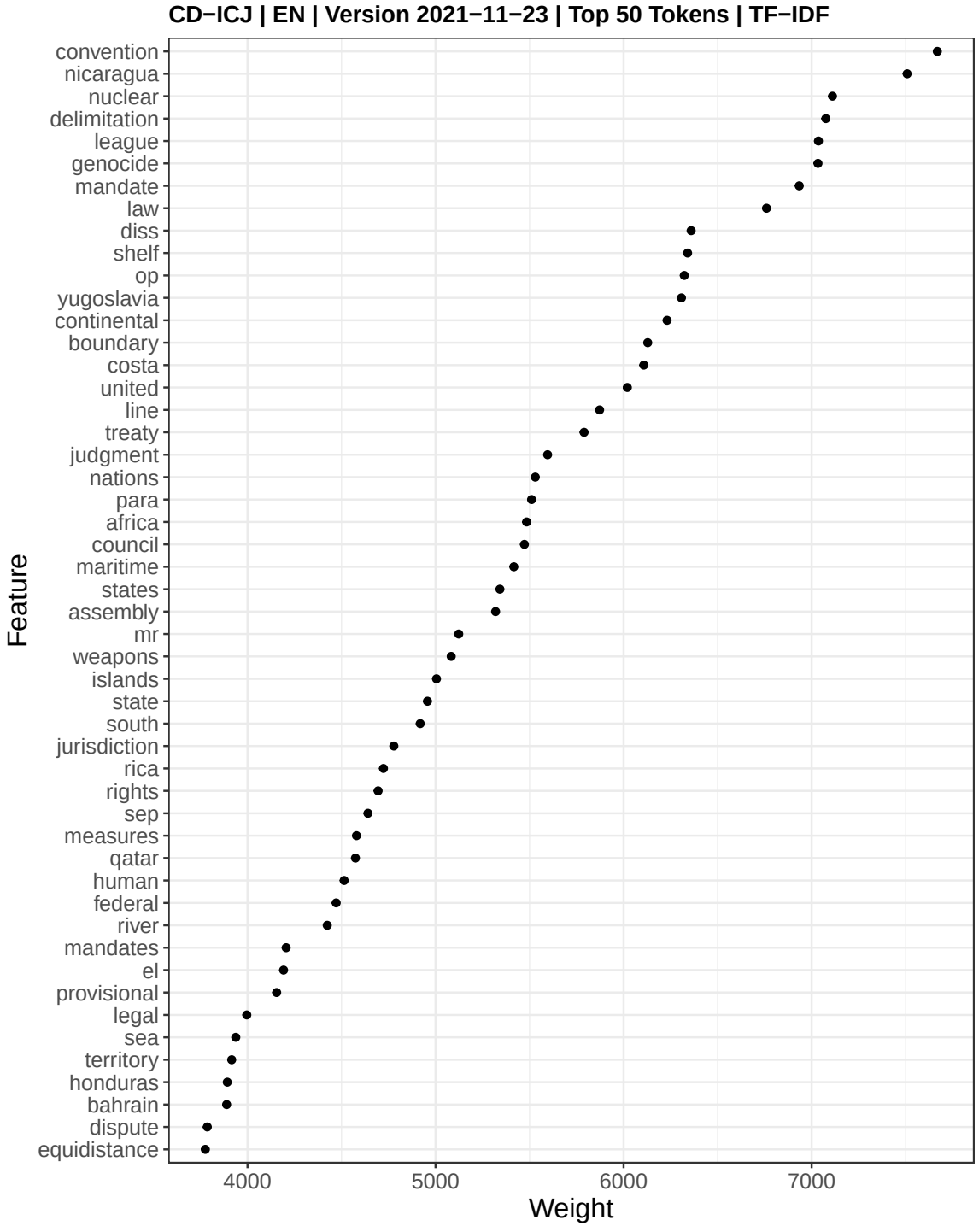


DOI: 10.5281/zenodo.3826445

25.7 Most Frequent Tokens | TFIDF Weighting | Scatterplots

25.7.1 English

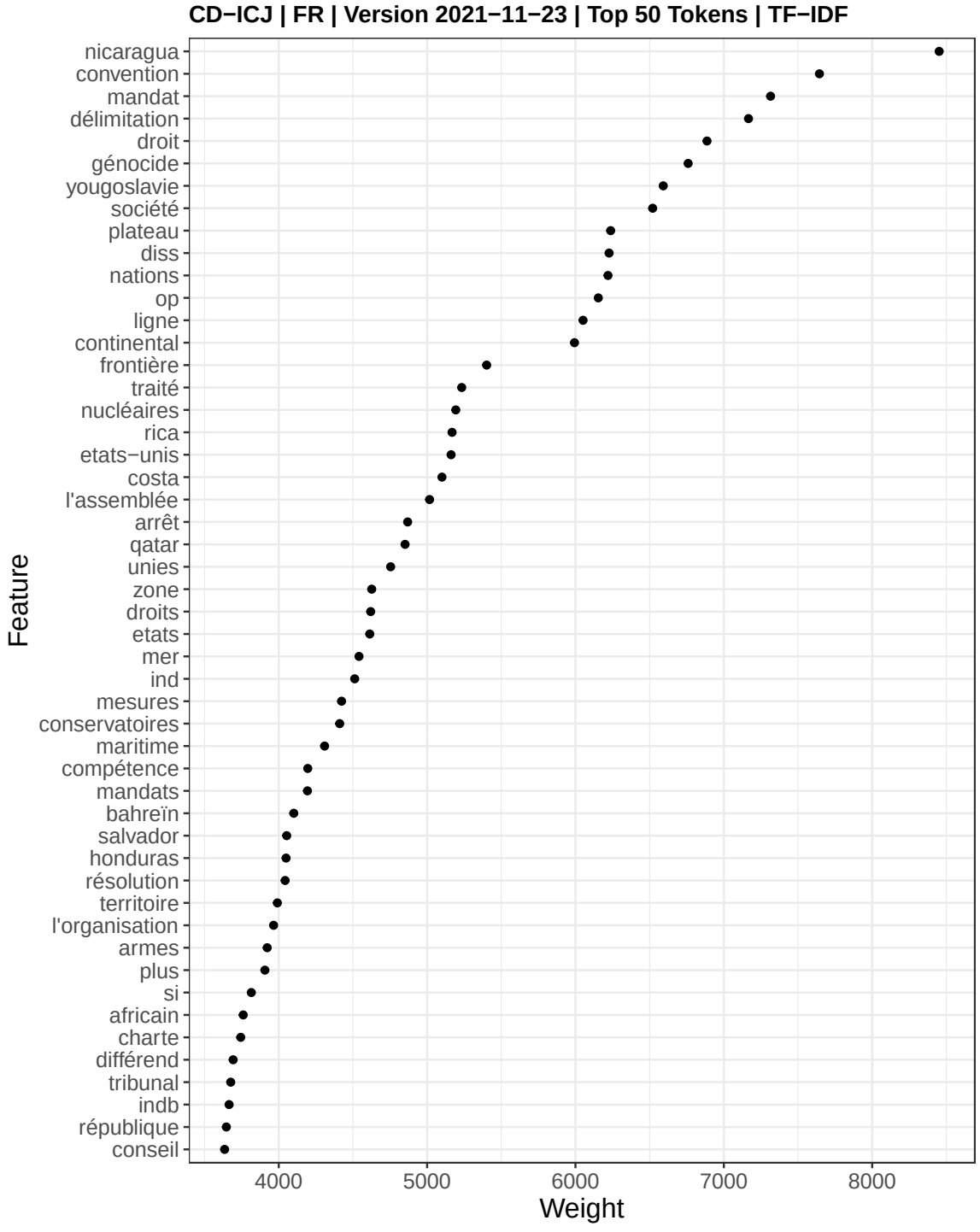
```
print(
    ggplot(data = tstat.tfidf.en[1:50, ],
           aes(x = reorder(feature,
                           frequency),
               y = frequency))+
    geom_point()+
    coord_flip()+
    theme_bw()+
    labs(
        title = paste(datashort,
                      "| EN | Version",
                      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.3826445

25.7.2 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.3826445

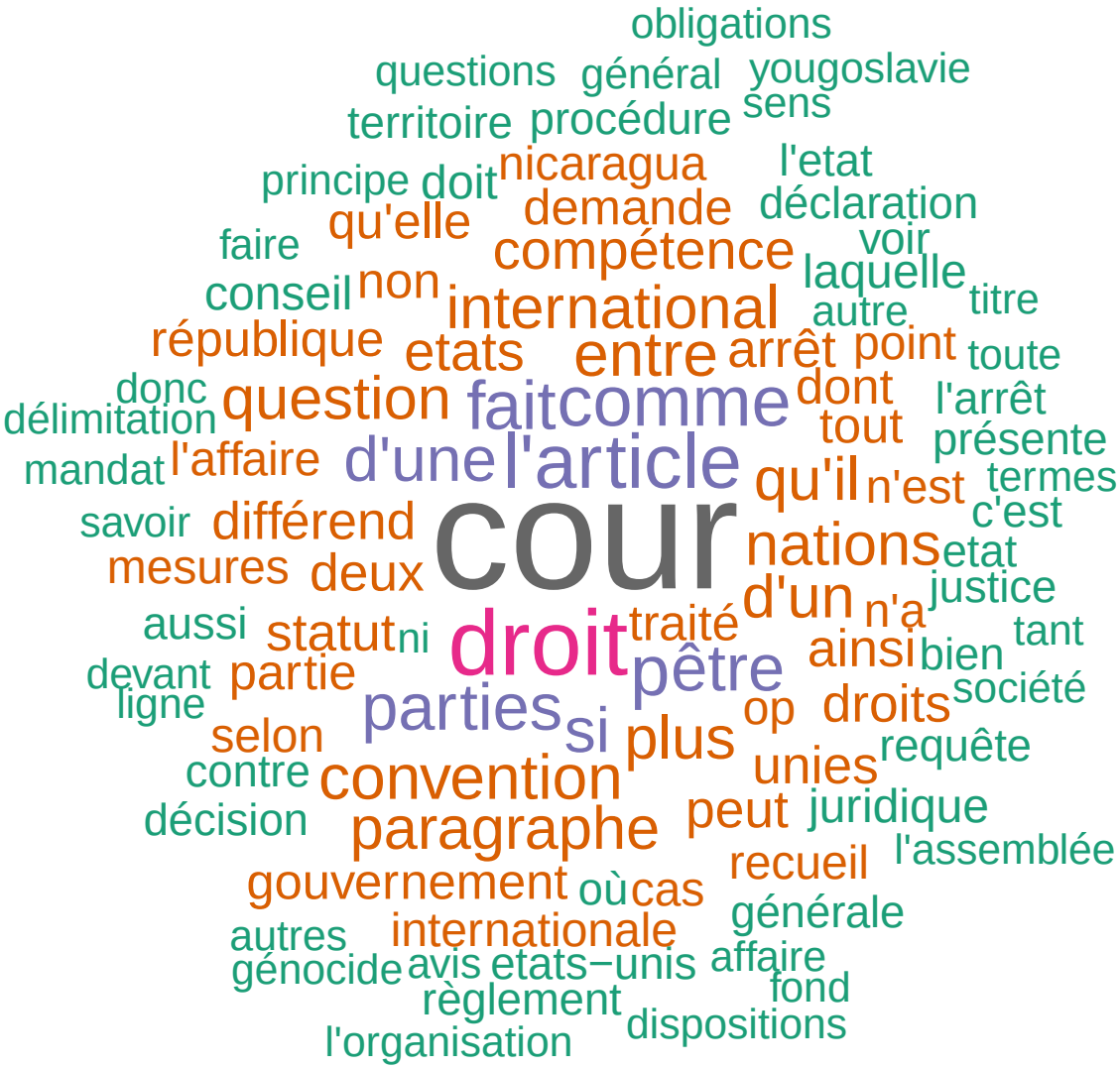
