Assignment 1: JSON Flattening and Analysis

Venkata Lakshmi Parimala Pasupuleti

2025-07-15

In this assignment, we process a nested JSON file containing DMCA notices. The workflow involves flattening the data, extracting domain and IP information, parallelizing IP resolution, and generating summaries to understand patterns in the data.

Step 1: Flatten the JSON

```
json_path <- "D:/projects/GTAGRA project/Assignment-1/response.json"
json_data <- fromJSON(json_path, simplifyVector = FALSE)

notices <- json_data$notices
records <- list()

for (notice in notices) {
  works <- notice$works
  if (!is.null(works) && length(works) > 0) {
    for (work in works) {
      desc <- if (!is.null(work$description)) work$description else NA
      urls <- work$infringing_urls
      if (!is.null(urls) && length(urls) > 0) {
        for (u in urls) {
          if (!is.null(u$url)) {
            row <- list(
              id = notice$id,
              type = notice$type,
              title = notice$title,
              date_sent = notice$date_sent,
              date_received = notice$date_received,
              sender_name = notice$sender_name,
              principal_name = notice$principal_name,
              recipient_name = notice$recipient_name,
              work_description = desc,
              infringing_url = u$url
            )
            records <- append(records, list(row))
          }
        }
      }
    }
  }
}
```

```r
df <- bind_rows(records)
df <- df %>% mutate_all(as.character)
write.csv(df, "D:/projects/GTAGRA project/Assignment-1/R/flattened_step1_R.csv", row.names = FALSE, file

head(df)
```

```
## # A tibble: 6 x 10
##   id       type  title        date_sent date_received sender_name principal_name
##   <chr>    <chr> <chr>        <chr>     <chr>         <chr>       <chr>
## 1 23878572 DMCA  DMCA (Copyr~ 2021-05-~ 2021-05-24T0~ 3Ants Deve~ Netflix
## 2 23878572 DMCA  DMCA (Copyr~ 2021-05-~ 2021-05-24T0~ 3Ants Deve~ Netflix
## 3 23878572 DMCA  DMCA (Copyr~ 2021-05-~ 2021-05-24T0~ 3Ants Deve~ Netflix
## 4 23878572 DMCA  DMCA (Copyr~ 2021-05-~ 2021-05-24T0~ 3Ants Deve~ Netflix
## 5 23878572 DMCA  DMCA (Copyr~ 2021-05-~ 2021-05-24T0~ 3Ants Deve~ Netflix
## 6 23878572 DMCA  DMCA (Copyr~ 2021-05-~ 2021-05-24T0~ 3Ants Deve~ Netflix
## # i 3 more variables: recipient_name <chr>, work_description <chr>,
## #   infringing_url <chr>
```

Step 2 & 3: Create domain and IP columns using 4 CPUs

```r
library(urltools)
library(parallel)
library(dplyr)

# Extract domain
df$domain <- domain(df$infringing_url)

# Function to resolve IP address
get_ip <- function(domain) {
  tryCatch({
    ip <- system(paste("nslookup", domain), intern = TRUE)
    addr <- grep("Address", ip, value = TRUE)
    ip_value <- if (length(addr) > 0) {
      gsub("Address: ", "", tail(addr, 1))
    } else {
      NA_character_
    }
    ip_value
  }, error = function(e) NA_character_)
}

# Get unique domains
unique_domains <- unique(df$domain)

# Parallel IP resolution
cl <- makeCluster(4)
clusterExport(cl, varlist = c("get_ip"))
ips <- parSapply(cl, unique_domains, get_ip)
stopCluster(cl)

# Create IP mapping
df_ip <- data.frame(domain = unique_domains, ip_address = ips, stringsAsFactors = FALSE)
df <- left_join(df, df_ip, by = "domain")
```

```
write.csv(df, "D:/projects/GTAGRA project/Assignment-1/R/flattened_step2-3_R.csv", row.names = FALSE, f
head(df[, c("infringing_url", "domain", "ip_address")])
```

```
## # A tibble: 6 x 3
##   infringing_url                                                domain ip_address
##   <chr>                                                         <chr>  <chr>
## 1 https://www.poseidonhd.in/pelicula/ver-online-el-baile-de-l~ www.p~ " 95.215.~
## 2 https://www1.cuevana3.video/13445/el-baile-de-los-41         www1.~ "Addresse~
## 3 https://pelisplus.live/el-baile-de-los-41-2021-online-latin~ pelis~ " 185.130~
## 4 https://pelis28.nu/ver-pelicula/el-baile-de-los-41-a2c3t5y6~ pelis~ " 104.247~
## 5 https://pelis24.app/ver-el-baile-de-los-41-online-espanol/   pelis~ "Addresse~
## 6 https://pelis-123.com/peliculas/el-baile-de-los-41/          pelis~ "Addresse~
```

Step 4: Summarizations

In this final step, we generate three summaries to better understand the dataset:

- **Summary 1**: Top 10 domains by number of infringing URLs.
- **Summary 2**: Number of unique notices sent by each sender.
- **Summary 3**: Number of infringing URLs per work description.

```
summary1 <- df %>%
  count(domain, sort = TRUE) %>%
  head(10)
write.csv(summary1, "D:/projects/GTAGRA project/Assignment-1/R/summary_urls_per_domain_R.csv", row.name
summary1
```

```
## # A tibble: 10 x 2
##    domain                              n
##    <chr>                           <int>
##  1 chomikuj.pl                     23605
##  2 watchepisodeseries.unblockit.onl  5422
##  3 rapidgator.net                    1760
##  4 www.torlock.cc                     932
##  5 ul.to                              851
##  6 drive.google.com                   730
##  7 www.filefactory.com                701
##  8 ok.ru                              662
##  9 vidlox.me                          614
## 10 1337x.mrunblock.surf               555
```

```
summary2 <- df %>%
  group_by(sender_name) %>%
  summarise(notice_count = n_distinct(id)) %>%
  arrange(desc(notice_count))
write.csv(summary2, "D:/projects/GTAGRA project/Assignment-1/R/summary_notices_per_sender_R.csv", row.na
head(summary2, 10)
```

```
## # A tibble: 10 x 2
##    sender_name                          notice_count
##    <chr>                                       <int>
##  1 Vobile Inc                                    220
##  2 3Ants Development & Strategies S.L.           143
##  3 Marketly llc                                   63
##  4 MarkScan                                       46
```

```
##  5 MEDIA STORY                                    10
##  6 MEDIASTORY                                      6
##  7 3ants D&S                                       5
##  8 MediaStory                                      2
##  9 Brad Bo                                         1
## 10 Media Story                                     1
```

```
summary3 <- df %>%
  group_by(work_description) %>%
  summarise(url_count = n()) %>%
  arrange(desc(url_count))
write.csv(summary3, "D:/projects/GTAGRA project/Assignment-1/R/summary_urls_per_work_R.csv", row.names =
head(summary3, 10)
```

```
## # A tibble: 10 x 2
##    work_description                                                  url_count
##    <chr>                                                                 <int>
##  1 "Stranger Things"                                                      6566
##  2 "Jupiter's Legacy"                                                     3745
##  3 "The Witcher"                                                          2461
##  4 "The Crown"                                                            2360
##  5 "House of Cards: Chapter 1\n"                                          1850
##  6 "Army of the Dead"                                                     1482
##  7 "Sacred Games"                                                         1426
##  8 "Stranger Things: Chapter One: The Vanishing Of Will Byers\n"          1403
##  9 "Sense8\n"                                                             1204
## 10 "Things Heard & Seen\n"                                                1035
```

Done!