itle: “Group Activity”
uthor: “Renz Moquete, Adrian Sayson, Rashir John Castillano, Michael Simpron”
ate: “2025-12-01”
utput:
pdf_document:
latex_engine: xelatex
html_document: default

Loading the library

```
library(rvest)
library(dplyr)
library(stringr)
library(lubridate) ## For date formats
library(ggplot2)
```

1. Creating an object

```
titles <- character(0)
authors <- character(0)
submission_dates <- character(0)
originally_announced <- character(0)
doi <- character(0)
```

2. Importing the url and created a structure

```
# Base URL for Nuclear Theory
base_url <- "https://arxiv.org/search/?query=Nuclear+Theory&searchtype=all&source=header&start="

all_papers <- list()

# Loop 4 times to get 200 papers (0, 50, 100, 150)
starts <- seq(from = 0, to = 150, by = 50)

for (i in starts) {

  # Construct URL
  url <- paste0(base_url, i)
  print(paste("Scraping:", url))

  # STANDARD SCRAPING
  tryCatch({
    page <- read_html(url)

    # Extract containers
    papers_html <- page %>% html_nodes("li.arxiv-result")

    # Extract Data Elements
    titles <- papers_html %>%
      html_node("p.title.is-5.mathjax") %>%
      html_text(trim = TRUE)
```

```
    authors <- papers_html %>%
      html_node("p.authors") %>%
      html_text(trim = TRUE) %>%
      str_remove("Authors:\n")

    # CLEANER: Removes "Less" and any invisible/weird symbols like triangles
    abstracts <- papers_html %>%
      html_node("span.abstract-full") %>%
      html_text(trim = TRUE) %>%
      str_remove("Less") %>%
      str_remove_all("[^[:ascii:]]")

    meta_raw <- papers_html %>%
      html_node("p.is-size-7") %>%
      html_text(trim = TRUE)

    # Store in temporary dataframe
    temp_df <- data.frame(
      title = titles,
      author = authors,
      abstract = abstracts,
      meta_raw = meta_raw,
      stringsAsFactors = FALSE
    )

    all_papers[[length(all_papers) + 1]] <- temp_df

  }, error = function(e) {
    print(paste("Error on page starting at", i))
  })

  # Wait 3 seconds
  Sys.sleep(3)
}
```

```
## [1] "Scraping: https://arxiv.org/search/?query=Nuclear+Theory&searchtype=all&source=header&start=0"
## [1] "Scraping: https://arxiv.org/search/?query=Nuclear+Theory&searchtype=all&source=header&start=50"
## [1] "Scraping: https://arxiv.org/search/?query=Nuclear+Theory&searchtype=all&source=header&start=100
## [1] "Scraping: https://arxiv.org/search/?query=Nuclear+Theory&searchtype=all&source=header&start=150
```

```
# Combine all lists into one dataframe
df_papers <- bind_rows(all_papers)

# Check count
print(paste("Total papers extracted:", nrow(df_papers)))
```

```
## [1] "Total papers extracted: 200"
```

3. Cleaning the data

```
df_clean <- df_papers %>%
  mutate(
    # 1. Extract Submission Date
    submission_date_text = str_extract(meta_raw, "Submitted.*?(=?;)"),
```

```
    submission_date_text = str_remove_all(submission_date_text, "Submitted |;"),
    submission_date = dmy(submission_date_text),

    # 2. Extract DOI
    doi = str_extract(meta_raw, "doi:.*"),
    doi = str_remove(doi, "doi:"),

    # 3. Extract Announced Date (Backup)
    announced_date_text = str_extract(meta_raw, "originally announced [A-Za-z]+ [0-9]{4}"),
    announced_date_text = str_remove(announced_date_text, "originally announced "),
    originally_announced = my(announced_date_text)
  )

# Remove rows where date might have failed
df_clean <- df_clean %>% filter(!is.na(submission_date))

# Show the first few rows to check data
head(df_clean %>% select(title, submission_date, doi))
```

```
##                                                                                title
## 1                                          Short-range production of three bottom mesons
## 2               New Physics Searches via Beam Normal Spin Asymmetry in Bhabha Scattering
## 3 Pion photoproduction of nucleon excited states with Hamiltonian effective field theory
## 4             On the possibility of superradiant neutrino emission by atomic condensates
## 5  Relativistic recoil as a key to the fine-structure puzzle in muonic $^{90}\\text{Zr}$
## 6                   Study of $0^+$ and $8^-$ states in even-even $^{250-260}$No isotopes
##   submission_date  doi
## 1      2025-11-27 <NA>
## 2      2025-11-27 <NA>
## 3      2025-11-27 <NA>
## 4      2025-11-27 <NA>
## 5      2025-11-27 <NA>
## 6      2025-11-27 <NA>
```

Arangging The dates

```
df_sorted <- df_clean %>%
  arrange(submission_date)

# Display sorted dates to verify
head(df_sorted$submission_date)
```

```
## [1] "2025-11-01" "2025-11-01" "2025-11-02" "2025-11-02" "2025-11-02"
## [6] "2025-11-03"
```

5. Turning into a plot time series

```
# Count papers per month
papers_per_day <- df_sorted %>%
  mutate(day = floor_date(submission_date, "day")) %>%
  group_by(day) %>%
  summarise(count = n())

# Plot
```

```r
ggplot(papers_per_day, aes(x = day, y = count)) +
  geom_line(color = "darkblue", linewidth = 0.8) + # Connect the days
  geom_point(color = "red", size = 2) +            # Show the dots
  labs(title = "Time Series: Nuclear Theory Papers (arXiv)",
       subtitle = "Number of papers submitted per Day (Last 200 papers)",
       x = "Date",
       y = "Number of Papers") +
  scale_x_date(date_breaks = "1 week", date_labels = "%b %d") + # Make dates easier to read
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```