25.8 Most Frequent Tokens | TF Weighting | Wordclouds

25.8.1 English

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

25.8.2 French

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



25.9 Most Frequent Tokens | TFIDF Weighting | Wordclouds

25.9.1 English

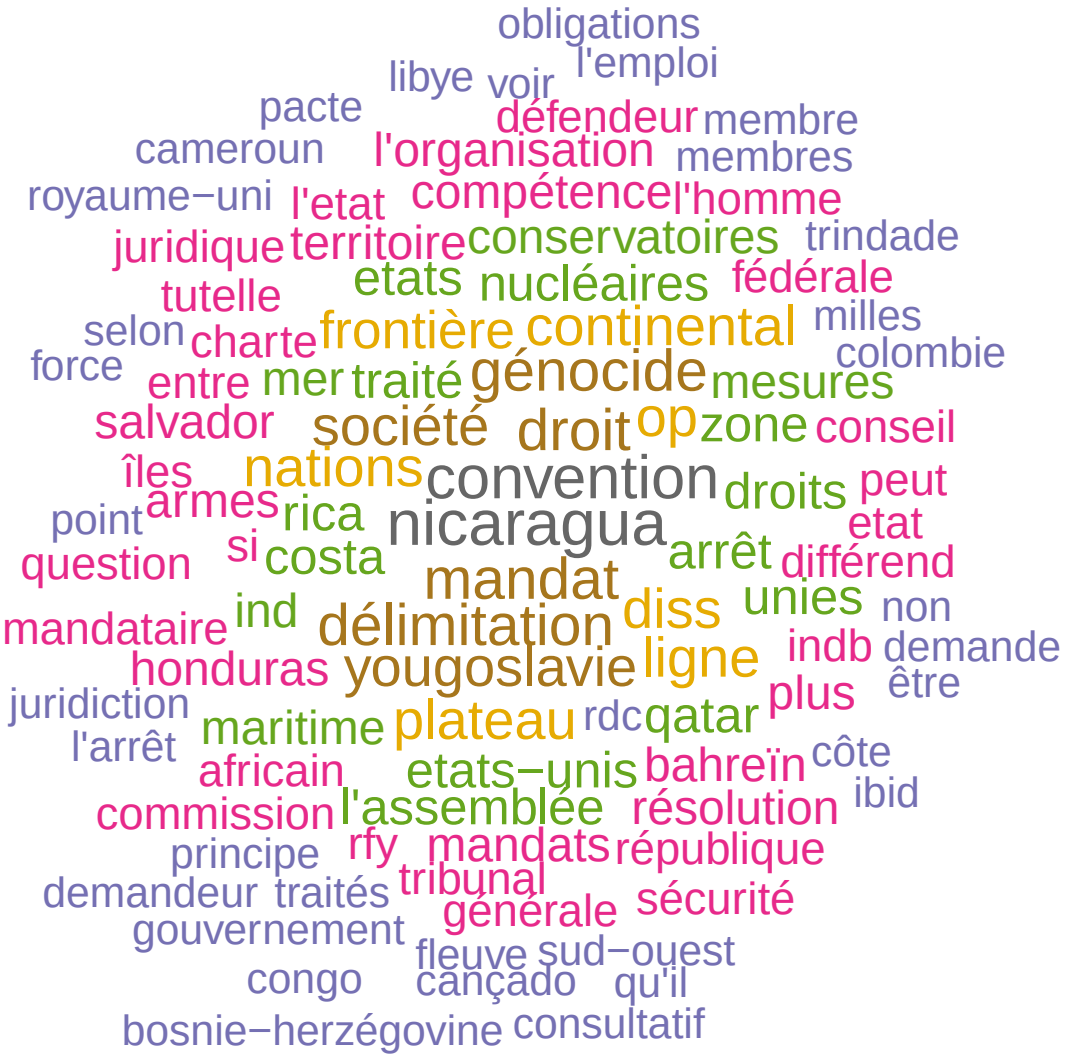
```
textplot_wordcloud(dfm.tfidf.en,
                   max_words = 100,
                   min_size = 1,
                   max_size = 2,
                   random_order = FALSE,
                   rotation = 0,
                   color = brewer.pal(8, "Dark2"))
```

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

25.9.2 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
```



26 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.

26.1 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
```

26.2 English

26.2.1 Calculate Similarity

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

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

26.2.2 Create Empty Lists

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

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

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

26.2.3 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)
```

26.2.4 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"))
}
```

26.2.5 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"))
}
```

26.2.6 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")
```

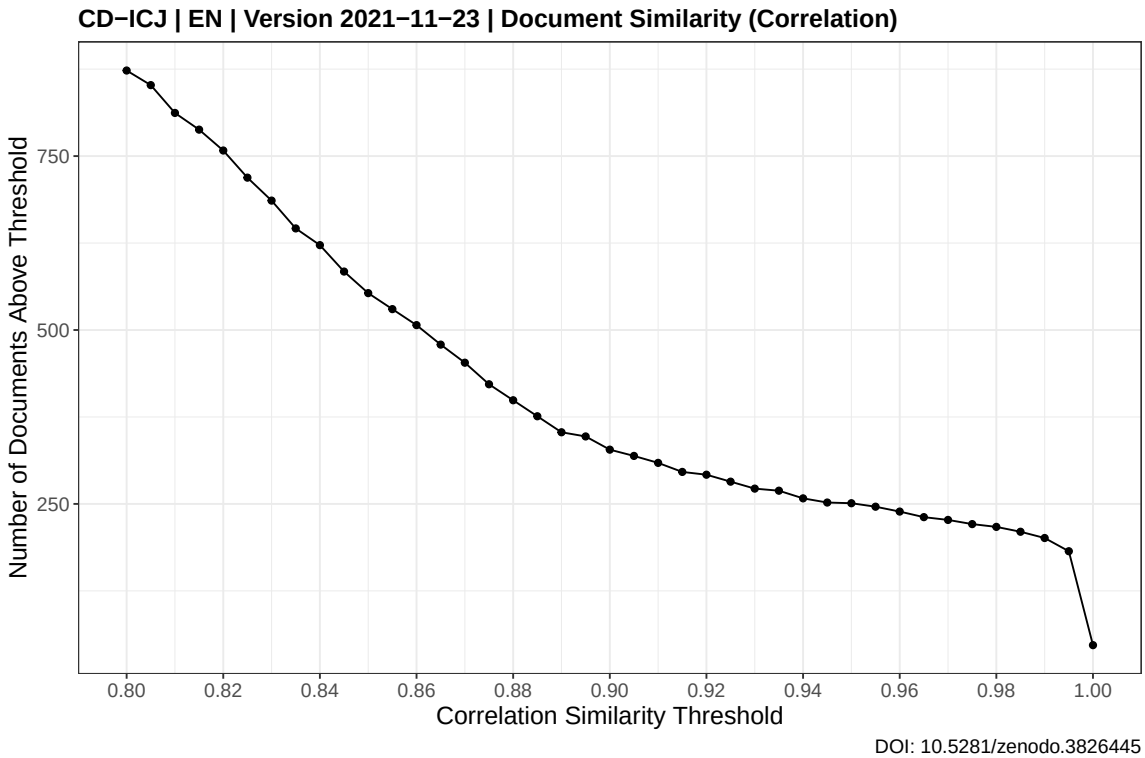
Threshold	Number to Drop
0.800	873
0.805	852
0.810	812
0.815	788
0.820	758
0.825	719
0.830	686
0.835	646
0.840	622
0.845	584
0.850	553
0.855	530
0.860	507
0.865	479
0.870	453
0.875	422
0.880	399
0.885	376
0.890	353
0.895	347
0.900	328
0.905	319
0.910	309
0.915	296
0.920	292
0.925	282
0.930	272
0.935	269
0.940	258
0.945	252

(continued)

Threshold	Number to Drop
0.950	251
0.955	246
0.960	239
0.965	231
0.970	227
0.975	221
0.980	217
0.985	210
0.990	201
0.995	182
1.000	47

```
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"
    )
)
```



26.3 French

26.3.1 Calculate Similarity

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

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

26.3.2 Create Empty Lists

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

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

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

26.3.3 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)
```

26.3.4 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"))
}
```

26.3.5 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"))

}
```

26.3.6 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")
```

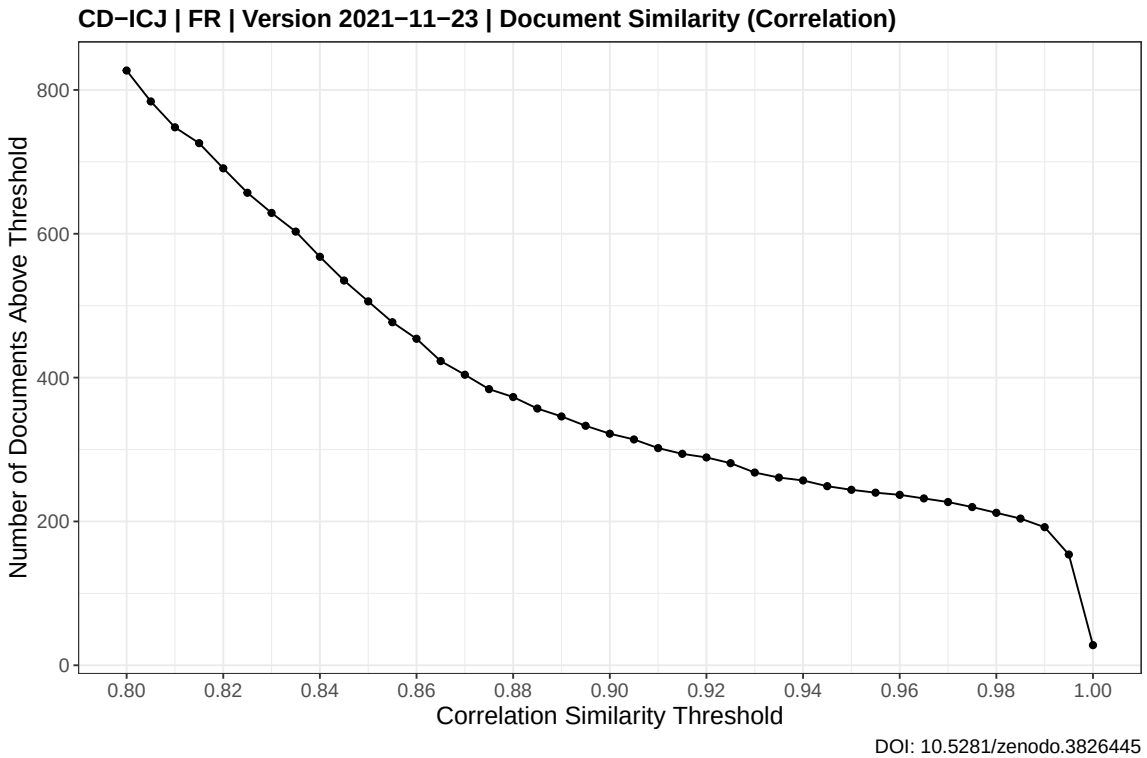
Threshold	Number to Drop
0.800	827
0.805	784
0.810	748
0.815	726
0.820	691
0.825	657
0.830	629
0.835	603
0.840	568
0.845	535
0.850	506
0.855	477
0.860	454
0.865	423
0.870	404
0.875	384
0.880	373
0.885	357
0.890	346
0.895	333
0.900	322
0.905	314
0.910	302
0.915	294
0.920	289
0.925	281
0.930	268
0.935	261
0.940	257
0.945	249

(continued)

Threshold	Number to Drop
0.950	244
0.955	240
0.960	237
0.965	232
0.970	227
0.975	220
0.980	212
0.985	204
0.990	192
0.995	154
1.000	28

```
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"
    )
)
```



27 Create CSV Files

27.1 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")
```

27.2 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")
```

28 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	2169
FR_PDF_ORIGINAL_FULL	2160
EN_PDF_ENHANCED_max2004	1484
FR_PDF_ENHANCED_max2004	1482
EN_PDF_BEST_FULL	2169
FR_PDF_BEST_FULL	2160
EN_PDF_BEST_MajorityOpinions	765
FR_PDF_BEST_MajorityOpinions	763
EN_TXT_BEST_FULL	2169
FR_TXT_BEST_FULL	2160
EN_TXT_TESSERACT_max2004	1484
FR_TXT_TESSERACT_max2004	1482
EN_TXT_EXTRACTED_FULL	2169
FR_TXT_EXTRACTED_FULL	2160

29 File Size Distribution

29.1 English

29.1.1 Corpus Object in RAM

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

```
## 80.1 Mb
```

29.1.2 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")
```

29.1.3 Total Size Comparison

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

variant	MB_total
BEST	2733.248
ORIGINAL	1346.270

29.1.4 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   63.09
```

```
# 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	638.3627
ORIGINAL	151.1586

```
# 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

10.17789	BEST
10.18234	BEST
10.27223	BEST
10.75333	ORIGINAL
10.96772	BEST
10.96772	ORIGINAL
10.96773	BEST
10.96773	ORIGINAL
11.61704	ORIGINAL
11.74513	BEST
11.74513	BEST
11.84531	ORIGINAL
11.87263	BEST
12.03522	BEST
12.35256	BEST
12.74965	BEST
12.75578	BEST
13.06277	BEST
13.12442	BEST
13.12442	ORIGINAL
13.29205	BEST
14.08529	BEST
14.25130	BEST
14.63674	ORIGINAL
14.90167	BEST
15.05370	ORIGINAL
15.49853	BEST
15.49977	BEST
15.91648	ORIGINAL
16.22027	BEST
16.22615	ORIGINAL

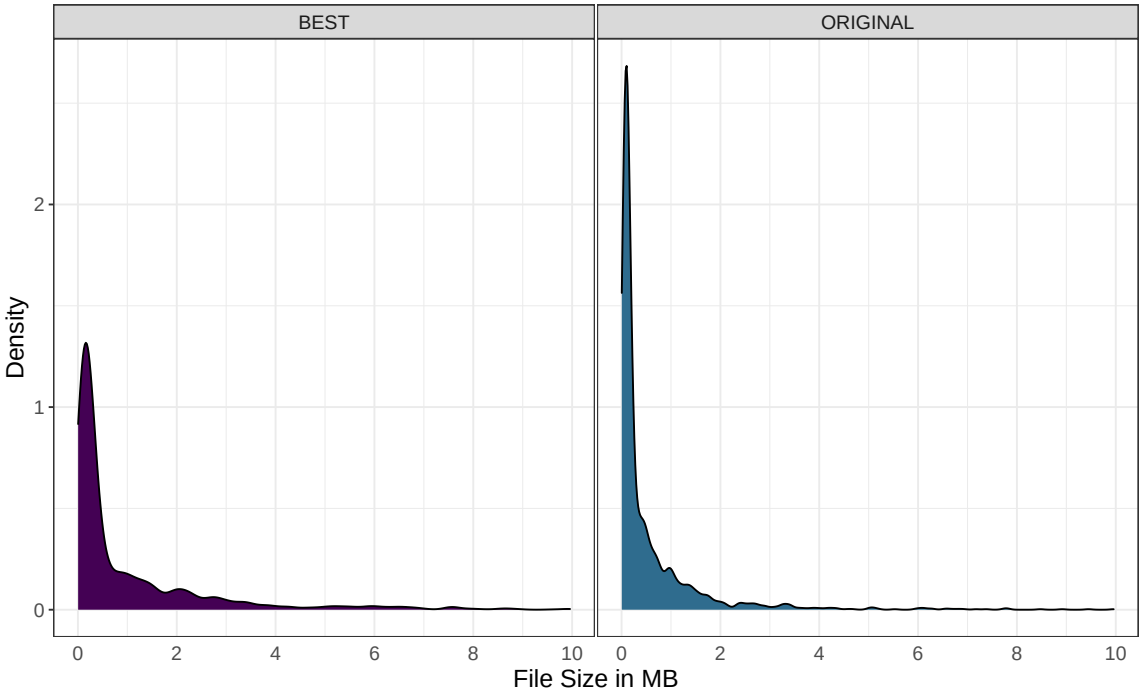
18.04835	BEST
18.77928	BEST
19.32346	BEST
19.79251	BEST
19.79428	BEST
20.80175	BEST
24.11547	BEST
24.76550	BEST
31.08605	BEST
34.08719	BEST
34.24570	BEST
42.66809	BEST
63.09498	BEST

29.1.5 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 2021−11−23 | Distribution of File Sizes up to 10 MB



DOI: 10.5281/zenodo.3826445

29.2 French

29.2.1 Corpus Object in RAM

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

```
## 80.1 Mb
```

29.2.2 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")
```

29.2.3 Total Size Comparison

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

variant	MB_total
BEST	2949.560
ORIGINAL	1416.258

29.2.4 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	837.7162
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
10.27292	BEST

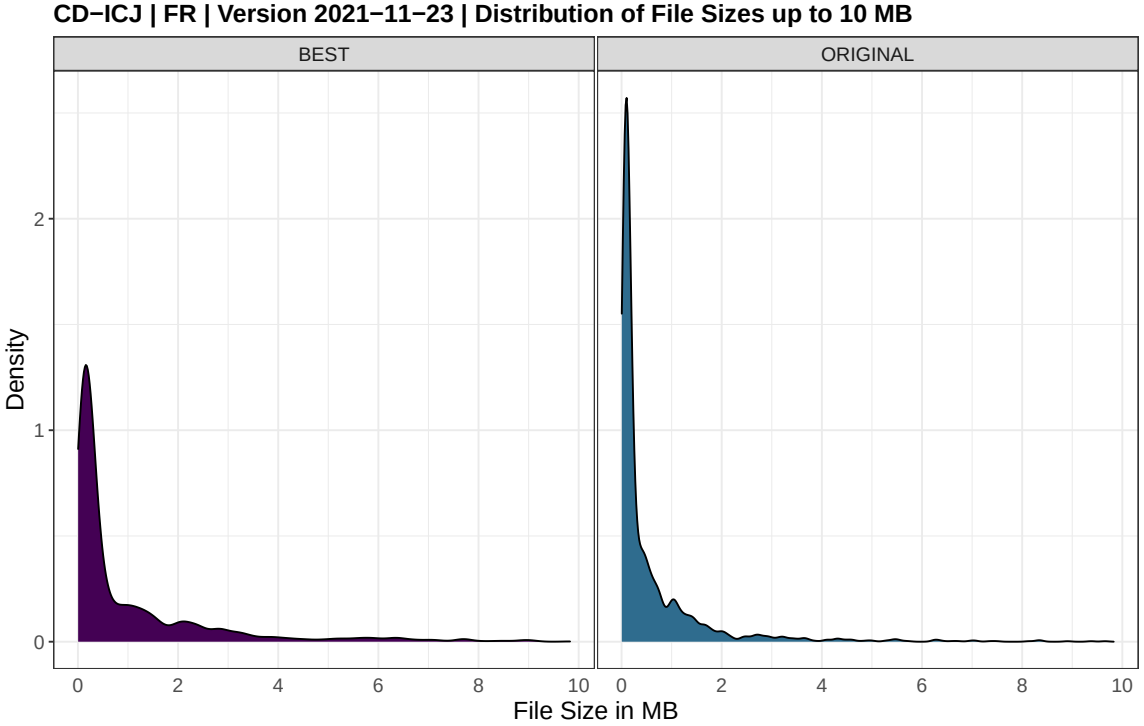
10.37768	BEST
10.37768	BEST
10.62218	ORIGINAL
10.62229	ORIGINAL
10.65338	BEST
10.66948	BEST
11.00723	ORIGINAL
11.11981	BEST
11.32830	BEST
11.32830	ORIGINAL
11.34529	BEST
11.34529	ORIGINAL
11.79728	BEST
11.91450	ORIGINAL
12.17450	BEST
12.17450	BEST
12.81959	BEST
13.01564	BEST
13.06957	ORIGINAL
13.21799	BEST
13.51871	BEST
13.51871	BEST
13.66376	BEST
13.92357	BEST
14.76233	ORIGINAL
15.17990	BEST
15.30099	BEST
15.69419	BEST
16.01056	ORIGINAL
16.44052	BEST
16.44053	BEST

16.69992	BEST
17.43610	ORIGINAL
17.47649	BEST
17.61179	ORIGINAL
18.52546	BEST
18.61407	BEST
19.10714	BEST
19.86303	BEST
20.01578	BEST
20.71829	BEST
20.71829	BEST
21.13854	BEST
21.41222	BEST
23.48634	BEST
23.85613	BEST
23.85624	BEST
26.41965	BEST
36.38976	BEST
36.89044	BEST
44.06448	BEST
44.30212	BEST
69.13688	BEST

29.2.5 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.3826445

30 Create ZIP Archives

30.1 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)
```

30.2 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)
}
```

30.3 ZIP ANALYSIS Directory

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

30.4 ZIP Unlabelled Files Directory

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

30.5 ZIP Source Files

```
files.source <- c(list.files(pattern = "Source"),
                  "data",
                  "functions",
                  "buttons")


files.source <- grep("spin",
                     files.source,
                     value = TRUE,
                     ignore.case = TRUE,
                     invert = TRUE)

zip(paste(datashort,
          datestamp,
          "Source_Files.zip",
          sep = "_"),
    files.source)
```

31 Delete CSV and Directories

The metadata CSV files are retained for Codebook generation.

31.1 Delete CSVs

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

31.2 Delete Data Directories

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

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

32 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.

32.1 Create Set of ZIP Archives

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

32.2 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)
```

}

32.3 Compute Hashes

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

```
## [1] "Parallel processing using 16 threads."
## [1] "Processed 21 files. Runtime was 13.19 secs."
```

32.4 Convert to Data Table

```
setDT(multihashes)
```

32.5 Add Index

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

32.6 Save to Disk

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

32.7 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))
```

32.8 Print to Report

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

index	filename
1	CD-ICJ_2021-11-23_EN_CSV_BEST_FULL.zip
2	CD-ICJ_2021-11-23_EN_CSV_BEST_META.zip
3	CD-ICJ_2021-11-23_EN_PDF_BEST_FULL.zip
4	CD-ICJ_2021-11-23_EN_PDF_BEST_MajorityOpinions.zip
5	CD-ICJ_2021-11-23_EN_PDF_ENHANCED_max2004.zip
6	CD-ICJ_2021-11-23_EN_PDF_ORIGINAL_FULL.zip
7	CD-ICJ_2021-11-23_EN_TXT_BEST_FULL.zip
8	CD-ICJ_2021-11-23_EN_TXT_EXTRACTED_FULL.zip
9	CD-ICJ_2021-11-23_EN_TXT_TESSERACT_max2004.zip
10	CD-ICJ_2021-11-23_EN-FR_ANALYSIS.zip
11	CD-ICJ_2021-11-23_FR_CSV_BEST_FULL.zip
12	CD-ICJ_2021-11-23_FR_CSV_BEST_META.zip
13	CD-ICJ_2021-11-23_FR_PDF_BEST_FULL.zip
14	CD-ICJ_2021-11-23_FR_PDF_BEST_MajorityOpinions.zip
15	CD-ICJ_2021-11-23_FR_PDF_ENHANCED_max2004.zip
16	CD-ICJ_2021-11-23_FR_PDF_ORIGINAL_FULL.zip
17	CD-ICJ_2021-11-23_FR_TXT_BEST_FULL.zip
18	CD-ICJ_2021-11-23_FR_TXT_EXTRACTED_FULL.zip
19	CD-ICJ_2021-11-23_FR_TXT_TESSERACT_max2004.zip
20	CD-ICJ_2021-11-23_Source_Files.zip
21	CD-ICJ_2021-11-23_UnlabelledFiles.zip

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

index	sha2.256
1	f757eebaa85237f66c1677943b098c5ec965c92299aa10fa887950ba92338e9b
2	4e2465307b5ac250d6a1acad17cbe79516268d44c19cc25fad31708108be4f32
3	893941b49bf91c717ce47bcd77d9c616710aa40ed69d93b53c5a51d9d41ee2a2
4	7c8a86c9b6efcd5e881f885414cdcdffbdb3a3aa5322c083451c005fd69048f5
5	a65b30ad4c71dd68734697ee074187979f24c053368698db748e75de0b035167
6	faa15b177aecf5d2a5133c4d4ac40a3e04f569f210c2e1bb7f333c0b182bdf05
7	27ff0eb2a9969ea1bd91f2e768950ea0a1e86d1665ddbfb96588d1e3046fe52d
8	8ed98ec8cf561cfef84452fc425fd9147bae910c630fbf9e8c242f3d6d9ab009
9	f373dec1408ed11c479bcbc2f8365ba5aba7924e04a468ad9515a7782686f2fd
10	484780f2c5baba2f39c5a8d0a65d0aa1306bcae4571285f99ac1f9cee01cff72
11	f3529616ede3e5c281fb99ca2a447e27f81781a2adfa7e490f8e70598f2239ac
12	735c453e4e3b370ae4c832415c65ad12bb3c2025a157fdd00726fd5dae216fab
13	8b5832bbb922f29e1d306e35e778f4c5deeb04ed4017308026e8d2669f870f42
14	38845a570e4f37511b6e80a7e8baa03b30989cb5dd443edd044c4a433bc6f641
15	08a7e6e5e2a5421a73a93f3dfafa43e608e0ec39247f522af0b9e385f67c5955
16	5053da1a573ded855b0aeb6505d4d7ad18357d2069ee588a59e46b72c4eaab24
17	883e60970e4d6b4769759cad863377682b14c710ab407d0322487e139cea7394
18	e232049d8e8d1f036c51b51c4e7f41430eef46cd5feda185350ab91825f0dc75
19	25f985600ac73c76cfe100ff7814800ca241cb8387c9af75d588dc7e4b3e2da5
20	53ed89572ced5dc60fd1e728a19839e5dc9dbf5e2f48381bbb4a9f45446b1154
21	523dfaa49d821bcb6e046cd8ef334b3da07e16eb4b034e352dbe7c8f7a7b8810

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

index	sha3.512
1	5dbecc3afb6c8e43545ce932246e10f23f9a19c0d8e5f1c13ed3b7c894fe54a1 e15ea7de2c82cb8952470f339f2cfda14a878b621b65f8f1a151cb9b33afbced
2	97bd154976b92fc2a62f01e4e20d0f6e4352ed6314c6fafc848b7dfe3ea32b40 e0daf7434348f439bcd00bb902e131684792e6d2de578180584294c0c04a8b16
3	58eb24f5c534994d6ad7a4ed3a70368237d831faf273f2de057a89e72eb91591 ecfe69eda98427393479c8885cb8ee0997430383ca056a71ed0d9e13e8d6ff1a
4	cc7b8ef7e0c8a14740b8d629a4939c9c72d9d7f0621e00894015bb4acbee46c2 184a547ef4c6dfea39c9d141025d372c54bc8c703a6b9082b16df58ba82f9573
5	4db436c104f9312141b18a6496a19f60801614e5fe14b2c2d54dced5bec3983a ad241ed351ec42ceeef9f159f1fd31b3207ca76aa44d651563b70fe101a95949
6	712e5c958ff247c1d5c40ba23e8aa7b009d26d5bd338f3e38fd4a07c6fcef9f2 568089f9142128553f516c74636fe64527e81fa15d62b39a2aeed81fa7b3335f
7	135af5a50e05632931f92ac68a487af2d65f5368fe0b2f0c21666175bd88286a d30b078dbc2df44e91b722224f674dd810c6dcdfacc036960e145fa13be0a0e6
8	9092fae5b4ddf415e33017c2778c7ecdd2654f2e1d9b482981fb156eb5dec959 026719c93242312dcbe76e3641c9ff85fcbb876d1bc794a318cb5cc9f9fa1110
9	3596ef6650361c5ae24b862933e3d77cfdd2e4769267c77245b1f033258bdd22 66c551b6097d973ab6ec86e1995fb4c805a2d976878dfc0d95e1b10d3d2d4700
10	d155b7000cd5b4345f96630faf329934baad322a764c2c23aa83d550dde5209d 070ad1d50ae872b9a59ab71907d28077d13b2b1bb2d4909e364e19e11209781e
11	e3df8e921ff31eca1853972171e1a7cdc37633da6f2988a30ff11eb004d51864 755df2b333fc03283edc6ffaff1ebbc763c7d192df01ba196e5a99b052b21729
12	1170c7bde43170d49e2da6f9c8b5f86fcf1d82f4810ff8dbd2ccacc8e36cca33 84494d6ddeb5c2b90d8bc046e8890dd8acc23d527b12f1fe1e5e03de2e3b77f2
13	24ccca52ec57c052e232a512806c6537d8e2a426fba372febde69edbba17f797 b14e17ad1df38bf9d9d07c1ee8a56ac77ae121ddac45d085fc1a2811a81a416f
14	61c52eab1b4ac2d8dfcf3d537474701130e7887dd72145ac5d8dd1abdc08c18e 948158312eee8cff74e7078ea291dbc14c2db9fca2c0c332e19a63a804e6413b
15	8137d752797b06859d1239550697951325710d702bfeabe508b40c49c2353ff0 2d29895ac56f2bf49789d0ca740d3c100ac33af90de00f89b53268a9a64e409f

16	c6146691ed1a32661be8500f2468c901941be8ff7b611f096b0e880fbf81952e 9a85229a1043546d0fb18e2cfcb9fc2800c2c06c87880857833263f4b83028fb
17	3fc01e0e5f439b7295ba0b96caa6660472000ec4b991cef94385d1cd1d348bb7 ca73edd6b66829077f2ea326fd4c49c6e8d306db847d708f742bf02786e25b14
18	e5f5b32f77276210e1c82012c38573a202911eff6a42436bd68608136300eb06 8bf2f50a74c73e2f7520c41a3292c125ac02519bdcbf408840fdca36eb98766c
19	7ba538cedf69c138583df0dc139a165b40f5f1937a920be4c2339cdd025b9b88 b7eb81137ab678898d8e7948e838977e7f4adc9a7fc6ddf57e43165d12d59ff8
20	695875b2e3d4197a0a1029d2a326ca36bec1beaa79c1505d41c4e3967d91d97f 1ab5c8676ea50ec8bf6cdd40cb4e539c69d87d6a721a69cc5d37dad628bcead4
21	61d8afa456e426af9d19cb6f7a40d3904885a1c6cebe4f7cbbc63a639552f74e 22f2a724a25c50ec7bc2b913b5c9e775f233ac8bd608139327e04e9aae5fd961

33 Finalize

33.1 Datestamp

```
print(datestamp)
```

```
## [1] "2021-11-23"
```

33.2 Date and Time (Begin)

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

```
## [1] "2021-11-23 03:59:58 CET"
```

33.3 Date and Time (End)

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

```
## [1] "2021-11-23 14:17:41 CET"
```

33.4 Script Runtime

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

```
## Time difference of 10.29522 hours
```

33.5 Warnings

```
warnings()
```

34 Strict Replication Parameters

```
sessionInfo()
```

```
## R version 4.0.5 (2021-03-31)
## Platform: x86_64-redhat-linux-gnu (64-bit)
## Running under: Fedora 34 (Workstation Edition)
##
## Matrix products: default
## BLAS/LAPACK: /usr/lib64/libflexiblas.so.3.0
##
## 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_1.0.16         iterators_1.0.13
##  [3] foreach_1.5.1             data.table_1.14.0
##  [5] textcat_1.0-7             quanteda.textplots_0.94
##  [7] quanteda.textstats_0.94.1 quanteda_3.1.0
##  [9] readtext_0.81             RColorBrewer_1.1-2
## [11] viridis_0.6.1             viridisLite_0.4.0
## [13] scales_1.1.1              ggplot2_3.3.5
## [15] rsvg_2.1                  DiagrammeRsvg_0.1
## [17] DiagrammeR_1.0.6.1        magick_2.7.3
## [19] kableExtra_1.3.4          knitr_1.34
## [21] fs_1.5.0                  pdftools_3.0.1
## [23] stringr_1.4.0             mgsub_1.7.3
## [25] rvest_1.0.1               httr_1.4.2
##
## loaded via a namespace (and not attached):
##  [1] jsonlite_1.7.2      RcppParallel_5.1.4  askpass_1.1
##  [4] highr_0.9           selectr_0.4-2       yaml_2.2.1
##  [7] slam_0.1-48         qpdf_1.1            pillar_1.6.2
## [10] lattice_0.20-44     glue_1.4.2          digest_0.6.27
## [13] tau_0.0-24          colorspace_2.0-2    htmltools_0.5.2
## [16] Matrix_1.3-4        pkgconfig_2.0.3     ISOcodes_2021.02.24
## [19] purrr_0.3.4         webshot_0.5.2       svglite_2.0.0
## [22] nsyllable_1.0       tibble_3.1.4        farver_2.1.0
## [25] generics_0.1.0      ellipsis_0.3.2      withr_2.4.2
## [28] magrittr_2.0.1      crayon_1.4.1        evaluate_0.14
## [31] stopwords_2.2       fansi_0.5.0         xml2_1.3.2
## [34] tools_4.0.5         lifecycle_1.0.0     V8_3.4.2
## [37] munsell_0.5.0       compiler_4.0.5      proxyC_0.2.1
## [40] systemfonts_1.0.2   rlang_0.4.11        grid_4.0.5
```

```
## [43] rstudioapi_0.13     htmlwidgets_1.5.4   visNetwork_2.0.9
## [46] labeling_0.4.2      rmarkdown_2.10      gtable_0.3.0
## [49] codetools_0.2-18    curl_4.3.2          R6_2.5.1
## [52] gridExtra_2.3       dplyr_1.0.7         fastmap_1.1.0
## [55] utf8_1.2.2          fastmatch_1.1-3     stringi_1.7.4
## [58] Rcpp_1.0.7          vctrs_0.3.8         tidyselect_1.1.1
## [61] xfun_0.25
```

```
system2("openssl",
        "version",
        stdout = TRUE)
```

```
## [1] "OpenSSL 1.1.1l  FIPS 24 Aug 2021"
```

```
system2("tesseract",
        "-v",
        stdout = TRUE)
```

```
## [1] "tesseract 4.1.1"
## [2] " leptonica-1.81.1"
## [3] "  libgif 5.2.1 : libjpeg 6b (libjpeg-turbo 2.0.90) : libpng 1.6.37 :
    libtiff 4.2.0 : zlib 1.2.11 : libwebp 1.2.1"
## [4] " Found AVX2"
## [5] " Found AVX"
## [6] " Found FMA"
## [7] " Found SSE"
```

```
system2("convert",
        "--version",
        stdout = TRUE)
```

```
## [1] "Version: ImageMagick 6.9.11-27 Q16 x86_64 2021-01-25 https://imagemagick.
    org"
## [2] "Copyright: (c) 1999-2020 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 fftw 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. 2020. *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. 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.

———. 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.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, Akitaka Matsuo, and William Lowe. 2021. *Quanteda: Quantitative Analysis of Textual Data*. https://quanteda.io.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2021. *Quanteda.textplots: Plots for the Quantitative Analysis of Textual Data*. https://CRAN.R-project.org/package=quanteda.textplots.

Corporation, Microsoft, and Steve Weston. 2020. *DoParallel: Foreach Parallel Adaptor for the Parallel Package*. https://CRAN.R-project.org/package=doParallel.

Dowle, Matt, and Arun Srinivasan. 2021. *Data.table: Extension of 'Data.frame'*. https://CRAN.R-project.org/package=data.table.

Ewing, Mark. 2021. *Mgsub: Safe, Multiple, Simultaneous String Substitution*. https://CRAN.R-project.org/package=mgsub.

Garnier, Simon. 2021a. *Viridis: Colorblind-Friendly Color Maps for R*. https://CRAN.R-project.org/package=viridis.

———. 2021b. *ViridisLite: Colorblind-Friendly Color Maps (Lite Version)*. https://CRAN.R-project.org/package=viridisLite.

Hester, Jim, and Hadley Wickham. 2020. *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.

———. 2020. *DiagrammeR: Graph/Network Visualization*. https://github.com/rich-iannone/DiagrammeR.

Neuwirth, Erich. 2014. *RColorBrewer: ColorBrewer Palettes*. https://CRAN.R-project.org/package=RColorBrewer.

Ooms, Jeroen. 2020. *Rsvg: Render Svg Images into Pdf, Png, Postscript, or Bitmap Arrays*. https://github.com/jeroen/rsvg#readme.

———. 2021a. *Magick: Advanced Graphics and Image-Processing in R*. https://CRAN.R-project.org/package=magick.

———. 2021b. *Pdftools: Text Extraction, Rendering and Converting of Pdf Documents*. https://CRAN.R-project.org/package=pdftools.

R Core Team. 2021. *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.

———. 2019. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://CRAN.R-project.org/package=stringr.

———. 2020. *Httr: Tools for Working with Urls and Http*. https://CRAN.R-project.org/package=httr.

———. 2021. *Rvest: Easily Harvest (Scrape) Web Pages*. https://CRAN.R-project.org/package=rvest.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. 2021. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://CRAN.R-project.org/package=ggplot2.

Wickham, Hadley, and Dana Seidel. 2020. *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/.

———. 2021. *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.